Pós-Graduação em Ciência da Computação

"DMBUILDING: UMA METODOLOGIA PARA MONTAGEM DE VISÕES EM BASES DE DADOS DIRIGIDAS A PROBLEMAS DE MINERAÇÃO DE DADOS"

Por

DANIELA CARGNIN

Dissertação de Mestrado

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
www.cin.ufpe.br/~posgraduacao

RECIFE, JUNHO/2008

Pós-Graduação em Ciência da Computação

"DMBUILDING: UMA METODOLOGIA PARA MONTAGEM DE VISÕES EM BASES DE DADOS DIRIGIDAS A PROBLEMAS DE MINERAÇÃO DE DADOS"

Por

DANIELA CARGNIN

Dissertação de Mestrado

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
www.cin.ufpe.br/~posgraduacao

RECIFE, JUNHO/2008

Cargnin, Daniela

***Dmbuilding*: Uma metodologia para montagem de visões em bases de dados dirigidas a problemas de mineração de dados / Daniela Cargnin. – Recife : O Autor, 2008.**

x, 141 folhas : il., fig., tab.

Dissertação (mestrado) – Universidade Federal de Pernambuco. CIn. Ciência da Computação, 2008.

Inclui bibliografia e glossário.

1. Mineração de dados. I. Título.

006.312 CDD (22.ed.) MEI2008-065

Agradecimentos

Em primeiro lugar, a Deus, Senhor de todas as coisas e Arquiteto do Universo. Sem Ele, nada disso seria possível.

Aos meus amados pais, Luiz e Maria, por toda a dedicação, amor, carinho e luz nesses anos de minha vida. Vocês são meu espelho de sabedoria, garra, determinação e amor. Amo-lhes e tudo que sou devo a vocês.

A restante da minha família (Sônia, Flávio, João, Tiago, Natália, Rafael, Bruna e Gabriel) por todo o carinho e amor. Família é a base se tudo e sou abençoada pela família maravilhosa que tenho. Além disso, aos tios e tias, primos e primas que estiveram sempre acompanhando essa fase, entendendo principalmente a ausência.

Ao amigo Bruno Amorim, por ter sido muito mais que um amigo, mas sim um guia, uma inspiração, fonte de luz e esperança.

A amiga Cassandra Abrantes, pelo companheirismo, ajuda, compreensão, alegria, carinho e amor. Sem você Kau, tudo teria sido muito mais difícil. Obrigada a sempre me ajudar a tirar todas as pedras do meu caminho e nunca me deixar cair.

Ao professor Germano Vasconcelos pela orientação durante o desenvolvimento deste trabalho.

A Empresa NeuroTech, por ter gentilmente concedido o uso de suas ferramentas para a realização dessa dissertação, como também a todos os colegas da mesma, pela ajuda, amizade e alegria em todos os momentos desse processo.

E a todos os amigos, que muitas vezes sofreram e choraram comigo, mas que comigo hoje também sorriem: Alba, Alessandra, Aline, Antônio, Araken, Bianca, Bruno Sena, Coxinha, Cristiane, Cristini, Danielle, Fabio (RJ), Gabriela, Genis, Glaucya, Jackie, Jennefer, Júlio, Maíra, Marco, Mário, Marta Milanez, Mary, Natália, Priscila, Renata, Renato, Roberto, Rosana, Rozinha, Schoaba, Tarcísio, Tayza, Valci, Valdênia e Victor.

DMBUILDING: UMA METODOLOGIA PARA MONTAGEM DE VISÕES EM BASES DE DADOS DIRIGIDAS A PROBLEMAS DE MINERAÇÃO DE DADOS

Resumo

Os avanços tecnológicos têm aumentado drasticamente a magnitude dos dados armazenados em diversos domínios de aplicação. Esta abundância de dados tem excedido a capacidade de análise humana. Como conseqüência, algumas informações valiosas escondidas nestes grandes volumes de dados não são descobertas. Este cenário impulsionou a criação de várias técnicas capazes de extrair conhecimento de grandes volumes de dados. Algumas dessas técnicas são resultantes do emergente campo de Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases* - KDD).

O processo de KDD é composto de várias etapas. A etapa de preparação dos dados consome de 50% a 90% do tempo e esforço necessário para a realização de todo o processo. Quanto mais completa e consistente for a preparação, melhor será o resultado da mineração de dados. Uma forma de garantir a completude e a consistência dos dados é utilizar uma metodologia que aborde detalhadamente todas as atividades relacionadas à preparação dos dados.

Muitas metodologias foram propostas para o desenvolvimento de projetos de KDD. Apesar da maioria citar o processo de preparação dos dados, poucas metodologias específicas para montagem de visão de dados têm sido propostas. Diante deste cenário, esta dissertação tem como objetivos investigar as metodologias para o desenvolvimento de projetos de KDD, enfatizando os aspectos relacionados à preparação dos dados, e como resultado da investigação, propor uma metodologia para montagem de visões em bases de dados dirigidas a problemas de Mineração de Dados. Esta metodologia engloba, de forma detalhada, todo processo de preparação dos dados, desde o entendimento do problema até a geração da base.

A viabilidade prática da metodologia proposta, *DMBuilding,* é demonstrada através da realização de um estudo de caso que utiliza uma base de dados de um problema real de larga escala no domínio da análise de risco crédito. Os resultados ilustram os benefícios da metodologia, comprovando sua relevância para a montagem de visão em bases de dados.

Palavras-chave: Descoberta de Conhecimento em Bases de Dados, Preparação dos Dados, Metodologia para o Desenvolvimento de Projetos de Descoberta de Conhecimento em Bases de Dados, Metodologia para Montagem de Visão, Análise de Crédito.

***DMBUILDING:* A NEW METHODOLOGY FOR DATA MART GENERATION IN THE SOLUTION OF DATA MINING PROBLEMS**

Abstract

Technological advances have dramatically increased the volume of data stored in various domains of applications. Since this data abundance has highly exceeded the ability of humans to analyze and use the information available, some valuable hidden information are not understood and employed for efficient decision making processes. This scenario stimulated in the past few years the creation of several techniques for extracting knowledge from large amounts of data. Some of these techniques are the result of the emerging field of Knowledge Discovery in Databases - KDD.

The process of KDD is composed of several steps. The preparation of the data stage takes from 50% to 90% of the time and effort necessary for the completion of the whole process. The more complete and consistent is the preparation, the better the data mining results. One way to ensure the completeness and consistency of the data is to use a methodology that addresses in detail all the activities related to the preparation of the data.

Many methodologies have been proposed for the development of KDD but the majority of them only mention the data preparation process, and even fewer are devoted to the specific creations of data marts which integrate the whole data used to solve the KDD problem. This dissertation studies the methodologies for the development of KDD projects, emphasizing aspects related to the preparation of the data, and as a result, proposes a thorough methodology – *DMBuilding* – for the design of data marts in the context of KDD problems. This methodology includes, in detail, all of the data preparation process, from the understanding of the problem until the generation of the database.

The practical feasibility of the proposed methodology is demonstrated through the completion of a case study that uses a database with a real world complex problem in the field of credit risk analysis. The results illustrate the benefits of the methodology, showing its relevance for the design of data marts addressed to the data mining problems.

Keywords: Knowledge Discovery in Databases, Data Preparation, Methodology for the Development of Projects of Knowledge Discovery in Databases, Methodology for Data Marts Generation in the Solution of Data Mining Problems, Credit Analysis.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

CRISP-DM	*Cross-Industry Standard Process for Data Mining*
DER	*Diagrama de Entidade e Relacionamento*
DMLC	*Data Mining Life Cycle*
DW	*Data Warehouse*
IA	*Inteligência Artificial*
KDD	*Knowledge Discovery in Databases*
MD	*Mineração de Dados*
MLP	*Multi Layer Perceptron*
OLAP	*Online Analytical Processing*
PMBOK	*Project Management Body of Knowledge*
RNA	*Redes Neurais Artificiais*
SGBD	*Sistema Gerenciador de Banco de Dados*
TI	*Tecnologia da Informação*

Capítulo 1

Introdução

1.1 Considerações Iniciais

Atualmente, existem muitas metodologias para o desenvolvimento de projetos de Descoberta de Conhecimento em Base de Dados (*Knowledge Discovery in Databases* - KDD). Apesar de a maioria das metodologias citarem o processo de preparação dos dados, poucas metodologias específicas para montagem de visão de dados têm sido propostas. Diante deste cenário, esta dissertação tem como principal objetivo propor uma metodologia para montagem de visões em bases de dados dirigidas a problemas de Mineração de Dados (MD). Esta metodologia engloba, de forma detalhada, todo processo de Preparação dos Dados, desde o entendimento do problema até a geração da base de dados que será utilizada na fase de MD. Para demonstrar a aplicação prática da metodologia proposta foi desenvolvido um estudo de caso de análise de crédito, um problema real e de larga escala.

1.2 Motivação

No passado, a tecnologia limitada, tanto em relação ao *hardware* como em relação ao *software*, permitia o armazenamento de pequenos volumes de dados. Consultas simples eram suficientes para a análise destes dados. Nas últimas décadas, grandes avanços tecnológicos têm permitido o armazenamento e o processamento de grandes volumes de dados. Conseqüentemente, a análise manual se tornou impraticável, sendo necessária a utilização de métodos eficientes e auxiliados por computador para a análise dos dados e extração de conhecimento. Alguns desses métodos são resultantes do campo de KDD, que incorpora

técnicas utilizadas em áreas como Reconhecimento de Padrões, Inteligência Artificial (IA), Aquisição de Conhecimento, Estatística, Banco de Dados, *Data Warehousing* (DW), Visualização de Dados e outras áreas [Han & Kamber 2006].

Segundo [Fayyad et al. 1996a], KDD é um processo não trivial de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis. Este processo é iterativo (melhorias consecutivas podem ser realizadas a fim de obter melhores resultados) e interativo (várias decisões são tomadas pelos usuários durante o processo), composto das seguintes etapas:

- Seleção de Dados: Obter um conjunto de dados do qual se deseja extrair conhecimento;
- Pré-processamento: Tratar dados inconsistentes e incompletos;
- Transformação: Representar os dados de acordo com o algoritmo de MD;
- Mineração de Dados: Extração de conhecimento da base de dados;
- Interpretação e Avaliação dos Resultados: Interpretação e avaliação dos padrões interessantes.

A metodologia de Fayyad et al. foi a primeira metodologia desenvolvida para projetos de KDD. É uma metodologia útil, porém muito limitada em alguns aspectos, principalmente nas atividades relacionadas ao pré-processamento e transformação de dados.

A realização de um projeto de KDD é uma tarefa muito complexa. Para garantir soluções de qualidade é imprescindível o uso de alguma metodologia que sirva de guia, especificando o fluxo a ser seguido e as atividades que devem ser realizadas.

Muitas metodologias foram propostas para o desenvolvimento de projetos de KDD, entre as quais se destacam:

- Abordagem de Fayyad et al. [Fayyad et al. 1996a]: Considera o processo de KDD como sendo iterativo e interativo, composto das seguintes etapas: Seleção de Dados, Pré-processamento, Transformação, Mineração de Dados e Interpretação e Avaliação dos Resultados;
- CRISP-DM (*Cross-Industry Standard Process for Data Mining*) [Chapman et al. 2000]: Propõe um modelo de processo hierárquico que consiste de seis fases: Entendimento do Negócio, Entendimento dos Dados, Preparação dos Dados, Modelagem, Avaliação e Utilização;
- DMEasy [Cunha 2005]: Aborda uma metodologia genérica, interativa e iterativa para projetos de KDD integrada com um sistema de documentação de processos. É

fortemente baseada em CRISP-DM [Chapman et al. 2000], DMLC (*Data Mining Life Cycle*) [Hofmann & Tierney 2003] e PMBOK (*Project Management Body of Knowledge*) [PMBOK 2000], e está dividida em seis fases: Iniciação, Planejamento e Organização, Execução, Distribuição e Finalização, Acompanhamento e Controle do Projeto, e Processos Genéricos.

As atividades relacionadas à preparação dos dados consomem de 50% a 90% do tempo e esforço necessário para a realização de todo o processo de KDD [Two Crows 1999]. Quanto mais completa e consistente for a preparação, melhor será o resultado da mineração dos dados [Bigus 1996]. Uma forma de garantir a completude e a consistência dos dados é utilizar uma metodologia que aborde de forma detalhada todas as atividades relacionadas à preparação e a transformação dos dados. Por este motivo, algumas metodologias focam na etapa de Preparação dos Dados, entre as quais se destacam:

- Abordagem de Yu et al. [Yu et al. 2006]: Apresenta um esquema integrado de preparação de dados para análise de dados com a utilização de Redes Neurais Artificiais (RNA). Esta metodologia está dividida em três fases: Pré-análise dos Dados, Pré-processamento dos Dados e Pós-análise dos Dados. Cada fase é dividida em sub-processos;
- Metodologia da empresa Quadstone [Quadstone 2003]: Propõe uma metodologia composta de três fases: Captura de Comportamento do Cliente, Análise dos Dados e Modelagem, e Atribuição de *Scores* e Tomada de Decisão.
- Metodologia de Han & Kamber [Han & Kamber 2006]: Propõe uma metodologia focada no pré-processamento de dados e na explanação de técnicas de mineração de dados. Esta metodologia está dividida em quatro fases: Limpeza e Integração, Seleção e Transformação, Mineração de dados e Avaliação e Apresentação. Cada fase é dividida em sub-processos.

Apesar de a maioria das metodologias citadas fazerem referência a atividades relacionadas à preparação dos dados, poucas metodologias específicas para montagem de visão de dados têm sido propostas. A falta de metodologias específicas para a fase de Montagem de Visões de dados é o foco da motivação dessa dissertação.

1.3 Objetivos

Motivada pela importância das atividades relacionadas à preparação dos dados em projetos de KDD e pela escassez de metodologias que abordem estas atividades de forma detalhada, esta dissertação tem como objetivos:

- Investigar as metodologias para o desenvolvimento de projetos de KDD, enfatizando os aspectos relacionados à preparação dos dados;
- Como resultado da investigação, propor uma metodologia para montagem de visões em bases de dados dirigidas a problemas de MD, com foco em dados cadastrais e comportamentais;
- Aplicar a metodologia proposta em um problema de larga escala no domínio de análise de risco de crédito, para análise e validação da metodologia proposta.

1.4 Justificativas

Apesar de KDD ser uma área recente com muitos aspectos que ainda precisam ser pesquisados em profundidade, várias metodologias para o desenvolvimento de projetos de KDD já foram propostas. A maioria dessas metodologias visa à definição das etapas do processo de KDD, especificando a seqüência de execução e a interação entre elas. Nenhuma etapa é completamente explorada, pois o foco é descrever o processo como um todo. Com o crescimento da aplicação de KDD em diversos domínios, percebeu-se que as atividades relacionadas à preparação dos dados consomem de 50% a 90% do tempo e esforço necessário para a realização de todo o processo de KDD [Two Crows 1999]. Portanto, quanto mais completa e consistente for a preparação, melhor será o resultado da mineração de dados [Bigus 1996]. Apesar de a maioria das metodologias citarem o processo de preparação dos dados, poucas metodologias específicas para montagem de visão de dados têm sido propostas. Portanto, a principal contribuição dessa dissertação é a proposição de uma metodologia específica para montagem de visões em bases de dados dirigidas a problemas de MD. Esta metodologia engloba, de forma detalhada, todo processo de preparação dos dados, desde o entendimento do problema até a geração da base de dados. Para cada fase são especificados os processos, as atividades, as entradas, as saídas e os responsáveis. Este nível de detalhe permite que a metodologia seja utilizada como guia completo para montagem de visões em bases de dados.

A metodologia proposta nesta dissertação é fortemente baseada nas seguintes metodologias: abordagem de [Fayyad et al. 1996a], CRISP-DM [Chapman et al. 2000], DMEasy [Cunha, 2005], abordagem de [Yu et al. 2006] e metodologia da empresa Quadstone [Quadstone 2003]. As três primeiras metodologias são voltadas para o desenvolvimento de projetos de KDD, englobando todas as etapas do processo. As demais são mais focadas no processo de preparação dos dados.

A abordagem de [Fayyad et al. 1996a] e a CRISP-DM [Chapman et al. 2000] foram escolhidas por serem referências relevantes na área de KDD. A abordagem de [Fayyad et al. 1996a] foi um marco para a área de MD, pois introduziu o conceito de MD em um contexto mais amplo (KDD), focando muito mais na solução do problema do que nos resultados das técnicas de MD [Cunha 2005]. A metodologia CRISP-DM [Chapman et al. 2000] foi concebida por um consórcio de empresas interessadas em MD que decidiram criar um modelo de processo padrão, não proprietário e disponível para todos, que fosse baseado em experiências reais.

As principais motivações para investigar a metodologia DMEasy [Cunha 2005] são: possibilidade de uma especificação mais detalhada do processo de KDD em relação as outras metodologias, suporte à documentação do processo e ênfase na especificação do negócio associado ao problema. Esta metodologia foi proposta com o intuito de ser uma metodologia geral, completa e que atenda à realidade dos desenvolvedores de soluções de KDD.

A abordagem de [Yu et al. 2006] e a metodologia da empresa Quadstone [Quadstone 2003] também foram consideradas como base para a metodologia proposta por apresentarem como foco o processo de preparação dos dados, característica ausente nas demais metodologias. A abordagem de [Yu et al. 2006] especifica os problemas relacionados aos dados e apresenta uma série de técnicas de processamento que podem ser aplicadas para resolução dos mesmos. A metodologia da empresa Quadstone [Quadstone 2003] está mais focada na modelagem de variáveis comportamentais que são de extrema importância para resolução de problemas de KDD.

A metodologia proposta nesta dissertação foi aplicada em um problema de larga escala no domínio de análise de crédito. Este domínio foi escolhido por se tratar de um problema de larga complexidade, envolver um grande número de informações a serem analisadas, com múltiplos atributos relacionados, e ser de interesse de diversas instituições e empresas. Tais características permitiram a verificação da viabilidade prática da metodologia e a avaliação das suas virtudes e deficiências.

1.5 Estrutura da Dissertação

Esta dissertação está organizada em seis capítulos. O Capítulo 1 apresenta uma visão geral da área de KDD e das metodologias para o desenvolvimento de projetos de KDD. Este capítulo também expõe as motivações, os objetivos e as justificativas desta dissertação. O Capítulo 2 descreve detalhadamente todas as etapas do processo de KDD e destaca algumas áreas de aplicação. O Capítulo 3 aborda as principais metodologias para o desenvolvimento de projetos de KDD e para montagem de visão de dados, enfatizando os aspectos relacionados à preparação dos dados. O Capítulo 4 descreve de forma bastante detalhada a metodologia proposta nesta dissertação para montagem de visões em bases de dados dirigidas a problemas de MD. O Capítulo 5 descreve a aplicação da metodologia em um problema de análise de crédito. Por fim, o Capítulo 6 apresenta as conclusões, as contribuições, as limitações e os trabalhos futuros relacionados à dissertação.

Capítulo 2

Descoberta de Conhecimento em Bases de Dados

2.1 Introdução

O crescente avanço da tecnologia e o uso extensivo de Sistemas Gerenciadores de Banco de Dados (SGBD) proporcionam, gradativamente, mais recursos computacionais para a coleta, armazenamento, processamento e distribuição dos dados [Kimball et al. 1998].

A geração de informações corretas a partir dos dados armazenados é de essencial importância para qualquer organização. [Almeida 1995] afirma que ao ter acesso à informação uma empresa é capaz de aumentar o valor agregado de seu produto ou reduzir seus custos. Para [Freitas 1993], as informações e o conhecimento compõem um recurso estratégico essencial para o sucesso da empresa em um ambiente competitivo. Segundo [Oliveira 1999], toda empresa tem informações que proporcionam a sustentação para as suas decisões. Entretanto apenas algumas empresas conseguem otimizar o seu processo decisório e aquelas que estão neste estágio evolutivo seguramente possuem grande vantagem.

Apesar de as organizações reconhecerem a importância das informações "embutidas" em suas bases de dados, devido ao grande volume de dados, percebeu-se que era impraticável a realização de análise puramente humana [Romão 2002].

Data Warehousing (DW) foi um dos primeiros passos a fim de aprimorar a análise de grandes quantidades de dados [Kimball et al. 1998]. O *Data Warehouse* (DW) é o resultado de um processo de preparação de dados para exploração analítica. Este termo é utilizado para descrever um conjunto muito grande de dados orientados a assuntos, integrados e não voláteis, que suportam as tomadas de decisão [Srivastava & Chen 1999]. O desenvolvimento e a implantação de um DW envolvem a integração de dados de diversas fontes, realizando transformações sobre os mesmos a fim de obter dados limpos, agregados e consolidados. O

principal resultado esperado da utilização de um DW é um ambiente de consultas analíticas, onde as informações podem ser exploradas (analisadas) segundo algumas dimensões. Estas consultas, chamadas OLAP (*Online Analytical Processing*). São orientadas pelo usuário, o qual possui uma hipótese que queira validar ou simplesmente executa consultas aleatórias [Rezende et al. 2003]. Isso faz com que padrões escondidos nos dados não sejam descobertos de forma "inteligente", visto que é exigida do usuário a capacidade de pensar em todas as relações e combinar diversas hipóteses a fim de extrair conhecimento em uma grande massa de dados.

Neste contexto, surgiu a área de Descoberta de Conhecimento em Base de Dados (*Knowledge Discovery in Databases - KDD*) [Fayyad et al. 1996a]. A área de KDD pode ser entendida como a exploração e a análise de grandes quantidades de dados, de maneira automática ou semi-automática, com o objetivo de descobrir padrões e regras relevantes [Berry & Linoff 2004], que podem ser usados no processo de tomada de decisão.

Este capítulo trata do processo de KDD, enfatizando a preparação e o pré-processamento dos dados, foco dessa dissertação.

2.2 O Processo de Descoberta de Conhecimento

O termo KDD surgiu em um *workshop*, em 1989, para enfatizar o produto final desse processo: o "conhecimento" [Fayyad et al. 1996a].

KDD é um processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis ao ser humano [Fayyad et al. 1996a]. Portanto, o conhecimento obtido deve ser considerado interessante, auxiliando no processo de tomada de decisão.

Todo o processo de KDD é orientado ao domínio de aplicação e aos repositórios de dados. Para usar os dados é necessário que os mesmos estejam estruturados de forma a serem consultados e analisados adequadamente [Rezende et al. 2003].

A aplicação de um projeto de KDD só é viável quando alguns aspectos são considerados. Um aspecto prático é o impacto que a aplicação irá provocar. Esse impacto pode ser medido através de critérios como rendimento, redução de custos, melhora na qualidade dos produtos, economia de tempo, qualidade do conhecimento descoberto, aumento da automação de processos de análises manuais [Romão 2002]. Outros aspectos são a disponibilidade e a qualidade dos dados.

Muitas vezes os termos "Mineração de dados" (MD) e "Descoberta de Conhecimento em Banco de dados" são confundidos como sinônimos. Porém, o termo KDD é empregado para descrever todo o processo de extração de conhecimento de um conjunto de dados e o termo MD refere-se a uma das etapas deste processo.

A MD é uma área interdisciplinar que resultou da junção de áreas como Estatística, Banco de Dados, Arquiteturas Paralelas, Visualização de Dados, Aprendizado de Máquina e Computação de Alto Desempenho, conforme pode ser visto na Figura 2.1 [Motta 2007].

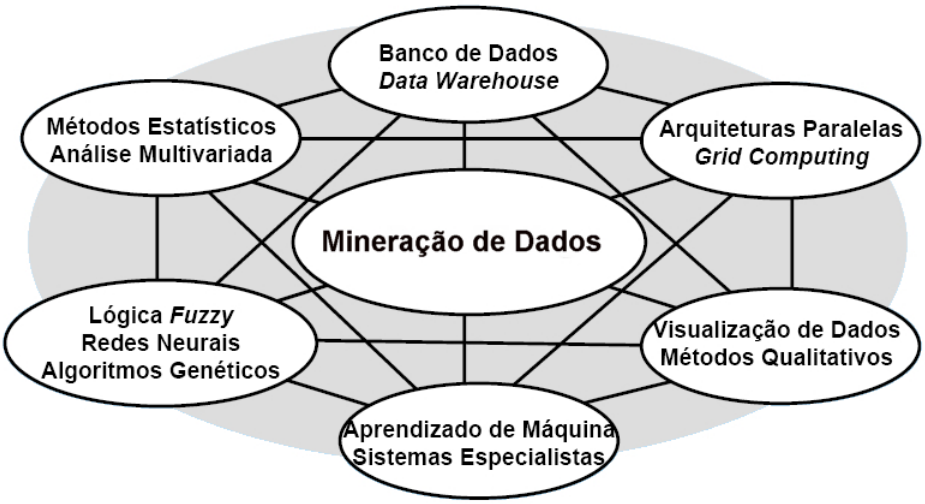


Figura 2.1 – Interdisciplinaridade da área de Mineração de Dados

No desenvolvimento de um projeto de KDD, muitos são os envolvidos. Cada atividade está diretamente ligada a um ou mais responsáveis. Não existe um padrão de definição de papéis, mas os papéis comumente citados nas metodologias de KDD são [Myatt 2007]:

- **Analista de dados**: entende as técnicas envolvidas no processo de KDD. Tem conhecimento sobre o funcionamento dos algoritmos e das ferramentas utilizadas no processo, mas não necessariamente conhece o domínio ao qual os dados pertencem;
- **Especialista no domínio:** conhece o domínio no qual o processo de KDD será aplicado. Por exemplo, pode-se utilizar KDD para encontrar padrões de vendas de produtos e, nesse caso, o especialista no domínio pode ser um especialista em *marketing* ou alguém que tenha maior familiaridade com relação ao domínio do problema a ser resolvido;

- **Usuário:** irá utilizar o resultado do processo de KDD. Normalmente, o usuário não é somente uma pessoa, mas uma instituição, uma empresa ou um departamento de uma empresa.
- **Especialista em TI (Tecnologia da Informação):** detém todo o conhecimento do funcionamento da área de informática da organização na qual será aplicado o resultado do processo de KDD. Este conhecimento é necessário para o levantamento e o entendimento dos dados que serão necessários para a realização do projeto.

Esses responsáveis possuem habilidades diferentes, mas não são necessariamente pessoas diferentes. Por exemplo, freqüentemente os papéis de usuário e especialista são exercidos por uma mesma pessoa, quando esta possui conhecimento detalhado do domínio da aplicação.

Várias metodologias de KDD foram propostas. Por ser uma das metodologias que descreve as abordagens comumente usadas pelos usuários de mineração de dados utilizadas para o desenvolvimento de projetos, a descrição do processo de KDD exposta nas próximas seções é baseada na metodologia CRISP-DM [Chapman et al. 2000]. Segundo esta metodologia, um projeto de KDD consiste de seis fases (entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e utilização), conforme descrito na Figura 2.2.

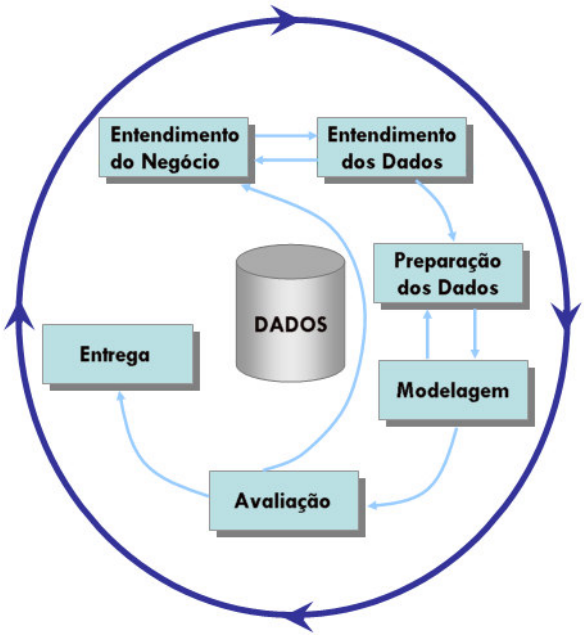


Figura 2.2 – Etapas do processo de KDD segundo a metodologia CRISP-DM [Chapman et al. 2000]

2.2.1 Entendimento do Negócio

Nesta primeira etapa, é realizado um estudo para analisar o problema a ser resolvido e se esse problema deve ser solucionado através de algum método de MD. Nessa fase todas as documentações necessárias serão realizadas para auxiliar nas demais fases do projeto.

O primeiro passo a ser realizado é um estudo do domínio da aplicação e a definição de objetivos e metas a serem alcançados, a fim de planejar as atividades necessárias para que o projeto seja finalizado dentro do prazo, custo e qualidade desejáveis [Cunha 2005].

Essa etapa é essencial porque irá criar um plano cuja execução será focada no problema a ser resolvido [Chapman et al. 2000].

Ao final dessa etapa deve-se ter:

- Uma definição clara do problema a ser resolvido;
- Um entendimento completo dos dados relevantes ao problema;
- Uma visão de como os resultados do projeto serão utilizados.

É fundamental que o especialista do domínio da aplicação faça parte dessa etapa, fornecendo conhecimento sobre o domínio. Esse conhecimento serve de subsídio para todas as outras fases do projeto de KDD, principalmente para a fase de pré-processamento dos dados, onde um conhecimento mais profundo dos dados contribui com a escolha do conjunto de dados mais relevante para o problema. O especialista do domínio colabora na definição do que possam ser padrões válidos, do grau de importância dos atributos, de relacionamento entre bases e informações para a geração de novos atributos, de informações sobre os melhores parâmetros do processo de mineração de dados, entre outras atividades [Rezende et al. 2003].

O processo de KDD difere de outros processos analíticos, pois, apesar da aparente facilidade de aplicar algoritmos de MD e obter algum resultado, sem uma definição clara do problema a ser resolvido, os resultados obtidos serão inúteis [John 1997].

2.2.2 Entendimento dos Dados

Uma vez que o problema foi definido, dados relevantes devem ser coletados. Portanto, é necessário verificar se os dados disponíveis são relevantes para a aplicação de algum algoritmo de MD e analisar se os mesmos são úteis para o problema [Alluri 2005].

O conhecimento sobre os dados irá contribuir para algumas etapas posteriores, como a fase de transformação dos dados [Chapman et al. 2000].

Na maioria dos casos, dados relevantes são extraídos de uma ou mais bases de dados ou de um DW que foi originalmente criado para realização de consultas analíticas [Baragoin et al. 2001] [Inmon 2005].

O entendimento dos dados que serão utilizados durante o processo de KDD pode ser documentado em um relatório com a descrição dos dados [Two Crows 1999]. Esse relatório deve conter propriedades como formato dos dados, número de exemplos e atributos; total de dados ausentes e diferentes; nome, domínio, descrição, valores máximo e mínimo dos atributos; e descrição das fontes de dados.

É de suma importância que esta fase seja executada utilizando-se todo o conhecimento do especialista do domínio e do especialista em TI. Além disso, é de extrema utilidade que exista uma descrição formal dos dados (geralmente em forma de Dicionário de Dados) e uma caracterização de como os dados estão agrupados e relacionados (no caso de banco de dados relacionais, através do DER - Diagrama de Entidade e Relacionamento [Ramakrishnan & Gehrke 2002]). No entanto, muitas instituições não possuem este tipo de documentação, e é nesse momento que os especialistas são chave fundamental para o entendimento dos dados.

Ao fim, devem ser analisadas a qualidade dos dados e a viabilidade do projeto diante dos dados disponíveis. Para aumentar a qualidade dos dados já disponíveis na organização, fontes externas (referentes ao negócio em questão) podem ser adquiridas. Caso a qualidade dos dados seja um fator conflitante, o projeto pode ser descontinuado.

2.2.3 Preparação dos Dados / Pré-processamento

Depois de entender o problema, levantar os objetivos e avaliar os dados disponíveis, chega-se ao processo de mapeamento do problema, que tem por objetivo planejar as tarefas de MD a serem aplicadas e construir a visão final dos dados. Entende-se por visão de dados a caracterização final dos dados a serem apresentados aos algoritmos de MD [Berry & Linoff 2004]. Esta fase engloba toda transformação realizada na base de dados, como mudança de granularidade e outras transformações que serão abordadas pela metodologia proposta nessa dissertação.

Normalmente, os dados disponíveis para análise não estão em um formato adequado para a extração de conhecimento e apresentam diversos problemas, como grande quantidade de valores desconhecidos, formatos diferentes para o mesmo atributo em diversas tabelas,

ruído (atributos com valores incorretos), entre outros. Além disso, em razão de limitações de memória ou tempo de processamento, muitas vezes não é possível à aplicação direta dos algoritmos de MD [Deschaine et al. 2001]. Dessa maneira, torna-se necessária a aplicação de métodos para tratamento, limpeza e redução do volume de dados antes de iniciar a etapa de mineração de dados. Tais métodos permitem que as informações escondidas nos dados se tornem mais acessíveis e disponíveis para as ferramentas de MD [Pyle 1999]. É importante salientar que a execução das transformações deve ser guiada pelos objetivos do projeto a fim de que o conjunto de dados gerado apresente as características necessárias para que os objetivos sejam cumpridos [Rezende et al. 2003].

Em um projeto de KDD, todas as fases são importantes, mas normalmente a fase de preparação dos dados é a mais relevante e crítica. Quanto mais elaborada for esta fase, maior a tendência de se produzir bons resultados. A fase de preparação dos dados tende a consumir de 50 a 90% do tempo de um projeto de KDD [Jermyn et al. 1999] [Two Crows 1999].

De forma geral, a preparação dos dados é um processo semi-automático, pois depende da capacidade do analista de dados em identificar os problemas presentes nos dados e os métodos apropriados para solucioná-los.

As tarefas realizadas nessa etapa podem ser divididas em dois grupos [Batista 2003]:

- **Tarefas fortemente dependentes do conhecimento do domínio:** Essas tarefas podem ser efetivamente realizadas apenas com o uso do conhecimento específico do domínio. Um método automático pode ser empregado para realizar este tipo de tarefa, entretanto, esse método depende de que um conhecimento específico seja fornecido. Um exemplo disso são as verificações de integridade dos dados. Por exemplo, em uma aplicação de transações bancárias onde as variáveis "*agência*" e "*conta*" não podem assumir valores nulos, através de um conjunto de regras dependentes de domínio é possível verificar a integridade dessas variáveis;
- **Tarefas fracamente dependentes do conhecimento do domínio:** Essas tarefas podem ser realizadas por métodos que extraem dos próprios dados as informações necessárias para tratar o problema de pré-processamento dos dados. Apesar de essas tarefas dependerem do conhecimento do domínio para selecionar o método correto para tratar o problema, elas podem ser realizadas por métodos que são mais automáticos do que aqueles utilizados em tarefas que dependem fortemente do conhecimento de domínio. O tratamento de valores desconhecidos e a identificação de valores extremos são exemplos de tarefas fracamente dependentes do conhecimento do domínio.

A fase de preparação de dados consiste em várias atividades: Extração e Integração dos Dados, Seleção dos Dados, Limpeza dos Dados e Transformação dos Dados.

2.2.3.1 Extração e Integração dos Dados

A qualidade dos dados é o principal fator de influência para a qualidade dos resultados de um projeto de KDD. Os dados devem ser confiáveis e representar fielmente o problema que se pretende resolver [Pyle 2003].

Apesar de parecer simples, a fase de extração e integração dos dados é uma fase custosa que demanda conhecimento profundo dos dados disponíveis e requeridos [Two Crows 1999].

Durante a extração dos dados pode haver a necessidade de junções de tabelas de diversas fontes, filtros, conversão de arquivos de textos para tabelas relacionais, entre outros. Além disso, dados provenientes de diversas partes de uma mesma organização podem variar com relação à qualidade e ao intervalo de atualizações [Adriaans & Zantinge 1996].

Unir dados de diversas fontes é uma situação comum em qualquer projeto de KDD. Ao unir esses dados, todo cuidado é pouco, pois erros são freqüentes nessa junção.

Os problemas encontrados nessa fase podem ser minimizados caso a organização já possua seus dados em um DW, pois todo o trabalho de extração, união e limpeza dos dados já foi realizado no momento da construção do DW. Além disso, o DW é periodicamente atualizado com dados de sistemas transacionais e/ou fontes externas. Essa tecnologia tem sido largamente utilizada, uma vez que os bancos de dados transacionais não são considerados adequados para fornecer repostas para análises estratégicas [Imnon 2005].

2.2.3.2 Seleção dos Dados

A seleção dos dados ocorre em duas dimensões: representatividade dos exemplos e relevância dos atributos [Monteiro 1999].

A redução do número de exemplos deve ser executada mantendo as características do conjunto de dados original. Isso pode ser feito através de métodos estatísticos, como a amostragem aleatória, que tendem a produzir amostras representativas dos dados [Rezende et al. 2003]. Se a amostra não for representativa ou a quantidade de exemplos for insuficiente para caracterizar os padrões escondidos nos dados, os modelos obtidos não representarão a

realidade ou poderão apresentar *overfitting* (ajuste excessivo do modelo ao conjunto de treinamento, afetando a generalização) [Fayyad et al. 1996b]

A grande maioria dos algoritmos de MD são desenvolvidos para aprender quais são os atributos apropriados para a aprendizagem. Ter muitos atributos, na teoria, deveria resultar em um aumento do poder discriminante, mas, na prática, a adição de atributos irrelevantes ou redundantes geralmente "confundem" os algoritmos [Witten & Frank 2005].

Por causa do efeito negativo dos atributos irrelevantes ou redundantes na maioria dos algoritmos de MD, é comum realizar a fase seleção de atributos. Uma das formas amplamente utilizadas para selecionar os atributos relevantes é a forma manual, baseada no profundo entendimento do problema e do significado real dos atributos [Berry & Linoff 2004].

Além da forma manual que tende a ser difícil e demandar tempo, métodos automáticos também podem ser úteis. A idéia geral é aplicar alguma medida que é usada para quantificar a relevância do atributo de acordo com uma classe ou conceito. Alguns exemplos dessas medidas são: ganho de informação, índice de Gini e coeficiente correlação [Han & Kamber 2006].

Algumas variáveis podem ser supérfluas, apresentando a mesma informação do que outras variáveis, fazendo assim com que o tempo de execução do modelo seja maior [Baragoin et al. 2001]. Para tanto, testes estatísticos podem ser realizados para saber o quão relacionadas estão as variáveis (como por exemplo, análises estatísticas bi-variadas, regressão linear e polinominal) [Han & Kamber 2006]. Variáveis correlacionadas devem ser reduzidas através da escolha de uma das variáveis ou através da composição de uma nova variável a partir daquelas correlacionadas, utilizando técnicas como análise fatorial ou análises de componentes [Baragoin et al. 2001].

Segundo [Han & Kamber 2006], uma árvore de decisão gerada a partir de um conjunto de dados também pode ser utilizada na seleção de atributos. Todos os atributos que não aparecerem na árvore são tidos como irrelevantes e podem ser eliminados.

Outros critérios para a exclusão de variáveis incluem problemas na aquisição dos dados no momento da utilização do modelo construído, custo para a aquisição, restrição do uso e problemas de qualidade [Two Crows 1999].

A seleção de atributos proporciona uma representação de dados mais compacta e uma melhor representatividade do conceito do alvo, focando a atenção do usuário nas variáveis relevantes. Além disso, otimiza o tempo de processamento, visto que o algoritmo de MD irá trabalhar com um subconjunto dos atributos, diminuindo o seu espaço de busca [Witten & Frank 2005].

2.2.3.3 Limpeza dos Dados

Problemas como erro de digitação, erro na captação e formatação incorreta realizada pelos sistemas de informação, normalmente geram dados incompletos, com ruídos e inconsistentes. Portanto, após a seleção dos exemplos e dos atributos relevantes, é necessário realizar a limpeza dos dados [Dasu & Johnson 2003]. Nessa fase são verificadas características que podem afetar a qualidade dos dados e, como conseqüência, a qualidade do modelo.

Um dos problemas mais comuns é o de valores ausentes (*missing values*). Segundo [Han & Kamber 2006] [Two Crows 1999], várias técnicas podem ser utilizadas para a resolução desse problema, tais como: ignorar o atributo (quando a grande maioria dos valores do atributo são ausentes), substituição por um valor constante, substituição pela média (atributos contínuos), substituição pela moda (atributos nominais), substituição pela mediana (atributos ordinais) ou substituição pelo valor mais provável (através do uso de técnicas de regressão ou de árvores de decisão). O método de previsão tende a gerar melhores resultados, no entanto, requer muito mais tempo.

Outro problema encontrado nos dados é o ruído (*outlier*). O ruído é a ocorrência de um valor único ou de baixa freqüência que está longe da média da maioria dos valores presentes na variável [Dasu & Johnson 2003]. A definição de valores fora da média está totalmente relacionada com o conhecimento do domínio dos dados [Rud 2001]. O problema de *outliers* pode ser resolvido através da substituição pela média, substituição de acordo com uma distribuição específica ou remoção do exemplo [Paul et al. 2003].

Técnicas como *clustering* podem ser utilizadas na identificação de *outliers* [Duda et al. 2002] [Theodoridis & Koutroumbas 1999]. Na técnica de *clustering* valores similares são organizados em grupos (*clusters*) e os valores isolados são considerados *outliers* [Han & Kamber 2006].

Além de eliminar o ruído dos dados, é interessante verificar a fonte desses problemas e então corrigi-los, para que no futuro esses erros deixem de existir.

2.2.3.4 Transformação dos Dados

O principal objetivo dessa etapa é transformar a representação dos dados de acordo com o algoritmo de MD, a fim de superar quaisquer limitações existentes no algoritmo [Batista 2003]. É importante salientar que a execução das transformações deve ser guiada pelos objetivos do projeto a fim de que o conjunto de dados gerado apresente as características

necessárias para que os objetivos sejam alcançados. Além disso, é essencial ter conhecimento profundo do domínio do problema para que as transformações sejam realizadas de maneira adequada.

Existe uma série de técnicas de transformação de dados, como por exemplo:

- **Normalização:** consiste em transformar a escala dos valores do atributo, alterando seus intervalos originais para um intervalo específico, como por exemplo de -1,0 a 1,0 ou de 0,0 a 1,0 [Han & Kamber 2006]. Essa transformação é importante para métodos que calculam distâncias entre atributos (como *K-Nearest Neighbours*, *clustering* e RNA). Os principais tipos de normalização são: *Min-Max* (realiza uma transformação linear nos dados), *z-score* (normaliza os valores baseando-se na média de um conjunto de dados) e *decimal scaling* (normaliza através da mudança do ponto decimal). Esses tipos podem alterar os dados originais [Han & Kamber 2006];
- **Discretização:** converte um dado contínuo em um dado discreto, dividindo a amplitude do atributo em intervalos. Esta técnica é útil para algoritmos de classificação e de clustering que utilizam apenas atributos nominais [Witten & Frank 2005]. É recomendável converter os dados em categorias que reflitam a verdadeira variação dos dados [Myatt 2007]. Substituindo inúmeros valores de um atributo contínuo por uma pequena gama de rótulos representativos dos intervalos reduz e simplifica os dados originais. Isto leva a uma representação de resultados de mineração concisa e de fácil utilização [Han & Kamber 2006];
- **Agregação:** operações de agregação são aplicadas aos dados. Pode ser utilizada para a criação de atributos através de derivações de atributos existentes na base de dados. Qualquer operação matemática, como média ou soma, pode ser aplicada a uma ou mais variáveis para geração de novas variáveis [Cabena et al. 1997];
- **Suavização:** tem como objetivo diminuir o número de valores de um atributo sem que seja necessário discretizá-lo. Os valores de um determinado atributo são agrupados e cada grupo é substituído por um valor numérico que o represente. Esse novo valor pode ser a média ou a mediana do grupo [Rezende et al. 2003]. Esta técnica tende a remover o ruído dos dados. Técnicas de *clustering* e regressão podem ser usadas para suavização [Han & Kamber 2006];
- **Generalização:** tem como objetivo substituir os dados de baixo nível (ou primitivos) por dados de alto nível através do uso de conceitos de hierarquia. Por

exemplo, atributos categóricos, como *rua*, podem ser generalizados para conceitos de alto nível, como *cidade* ou *país*, e atributos numéricos, como *idade*, podem ser mapeados para um conceito de alto nível como *jovem, meia-idade* e *idoso* [Monteiro 1999];

- **Padronização:** Em variáveis de escala nominal ou ordinal é bastante comum a apresentação de erros tipográficos, como a escrita de um determinado valor de diversas formas. Por exemplo, a variável "*cidade*" pode conter vários registros significando um mesmo valor (como por exemplo: São Paulo, S. Paulo, S.P., SP). É importante que uma lista de possíveis valores de cada atributo seja analisada a fim de realizar uma padronização dos valores com o mesmo conceito [Witten & Frank 2005] [Myatt 2007].

2.2.4 Modelagem

Após a preparação dos dados, inicia-se a fase onde algum algoritmo de MD é aplicado para resolução do problema proposto. Além da escolha do algoritmo de MD, essa fase define a melhor configuração para o mesmo. Como esta etapa é iterativa, pode ser necessário que a mesma seja executada diversas vezes para ajustar o conjunto de parâmetros a fim de obter resultados mais adequados aos objetivos pré-estabelecidos [Rezende et al. 2003].

Os tipos de padrões que podem ser descobertos dependem da técnica de MD que será aplicada [Two Crows 1999]. A maioria das técnicas usadas atualmente pelas ferramentas de mineração de dados são provenientes da área de Inteligência Artificial (IA) [Curoto 2003].

Existem dois tipos de tarefas de MD: tarefas descritivas e as tarefas preditivas [Zaïne 1999]. As tarefas descritivas consistem no descobrimento de informações intrínsecas ao conjunto de dados. A informação descoberta é representada numa forma compreensível ao ser humano [Fayyad et al. 1996a] [Rezende et al. 2003]. Regras de Associação, Sumarização e *Clustering* são exemplos de tarefas descritivas. As tarefas preditivas consistem em fazer previsões baseadas em inferências dos dados disponíveis [Amorim 2004]. Os principais tipos de tarefas preditivas são classificação e previsão.

Essa seção irá abordas três das técnicas citadas acima: *Clustering*, Classificação e Previsão.

2.2.4.1 *Clustering*

Clustering (ou Segmentação) consiste em segmentar uma população heterogênea em número finito de subgrupos mais homogêneos chamados de *clusters* [Duda et al. 2002]. A divisão em grupos é importante para [Myatt 2007]:

- **Encontrar relações escondidas:** Ao analisar os subgrupos homogêneos, é possível encontrar relações que anteriormente não eram óbvias [Duda et al. 2002];
- **Segmentações:** Técnicas de agrupamento geram divisões que simplificam os dados para análises. Estes subgrupos podem ser usados para criar outros modelos mais simplificados, baseados especificamente nas características de cada grupo [Theodoridis & Koutroumbas 1999].

Como os registros não possuem rótulos, os mesmos são agrupados de acordo com a similaridade entre eles (e a dissimiralidade entre outros grupos). Cabe ao algoritmo de *clustering* descobrir as classes, que podem ser mutuamente exclusivas ou sobrepostas [Zaïne 1999]. Ao usuário, cabe definir o significado de cada classe descoberta. A medição da qualidade do esquema de classificação produzido não é trivial [Fayyad et al. 1996a].

O uso de técnicas de *clustering* tem como vantagens a flexibilidade (dependendo da configuração do algoritmo, *clusters* diferentes são obtidos) e a possibilidade de utilizar uma abordagem hierárquica em que os *clusters* podem ser agrupados hierarquicamente [Myatt 2007]. Como desvantagens destacam-se: a interpretação dos grupos formados requer análises adicionais realizadas por especialista do domínio, a geração de grupos em grandes bases de dados geralmente tende a ser demorada e a limitação do número de observações que podem ser processadas [Duda et al. 2002].

2.2.4.2 Classificação

Classificação é uma função de aprendizagem que mapeia (classifica) um determinado exemplo em uma das classes pré-definidas [Fayyad et al. 1996a]. É uma das técnicas mais utilizadas.

A classificação prevê o valor de um atributo nominal ou categórico chamado de "classe" [John 1997]. Por exemplo, dada uma base de dados com informações de clientes, um modelo de classificação pode ser construído para identificar os clientes BONS e MAUS.

Um modelo é construído através da análise de registros da base de dados descritos por atributos. Cada registro pertence a uma classe pré-definida, determinada por um atributo chamado de classe. O algoritmo de classificação aprende a partir da base de treinamento, que contém exemplos com valores para as variáveis de entrada (descritivas) e para a variável de classe. Portando, a base de treinamento é utilizada para determinar e quantificar as relações entre as variáveis descritivas e a classe. Se o algoritmo foi desenvolvido de forma eficiente, ele será usado de forma preditiva para classificar novos registros nas classes pré-definidas [Zaïne 1999].

Como a classe de cada exemplo de treinamento é fornecida, essa forma de aprendizagem é chamada de supervisionada [Rezende et al. 2003]. Esse tipo de aprendizagem contrasta com a aprendizagem não supervisionada dos algoritmos de *clustering*, onde não são conhecidos as classes dos exemplos de treinamento e o número de classes a ser aprendida [Han & Kamber 2006].

Normalmente, os modelos são representados em forma de regras de classificação, árvores de decisão ou fórmulas matemáticas. Dentre as técnicas mais utilizadas, destacam-se RNA, Classificadores *Bayesianos* e Árvores de Decisão [Carvalho 1999] [Michie et al. 1994].

2.2.4.3 Previsão

Previsão (ou regressão) se assemelha com a classificação, pois é uma função de aprendizagem que mapeia um determinado exemplo em uma variável de previsão [Fayyad et al. 1996a]. Na classificação, a variável de previsão é categórica. Na regressão, a variável de previsão é contínua. A previsão pode ser usada para executar uma tarefa de classificação, convencionando-se que diferentes faixas (intervalos) de valores contínuos correspondem a diferentes classes [Dias 2001].

Uma das razões de tratar a previsão e a classificação como tarefas separadas é que em um modelo de previsão existe um tratamento diferenciado de relações temporais entre as variáveis descritivas e a variável de previsão [Berry & Linoff 2004].

Modelos de previsão são usados em situações onde são requeridas estimativas ou prognósticos, como um projeto de vendas ou a previsão do tempo. Um modelo de previsão calcula uma estimativa para uma ou mais variáveis (resposta), baseando-se em outras variáveis (descritivas) [Han & Kamber 2006].

A maioria das técnicas utilizadas para resolver problemas de classificação também podem ser usadas para resolver problemas de previsão.

2.2.5 Avaliação

Após a utilização de técnicas de MD, é necessário aplicar alguma metodologia para estimar o desempenho do modelo construído e compreender suas vantagens e limitações.

Para estimar a precisão de um modelo normalmente utilizam-se conjuntos disjuntos de treinamento e teste. Geralmente, quanto maior o conjunto de treinamento, melhor o indutor e, quanto maior o conjunto de teste, mais precisa a estimativa do desempenho [Witten & Frank 2000].

A qualidade de um modelo de classificação pode ser medida através da contagem de exemplos classificados de forma correta e incorreta. Essa contagem normalmente é representada através de uma matriz de confusão. Esta é uma ferramenta bastante útil para interpretar os resultados de problemas de classificação, pois além de mostrar a precisão global do modelo, especifica detalhadamente os tipos de erro, permitindo uma verificação mais abrangente de sua precisão [Two Crows 1999].

Os resultados de uma matriz de confusão são totalizados em duas dimensões (classes verdadeiras e classes preditas), como mostra a Tabela 2.1 [Myatt 2005].

Tabela 2.1 – Matriz de confusão

		Classe predita	
	Verdadeira/Predita	Verdadeiro (1)	Falso (0)
Classe verdadeira	Verdadeiro (1)	Contador11	Contador01
	Falso (0)	Contador10	Contador00

Onde:

- Contador11: o número de observações que são verdadeiras e foram classificadas como verdadeiras (verdadeiro positivo);
- Contador10: o número de observações que são falsas e foram classificadas como verdadeiras (falso negativo);
- Contador01: o número de observações que são verdadeiras e foram classificadas como falsas (falso positivo);
- Contador00: o número de observações que são falsas e foram classificadas como falsas (verdadeiro negativo).

O modelo ideal seria aquele que não existiriam ocorrências do Contador10 e Contador01, o que é praticamente impossível. De qualquer maneira o objetivo é minimizar o número de falso negativos e falso positivos.

Existem quatro fórmulas comumente utilizadas para verificar a qualidade de um modelo [Myatt 2007]:

- **Concordância:** mede a exatidão de um modelo e é calculada através da Equação 2.1:

$$\frac{(\text{Contador11} + \text{Contador 00})}{(\text{Contador11} + \text{Contador10} + \text{Contador01} + \text{Contador00})} \quad (2.1)$$

- **Taxa de Erro:** representa o número de predições erradas e é calculada através da Equação 2.2:

$$\frac{(\text{Contador10} + \text{Contador 01})}{(\text{Contador11} + \text{Contador10} + \text{Contador01} + \text{Contador00})} \quad (2.2)$$

- **Sensibilidade:** calcula quão apto o modelo está para classificar valores positivos e calculada através da Equação 2.3:

$$\frac{\text{Contador11}}{(\text{Contador11} + \text{Contador01})} \quad (2.3)$$

- **Especificidade:** calcula quão apto o modelo está para classificar valores negativos e calculada através da Equação 2.4:

$$\frac{\text{Contador00}}{(\text{Contador10} + \text{Contador00})} \quad (2.4)$$

Além da precisão, os algoritmos de MD podem ser analisados com relação a diversos aspectos, tais como velocidade de treinamento, robustez, escalabilidade e interpretabilidade [Han & Kamber 2006].

2.2.6 Utilização

Uma vez que o modelo de mineração de dados está construído e validado, ele pode ser utilizado de duas maneiras. Na primeira maneira um analista recomenda ações baseando-se simplesmente nos resultados obtidos do modelo. A segunda maneira é a aplicação de diferentes bases de dados ao modelo gerado. Neste caso, o modelo pode ser usado para criação de indicativos baseados na classificação dos exemplos ou para retornar uma pontuação como a probabilidade de uma determinada ação (exemplo, a resposta para uma determinada campanha de *marketing*) [Two Crows 1999].

Freqüentemente o modelo faz parte de um processo de negócio, tais como análise de risco de crédito e detecção de fraude.

2.3 Aplicações

Inicialmente, o conjunto de áreas de aplicação de KDD era bem restrito. Nos últimos anos, as organizações perceberam o potencial desta área e passaram a executar projetos de KDD em um conjunto cada vez maior de áreas de aplicação, entre as quais se destacam:

- Medicina: As bases de dados médicas possuem uma grande quantidade de dados históricos de doenças e exames. Tais dados são extremamente úteis no processo de tomada de decisão. No entanto, apresentam características que dificultam a análise, tais como: incompletude, incorretude, registros não representativos, falta de exatidão e valores coletados em intervalos distintos [Lavraè 1999]. Este cenário é bastante propenso a aplicações de KDD (como por exemplo, previsão de um paciente contrair determinada doença e previsão da efetividade de medicamentos) [Two Crows 1999];
- Setores Químico e Farmacêutico: Projetos de KDD vêm sendo desenvolvidos nestes setores para, entre outros fins, descobrir novas substâncias e inter-relações entre componentes químicos, e aprimorar fórmulas e componentes existentes [Two Crows 1999];
- Genética: Os seres humanos possuem milhares de genes que se apresentam de diversas formas. Devido a esta complexidade, os algoritmos de MD voltados para a análise de padrões seqüenciais, análise de associação (para identificação de seqüências de genes co-relacionadas), análise de caminho (para identificar a

influência dos genes nos diversos estágios de uma determinada doença) e busca por similaridade (entre seqüência de DNA) têm se tornado bastante úteis nesta área [Han & Kamber 2006];

- Internet: Os sites e as bases de dados associadas aos sistemas web possuem grande quantidade de dados que podem ser utilizados em projetos de KDD com diversas finalidades, tais como: descobrir padrões de navegabilidade dos visitantes, e identificar comportamentos e preferências dos consumidores [Witten & Frank 2000];
- Setor Financeiro e Varejista: Existe uma grande variedade de aplicações de KDD nos setores financeiro e varejista, entre as quais se destacam: previsão de vendas, análise *market-basket* (descobre associações entre produtos que são comprados em conjunto ou seqüência), segmentação de mercado (divide o mercado em segmentos cujos membros apresentam hábitos, necessidades e preferências comuns), análise de propostas de cartões de crédito e empréstimos, detecção de fraudes e tendências, e previsão de índices de bolsas de valores [Bigus 1996] [Fayyad et al. 1996a] [Monteiro 1999] [Witten & Frank 2000];
- Setor Elétrico: A aquisição de informações sobre demandas futuras é fundamental para as companhias elétricas. Se os resultados obtidos nos projetos de KDD forem precisos, as companhias poderão economizar em áreas como estabelecimento de reservas operacionais, planejamento de manutenção e gerenciamento do abastecimento [Witten & Frank 2000].

2.4 Considerações Finais

A área de KDD evoluiu, e continua evoluindo, devido aos benefícios alcançados em diversos domínios de aplicação. Este crescimento também tem sido impulsionado pela existência de tecnologias que proporcionam coleta, armazenamento e gerenciamento de grande quantidade de dados (fontes de informações valiosas que podem auxiliar no processo de tomada de decisão) [Rezende et al. 2003].

Apesar da evolução, KDD ainda possui uma série de tendências e perspectivas que podem motivar excelentes pesquisas. Entre as tendências e perspectivas desta área destacam-se [Fayyad et al. 1996a] [Rezende et al. 2003] [Amorim 2004]:

- Interação com o usuário: É importante que as ferramentas de KDD sejam interativas com o usuário, facilitando a incorporação de conhecimento prévio do domínio do problema e o entendimento do resultado obtido no final do processo;
- Aplicação de técnicas de Inteligência Artificial: Nos últimos anos, um número crescente de sistemas de KDD tem sido desenvolvido utilizando técnicas de IA. A maioria desses sistemas tem tido um bom desempenho em problemas considerados difíceis. Logo, a tendência é que os sistemas de KDD utilizem cada vez mais técnicas provenientes dessa área;
- Integração com outros sistemas: Os sistemas de KDD devem permitir a integração com outros sistemas, como SGBD, ferramentas analíticas e de visualização;
- Suporte a novos tipos de dados: Os SGBD orientados a objetos são capazes de manusear dados multimídia (imagem, vídeo e voz). Portanto, existe uma área em potencial para explorar dados de tipos complexos. Neste caso, duas soluções podem ser adotadas: extrair dados estruturados das bases de dados e minerar com as técnicas tradicionais; ou desenvolver ferramentas capazes de operar diretamente nos dados;
- Ambiente de rede e distribuído: O rápido crescimento de recursos disponíveis na Internet demanda uma grande necessidade por pesquisas para realizar o processo de KDD nesse ambiente conectado e distribuído. As pesquisas atuais de agentes inteligentes é um começo para atingir este desafio;
- Suporte a bases de dados de larga escala: Inicialmente o processo de KDD era validado com pequenos conjuntos de dados. À medida que bases de dados de larga escala (de problemas do mundo real) passaram a ser aplicadas, percebeu-se as limitações de alguns métodos que vinham sendo desenvolvidos. Portanto, é necessário que os métodos, técnicas e ferramentas aplicados em todo processo de KDD sejam capazes de lidar com bases de dados de larga escala;
- Automatização de todas as etapas do processo de KDD: À medida que as etapas passam a ser automatizadas, uma série de benefícios são obtidos, como por exemplo, diminuição do risco de inserção de erros humanos durante o processo e maior produtividade;
- Algoritmos incrementais: Em alguns domínios de aplicação os dados sofrem alterações contínuas (dados não estacionários), sendo necessária a utilização de métodos que sejam capazes de lidar com esta característica.

Dois aspectos que sempre devem ser considerados em projetos de KDD, independente das evoluções realizadas nesta área, são [Witten & Frank 2000] [Fayyad et al. 2001] [Rezende et al. 2003] [Amorim 2004]:

- Participação de especialistas do domínio: O conhecimento sobre o domínio do problema deve ser detalhadamente coletado no início do projeto, pois o mesmo servirá de subsídio para as demais etapas;
- Utilização de técnicas e ferramentas de visualização: Alcançar resultados eficientes não é o único objetivo da maioria dos projetos de KDD. É extremamente importante a obtenção de resultados compreensíveis para que o conhecimento descoberto seja aplicado no processo de tomada de decisões.

Capítulo 3

Metodologias de KDD e de Apoio à Montagem de Visões de Dados

3.1 Introdução

A realização de um projeto de KDD é uma tarefa muito complexa. Para garantir soluções de qualidade é imprescindível o uso de alguma metodologia que sirva de guia, especificando o fluxo a ser seguido e as atividades que devem ser realizadas.

Muitas são as metodologias propostas para o desenvolvimento de projetos de KDD. Apesar da maioria das metodologias citarem o processo de Preparação dos Dados, foco dessa dissertação, poucas metodologias específicas para montagem de visão de dados têm sido propostas.

A metodologia proposta nessa dissertação foi baseada em metodologias que serão apresentadas neste capítulo. As três primeiras são voltadas para todo o processo de KDD. As demais foram na Preparação de Dados. A seção 3.2 apresenta a metodologia de Fayyad et al. [Fayyad et al. 1996a], responsável pela origem do termo KDD. A seção 3.3 descreve a metodologia CRISP-DM [Chapman et al. 2000], metodologia não-acadêmica criada por um consórcio de empresas. A seção 3.4 aborda a metodologia DMEasy [Cunha 2005], que tem como proposta ser uma metodologia genérica para projetos de KDD integrada com um sistema de documentação de processos. Na seção 3.5 é apresentada a abordagem proposta de Yu et al. [Yu et al, 2006] apresenta um esquema integrado de preparação de dados para análise de dados com a utilização de RNA. Por último, na seção 3.6 é apresentada a metodologia da empresa Quadstone [Quadstone 2003], metodologia não acadêmica, usada como referência de utilização para os produtos desta empresa.

3.2 Abordagem de Fayyad et al

A abordagem proposta por Usama Fayyad, Gregory Piatetsky- Shapiro e Padhaic Smyth, em 1996, foi um marco para a área de Mineração de Dados, pois introduziu o conceito de MD em um contexto mais amplo, chamado de *Knowledge Discovery in Databases* (KDD), focando muito mais na solução do problema do que nos resultados das técnicas de mineração de dados [Cunha 2005].

Segundo [Fayyad et al. 1996a], KDD é um processo não trivial de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis. É um processo iterativo e interativo, composto das seguintes etapas: Seleção de Dados, Pré-Processamento, Transformação, Mineração de Dados e Interpretação e Avaliação dos Resultados, conforme Figura 3.1.

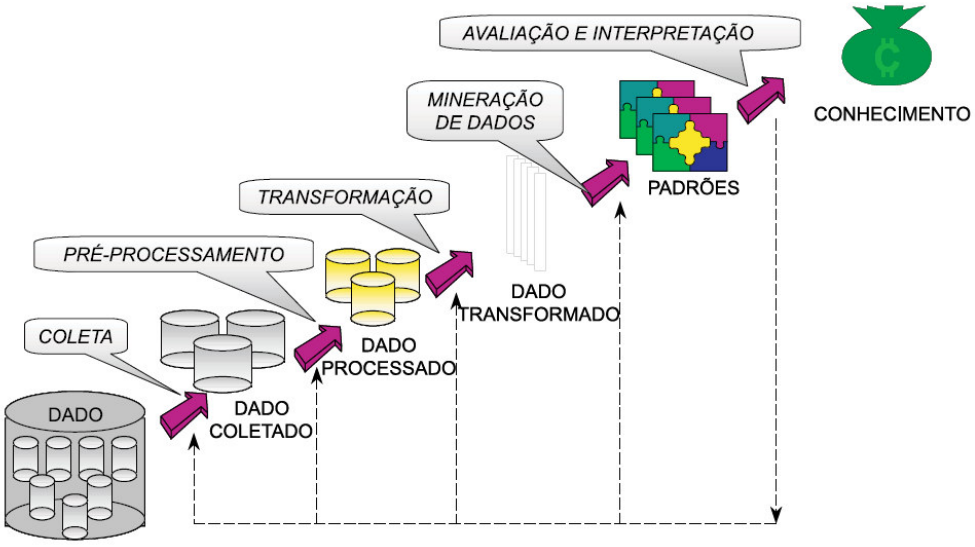


Figura 3.1 – Etapas do processo de KDD segundo Fayyad et al [Fayyad et al. 1996a]

Apesar de não ser mencionada como um dos passos da metodologia, os autores citam como a primeira etapa o entendimento do domínio da aplicação, a aquisição de conhecimentos relevantes e a definição do objetivo do processo de KDD do ponto de vista do usuário.

O segundo passo da metodologia é a criação/seleção da base de dados que será utilizada na etapa de MD para aquisição de conhecimento. A seleção da massa de dados deve considerar o objetivo que se deseja alcançar com o projeto.

O terceiro passo é o Pré-Processamento e Limpeza dos dados. É nesta etapa que dados com ruído, valores ausentes (*missing values*) e outras inconsistências devem ser tratados.

O quarto passo é a Transformação dos Dados, onde técnicas de redução de dados (para manter uma melhor representatividade dos dados) e transformação de dados (como a padronização para determinado algoritmo de MD) são aplicadas. No quinto passo (mineração de dados) seleciona-se o método que será utilizado para localizar padrões nos dados e aplica-se o método selecionado, ajustando seus parâmetros a fim de obter o resultado esperado.

O sexto passo é a Interpretação e Avaliação dos Padrões Minerados, com um possível retorno aos passos 1-5 para posterior iteração. Caso os resultados sejam satisfatórios, passa-se para o sétimo passo que é implantação do Conhecimento Descoberto. Nesse passo o conhecimento é incorporado ao desempenho do sistema ou é documentado e reportado as partes interessadas.

Do ponto de vista de Preparação de Dados e Montagem de Visão, essa abordagem não explora questões importantes, tais como:

1. Não especifica os papéis dos participantes do projeto, ou seja, não descreve as responsabilidades de cada participante nas fases do projeto.
2. É citada a aquisição de conhecimento do domínio, porém não trata claramente de levantamentos e estipulações essenciais para qualquer projeto, como especificação de recursos necessários e disponíveis, e elaboração de um planejamento completo, contendo definições de metas, critérios de sucesso e prazo.
3. Não aborda a documentação de processos realizados com sucesso ou falha entre as fases da metodologia. Apesar de definir o processo como iterativo, não cita como a abordagem lida com essa iteratividade [Cunha 2005].
4. Não se preocupa com a documentação de processos extras. Mesmo que o projeto tenha sido definido com maior quantidade de detalhes possíveis, é grande a possibilidade da execução de processos que fogem do escopo natural do projeto.
5. Não menciona a verificação de problemas existentes nas bases selecionadas no projeto. Tais problemas podem causar até o abortamento do projeto.
6. Não cita possíveis problemas que podem ser causados pela integração de diversas fontes de dados, nem soluções para os mesmos.
7. Não se preocupa com a homologação dos dados depois da realização de diversas transformações nas bases de dados.

3.3 CRISP-DM

A metodologia CRISP-DM [Chapman et al. 2000] foi concebida em 1996, por um consórcio de empresas de Mineração de Dados que estavam interessadas na expansão dessa área e na criação de uma metodologia que se tornasse referência na construção de projetos de KDD. Para tal, decidiram criar um modelo de processo padrão, não proprietário e disponível para todos, que fosse baseado em experiências reais de projetos de MD. Esta metodologia é descrita em termos de um modelo de processo hierárquico, que consiste de seis tarefas, conforme Figura 3.2.

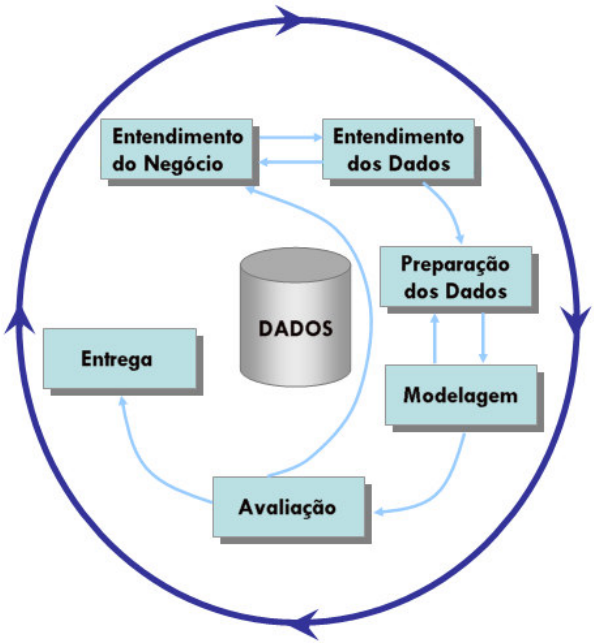


Figura 3.2 – Fases da metodologia CRISP-DM [Chapman et al. 2000]

A primeira fase é o Entendimento do Negócio, que é focado no entendimento dos objetivos do projeto e das necessidades sob a perspectiva de negócio. Neste passo define-se o problema de mineração de dados e o plano inicial do projeto para que todos os objetivos sejam alcançados.

A segunda fase é o Entendimento dos Dados, que se inicia com a coleta dos dados e prosegue com atividades voltadas para análise dos dados.

A terceira fase é a Preparação dos Dados, que contempla todas as atividades voltadas para construção da base final que será utilizada na fase de Modelagem. Atividades como transformação, seleção e limpeza são realizadas nessa fase.

A quarta fase é a Modelagem, onde uma ou mais técnicas de modelagem são aplicadas. Dependendo da técnica utilizada, é possível retroceder à fase de Preparação dos Dados para deixar os dados do modo apropriado para a aplicação da técnica.

A quinta fase é a Avaliação, onde os resultados gerados pelos modelos são analisados para garantir que os objetivos de negócio foram alcançados e decidir quanto ao uso do modelo no processo de tomada de decisão.

Por último, a sexta fase é a Utilização, que tem por objetivo a aplicação dos resultados obtidos com o projeto, seja em forma de relatórios, softwares ou simplesmente pela aquisição de conhecimento.

Do ponto de vista de Preparação de Dados e Montagem de Visão, a metodologia de CRISP-DM não explora questões importantes, tais como:

1. Não especifica os papéis dos participantes do projeto, ou seja, não descreve as responsabilidades de cada participante nas fases do projeto.
2. Aborda a documentação de processos realizados com sucesso ou falha entre as fases da metodologia, porém não especifica que tipo de ações devem ser tomadas caso a falha aconteça.
3. Não se preocupa com a documentação de processos extras. Mesmo que o projeto tenha sido definido com a maior quantidade detalhes possíveis, é grande a possibilidade da execução de processos que fogem do escopo natural do projeto.
4. Menciona o tratamento de dados, citando até a possibilidade de criação de variáveis, porém não aborda com detalhes sugestões para o tratamento e criação das mesmas, nem descreve o tipo de tratamento específico para o problema em questão.
5. Não considera a homologação dos dados depois da realização de diversas transformações nas bases de dados.

3.4 DMEasy

A DMEasy é uma metodologia interativa e iterativa para projetos de KDD proposta por Rodrigo Cunha [Cunha 2005]. Essa metodologia é fortemente baseada em CRISP-DM [Chapman et al. 2000], DMLC (*Data Mining Life Cycle*) [Hofmann & Tierney 2003] e PMBOK (*Project Management Body of Knowledge*) [PMBOK 2000], e foi proposta como

intuito de ser uma metodologia geral, completa e que atenda à realidade dos desenvolvedores de soluções de MD.

A DMEasy, ilustrada na Figura 3.3, está dividida em seis fases: Iniciação, Planejamento e Organização, Execução, Distribuição e Finalização, Acompanhamento e Controle do Projeto e Processos Genéricos.

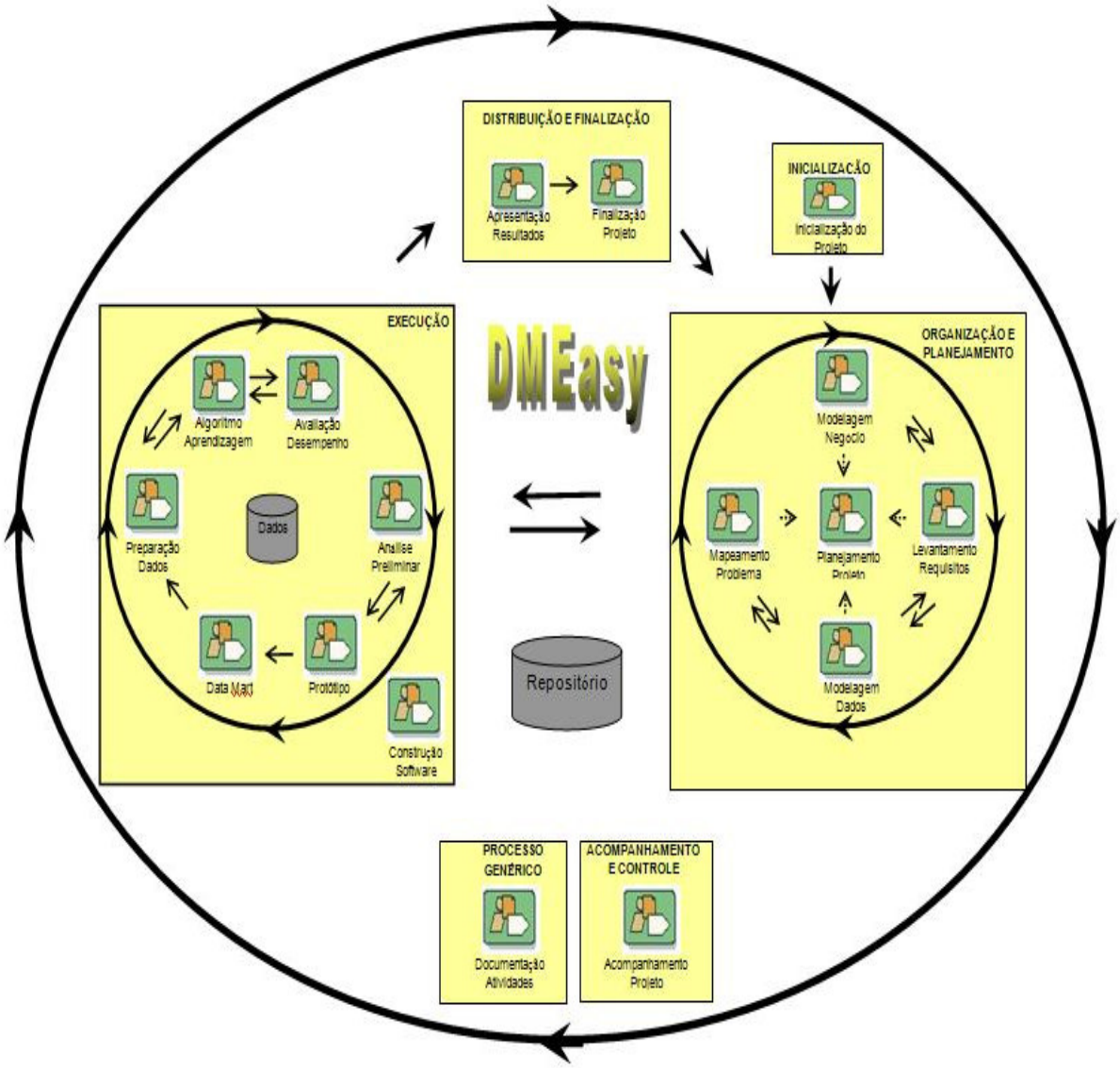


Figura 3.3 – Fases da metodologia DMEasy [Cunha 2005]

A primeira fase é a Inicialização, que tem como objetivo formalizar o inicio do projeto de KDD para todos os envolvidos. É nessa fase que os riscos relacionados ao projeto são analisados.

A segunda fase é o Planejamento e Organização, onde as informações e os requisitos são coletados a fim de elaborar um planejamento completo do projeto. É uma das fases mais importantes, pois é a base para todas as outras fases. Além do levantamento dos requisitos, são averiguadas informações sobre os dados disponíveis, o mapeamento do problema para alguma tarefa de MD, e os prazos e custos do projeto.

A terceira fase é a Execução, onde os dados são o centro do desenvolvimento. Nesta fase, os dados são manipulados e transformados em conhecimento estratégico, para apoiar o negócio do cliente. Um fator importante dessa fase é a análise dos dados com relação a inconsistências, pois dados com problemas dificultam a obtenção de resultados satisfatórios. Após a análise de consistência, os dados são preparados e transformados no formato específico do algoritmo de MD escolhido na segunda fase.

A quarta fase é a distribuição dos resultados obtidos e a formalização do encerramento do projeto. Após a entrega dos resultados, verifica-se se os mesmos atendem as expectativas especificadas inicialmente pelo cliente. Se os resultados não forem satisfatórios, pode ser necessário voltar para fase de Planejamento e Organização.

A quinta fase é o Acompanhamento e Controle do Projeto, que abrange todo o controle e monitoramento do projeto, desde a fase de Planejamento e Organização até a finalização do projeto.

Existe uma fase que nem sempre é executada, que é a fase de Processos Genéricos. Esta fase tem como finalidade documentar as atividades e/ou processos extras que não são abordados na metodologia DMEasy.

Esta metodologia aborda conceitos importantes de um projeto de KDD como:

- Estrutura das fases do projeto: cada fase do projeto possui um conjunto de entradas, um responsável, um conjunto de saídas e a descrição das atividades que devem ser executadas.
- Recursos envolvidos: são identificados os tipos de recursos envolvidos em projetos de KDD. Esta foi uma proposta de melhoria com relação a metodologia DMLC [Hofmann & Tierney 2003];
- Reuso do conhecimento;
- A importância do processo ser iterativo, abordando com clareza os motivos que fazem com que um projeto retroceda às fases anteriores.

Apesar de ser uma metodologia bem completa, ela aborda com mais clareza a natureza de um projeto de KDD, sem dar muita ênfase a problemas relacionados à montagem da visão

de dados. Do ponto de vista de Preparação de Dados e Montagem de Visão, alguns aspectos importantes são simplesmente mencionados, mas não abordados com detalhes, como por exemplo:

1. Menciona uma análise de inconsistências nos dados, porém não aborda, com detalhes, possíveis problemas na coleta de dados e na integração de bases (como por exemplo, cardinalidade e duplicidade). Além disso, não especifica que ações devem ser tomadas caso existam inconsistência nos dados, o que pode levar ao aborto de um projeto.
2. Menciona o desenvolvimento de um *Data Mart* (repositório de dados) [Inmon 2005] específico para o projeto (de acordo com a montagem da visão conceitual das variáveis), porém não cita que tipo de transformações podem ser realizadas.
3. Menciona o tratamento de dados, porém não aborda em detalhes que tipo de tratamento pode ser utilizado para o problema em questão.
4. Não considera a homologação dos dados, depois da realização de diversas transformações nas bases de dados.

3.5 Abordagem de Yu et al.

A proposta elaborada por Lean Yu, Shouyang Wang K.K. Lai [Yu et al. 2006], tem como intuito ser um esquema integrado de preparação de dados para análise de dados com a utilização de Redes Neurais Artificiais, focando em problemas com os dados e a apresentação das técnicas de processamento correspondentes.

Esta metodologia está dividida em três fases: Pré-Análise dos dados, Pré-Processamento dos Dados e Pós-Análise dos Dados, cada uma dessas fases também é dividida em sub-processos. Esse esquema pode ser observado na Figura 3.4.

A primeira fase é a de Pré-Análise dos Dados, que consiste na identificação e coleta de dados de interesse. Esta fase está subdividida em quatro sub-processos: Análise de Requisitos (Entendimento dos requisitos para o projeto em conjunção com a definição do problema e a definição dos objetivos), Coleta dos Dados (Coleta de dados baseados nos requisitos anteriormente definidos), Seleção de variáveis (Obtenção de um conjunto de variáveis que seja mais representativo para o problema) e Integração dos Dados (Realiza a integração de dados de diversas fontes e formatos).

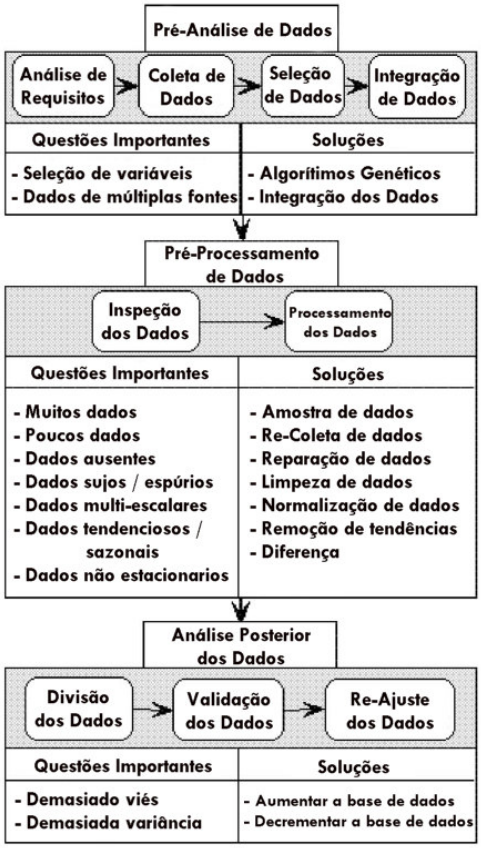


Figura 3.4 – Fases da abordagem de Yu et al [Yu et al. 2006]

A segunda fase é a de Pré-Processamento, onde os dados são analisados, transformados e estruturados de maneira a torná-los mais úteis para o problema. Essa fase é dividida em dois sub-processos: Verificação dos Dados (Verificação de problemas existentes nos dados, a fim de analisar a qualidade dos mesmos) e Pré-Processamento dos Dados (Aplicação de diversas técnicas para corrigir possíveis problemas encontrados no sub-processo anterior). Essa fase quer garantir que os dados sejam suficientes para o problema, limpos, com a correção de dados ausentes e espúrios, normalizados e no formato específico para a aplicação de RNA.

A terceira fase é a de Pós-Análise dos Dados, onde os dados obtidos da fase anterior são utilizados para o treinamento de uma RNA, contendo os seguintes sub-processos: Divisão dos Dados (Dividir os dados em subconjuntos para o treinamento de uma RNA), Validação dos Dados (Verificação dos resultados obtidos) e Ajuste dos Dados (Ajuste nos dados ou nos parâmetros da RNA para garantir mais qualidade dos resultados).

Analisando a proposta de Yu no ponto de vista de Preparação de Dados e Montagem de Visão, essa abordagem não explora questões importantes, tais como:

1. Não especifica os papéis dos participantes do projeto, ou seja, não descreve as responsabilidades de cada participante nas fases do projeto.
2. É citada a aquisição de conhecimento do domínio, porém não trata claramente levantamentos e estipulações essenciais para qualquer projeto, como especificação de recursos necessários e disponíveis, e elaboração de um planejamento completo, contendo definições de metas, critérios de sucesso e prazo.
3. Não menciona nenhum tipo de documentação de processos realizados com sucesso ou falha entre as fases da metodologia, nem a especificação do tipo de posição que deve ser tomadas caso a falhas aconteçam.
4. Não se preocupa com a documentação de processos extras.
5. Menciona o tratamento de dados, porém não guia na possível criação de novas variáveis que possam garantir a correção de erros existentes nos dados.
6. Não leva em consideração a homologação dos dados, depois da realização de diversas transformações nas bases de dados.
7. Não cita possíveis problemas que podem ser causados pela integração de diversas fontes de dados distintas, como também soluções para os mesmos.
8. Menciona uma análise de inconsistências nos dados, porém não aborda com detalhes, como possíveis problemas na coleta de dados, verificações de problemas com a integração de bases, cardinalidade, duplicidade e outros. Não deixa específico que ações são tomadas caso descoberto inconsistências dos dados, questão que pode levar ao aborto de um projeto.

3.6 Metodologia de Quadstone

Proposta em 2003 pela empresa escocesa Quadtsone Limited, tem por abordagem ser uma metodologia para guiar a criação de projetos utilizando o sistema Quadstone System.

Esta metodologia está dividida em três fases: Captura de Comportamento do Cliente, Análise dos Dados e Modelagem e Escoragem e Tomada de Decisão.

A primeira fase é a de Captura de Comportamento do Cliente. Esta fase é focada no levantamento das bases a serem utilizadas (verificando quais os tipos de dados disponíveis) e a criação da Visão de Dados, levando em consideração a obtenção de um arquivo contendo um registro por cliente. Nessa fase é proposto um padrão de variáveis geralmente criadas, levando em consideração o mercado de varejo.

A segunda fase é a de Análise dos Dados e Modelagem. A primeiro momento, consiste na verificação de problemas existentes com a massa de dados e como tratar alguns problemas como a existência de valores ausentes (*missing value*), valores espúrios (*outliers*) e outras transformações necessárias para a modelagem. A parte de Modelagem consiste na aplicação de algum método de mineração de dados, como também a validação do modelo e avaliação dos resultados.

A terceira fase é a de Escoragem e Tomada de decisão, que consiste na análise da melhor forma de tomar decisão perante os resultados obtidos pelo projeto e quais as preocupações a serem tomadas para realizar a escoragem de novos exemplos ao modelo criado.

Analisando a metodologia de Quadtsone no ponto de vista de Preparação de Dados e Montagem de Visão, essa abordagem não explora questões importantes, tais como:

1. Não especifica os papéis dos participantes do projeto, ou seja, não descreve as responsabilidades de cada participante nas fases do projeto.
2. Não é citada a aquisição de conhecimento do domínio do problema.
3. Não menciona nenhum tipo de documentação de processos realizados com sucesso ou falha entre as fases da metodologia, nem a especificação do tipo de posição que deve ser tomadas caso a falhas aconteçam.
4. Não se preocupa com a documentação de processos extras.
5. Não leva em consideração a homologação dos dados, depois da realização de diversas transformações nas bases de dados.
6. Não cita possíveis problemas que podem ser causados pela integração de diversas fontes de dados distintas, como também soluções para os mesmos.
7. Menciona uma análise de inconsistências nos dados, porém não aborda com detalhes, como possíveis problemas na coleta de dados, verificações de problemas com a integração de bases, cardinalidade, duplicidade e outros. Não deixa específico que ações são tomadas caso descoberto inconsistências dos dados, questão que pode levar ao aborto de um projeto

3.7 Metodologia de Han & Kamber

Esta proposta está inserida no livro de Jiawei Han e Micheline Kamber [Han & Kamber 2006], tem por abordagem ser uma metodologia focada na fase de pré-processamento de dados e na explanação sobre as técnicas de mineração de dados comumente em projetos de KDD.

Sua metodologia se assemelha muito a metodologia proposta por Fayyad et al [Fayyad et al. 1996a], porém com um detalhamento maior das fases que são mais relevantes para essa dissertação.

Essa metodologia está dividia em quatro fases: Limpeza e Integração, Seleção e Transformação, Mineração de dados e Avaliação e Apresentação, cada uma dessas fases também é dividida em sub-processos. Esse esquema pode ser observado na Figura 3.5.

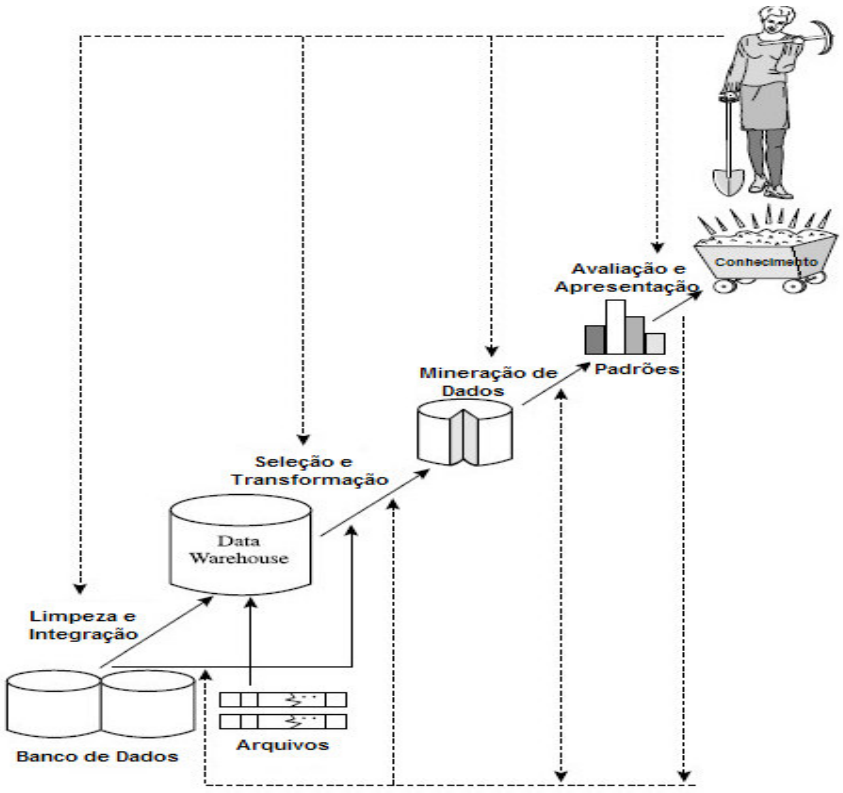


Figura 3.5 – Fases da abordagem de Han & Kamber [Han & Kamber 2006]

A primeira fase é a de limpeza e integração. Nessa fase são discutidas técnicas para limpeza de valores ausentes, redução de ruído e a explanação da limpeza dos dados como um processo em um projeto de mineração. Além da limpeza, são tratados conceitos de integração de dados, tendo como base a utilização de técnicas para evitar a redundância dos dados.

A segunda fase é a de seleção e transformação dos dados. Na fase de transformação dos dados, são explanadas técnicas para a transformação e a consolidação dos dados no formato apropriado para mineração. Técnicas como agregação, *smoothing*, generalização, normalização e construção de variáveis estão contidas nessa fase. Na fase de seleção dos dados, são explanadas técnicas para a obtenção de uma representação reduzida dos dados, que é menor no volume, mas que mantém a integridade original dos dados. Estratégias como agregação através de cubo de dados, seleção de um subconjunto de atributos, redução de dimensionalidade, discretização e conceitos de generalização de hierarquia são contidas nessa fase.

A terceira fase é a de mineração dos dados. Nessa fase, são explicadas em detalhe técnicas como regras de associação, classificação, previsão, análises de clusters, mineração de séries temporais e mineração de objetos.

A quarta e última fase é a de avaliação e apresentação de resultados. Em cada explanação das técnicas citadas na fase de mineração de dados, é explicado uma forma de avaliação dos resultados obtidos. Por fim, são explicadas algumas formas de apresentação de resultados e aplicações de mineração de dados.

Apesar de ser uma metodologia bem completa no aspecto de pré-processamento, alguns aspectos importantes como verificação de dados, homologação e documentação ficaram fora dessa metodologia. Do ponto de vista de Preparação de Dados e Montagem de Visão, essa abordagem não explora questões importantes, tais como:

1. Não especifica os papéis dos participantes do projeto, ou seja, não descreve as responsabilidades de cada participante nas fases do projeto.
2. Não é citada a aquisição de conhecimento do domínio do problema.
3. Não menciona nenhum tipo de documentação de processos realizados com sucesso ou falha entre as fases da metodologia, nem a especificação do tipo de posição que deve ser tomadas caso a falhas aconteçam.
4. Não se preocupa com a documentação de processos extras.
5. Não leva em consideração a homologação dos dados, depois da realização de diversas transformações nas bases de dados.
6. Menciona uma análise de inconsistências nos dados, porém não aborda com detalhes, como possíveis problemas na coleta de dados, verificações de problemas com a integração de bases, cardinalidade, duplicidade e outros. Não deixa específico que ações são tomadas caso descoberto inconsistências dos dados, questão que pode levar ao aborto de um projeto

3.8 Considerações Finais

As metodologias apresentadas neste capítulo demonstram características importantes do ponto de vista de Preparação e Transformação dos dados, que foram resumidas na Tabela 3.1. A notação "N" foi atribuída para as metodologias que não abordam a característica identificada. O valor "M" para as que mencionam, mas não entram em detalhes e "A" para as que abordam em detalhes.

As características identificadas como desejáveis em uma metodologia para montagem de visão em base de dados foram:

- **Levantamento e Definição do problema:** esta característica identifica se a metodologia aborda uma das fases essenciais de todo projeto de KDD que é a de levantamento e definição do problema;
- **Verificação dos dados:** identifica se a abordagem trata a questão de verificação de inconsistências existentes nos dados, quais as causas e citar alguma forma de tratá-los;
- **Integração dos dados:** identifica se a abordagem trata da fase de integração dos dados, citando os problemas que podem ocorrer nessa etapa e as formas de corrigir problemas originados nessa fase;
- **Transformação de variáveis:** identifica se a abordagem trata a fase de transformação de variáveis, que possibilita a utilização de variáveis com preenchimento muitas vezes considerados impróprios;
- **Agrupamento Transacional:** identifica se a abordagem especifica como tratar dados espalhados por diversas tabelas, inserindo maior conteúdo aos dados;
- **Homologação dos dados:** identifica se a abordagem especifica a homologação dos dados para cada fase em que se faz necessário qualquer tipo de transformação, inserção ou remoção de atributos;
- **Tratamento de variáveis:** identifica se a abordagem especifica que tipo de técnicas de tratamento de variáveis podem ser utilizadas;
- **Documentação:** identifica se a abordagem trata o processo de documentação de atividade no decorrer do projeto;
- **Processos Extras:** identifica se a abordagem especifica uma fase geral, onde é tratado todo processo extra que foge ao escopo da metodologia;

- **Abordagem a diversos problemas de MD:** identifica se a metodologia engloba aspectos relevantes aos diversos problemas de mineração de dados.

Tabela 3.1 – Comparativo de metodologias

Atividade	Fayyad	CRISP-DM	DMEasy	Yu	Quadstone	Han & Kamber
Levantamento e Definição do problema	M	A	A	A	N	N
Verificação dos dados	N	M	M	N	N	N
Integração dos dados	N	N	N	A	N	A
Transformação de variáveis	N	N	M	N	M	PA
Agrupamento Transacional	N	N	M	N	A	PA
Homologação dos dados	N	N	M	A	N	N
Tratamento de variáveis	A	M	M	A	N	A
Documentação	N	N	A	N	N	N
Processos Extras	N	N	A	N	N	N
Abordagem a diversos problemas de MD	N	N	N	N	N	A

N – Não Aborda; **M** – Menciona; **A** – Aborda; **PA** – Parcialmente Aborda

Apesar da maioria das metodologias estudadas para o desenvolvimento desta dissertação levar em consideração o levantamento e definição do problema, a aquisição detalhada do conhecimento do domínio não é muito abordada, principalmente no que se trata de aquisição de documentação de fontes de dados. Este ponto é muito importante para qualquer projeto, pois muitas transformações e pré-processamentos só se fazem possíveis com o conhecimento prévio e detalhado do problema e das bases de dados.

Outro fator importante é a documentação de processos realizados entre as fases do projeto, sejam estas realizadas com sucesso ou falha, para possíveis retornos ou justificativas de falha.

Observou-se que, apesar de ser comumente citado na literatura que a fase de Preparação de Dados é importante para qualquer projeto de KDD, nenhuma metodologia aborda com detalhes todas as atividades que fazem parte desta fase. Questões importantes como verificações dos dados, análise dos problemas de integração, aplicação de técnicas de transformações e agrupamento de dados são meramente mencionadas (e só são abordadas em duas metodologias).

A documentação de processos extras só é abordada em uma metodologia. Esta fase é importante, pois realiza a documentação de processos que fogem ao escopo do projeto e serve de apoio a melhorias futuras das metodologias.

A homologação dos dados também é dificilmente citada nas metodologias (somente abordada em uma metodologia). Este conceito é extremamente necessário para garantir a qualidade dos tratamentos realizados sobre as bases de dados.

Analisando todos os aspectos positivos e negativos das metodologias investigadas, verificou-se a necessidade de uma metodologia que englobasse o que há de melhor nessas metodologias, a fim de guiar melhor o processo de Montagem de Visão dos Dados. Além das principais características das metodologias investigadas, foi verificada a necessidade de abordagem de questões não relatadas em nenhuma metodologia, como a abordagem com detalhes de erros (e suas correções) em base de dados e o detalhamento de transformação de variáveis (além de abordar as técnicas convencionais de tratamento de variáveis), que fazem ser o diferencial da metodologia proposta.

Capítulo 4

DMBuilding: Metodologia Proposta

4.1 Introdução

Várias são as metodologias de apoio ao processo de KDD como um todo, porém nenhuma é específica para a preparação dos dados, tendo em vista os diversos problemas de mineração de dados existentes (classificação, segmentação, regressão, entre outros).

O objetivo desse capítulo é apresentar a metodologia proposta por essa dissertação, *DMBuilding,* metodologia específica para montagem de visão para problemas de mineração de dados.

A metodologia *DMBuilding* é fortemente baseada nos princípios das metodologias de KDD, como a proposta por Fayyad et al. [Fayyad et al. 1996a], CRISP-DM [Chapman et al. 2000] e DMEasy [Cunha, 2005], como também em metodologias mais focadas na área de preparação dos dados, como as abordagens de Yu et al. [Yu et al, 2006] e da empresa Quadstone [Quadstone 2003].

A idéia da metodologia *DMBuilding* é fornecer à equipe participante do projeto de mineração de dados os principais subsídios para preparar os dados, dando apoio a equipe desde a definição do problema até a aplicação de técnicas para o tratamento dos dados.

4.2 Metodologia

Conforme ilustrada na Figura 4.1, a metodologia *DMBuilding* está dividida em cinco fases:

- **Entendimento do problema:** esta é a fase onde são elaboradas as definições do projeto, que servirão de base para o resto do projeto. Garante que todos os

requisitos serão coletados e que o planejamento de todo o projeto será traçado, definindo o que vai ser resolvido, de que forma, os prazos e os responsáveis;

- **Verificação dos dados:** Após o entendimento do problema, os dados disponíveis serão analisados para verificar o que realmente pode ser utilizado, se os dados estão íntegros e se os mesmos são representativos para a resolução do problema proposto. Podem haver iterações com a fase anterior dependendo da natureza dos dados;
- **Montagem da visão:** Tendo o parecer positivo quanto à utilização dos dados para o projeto, chega-se à fase de montagem da visão, que aplica técnicas sobre os dados para agregar valor aos dados brutos;
- **Tratamento dos dados:** Com os dados já trabalhados, é realizado o tratamento final dos dados, para deixá-los no formato esperado pelo método de Inteligência Artificial;
- **Processos extras:** Fase responsável pela documentação de atividades realizadas que não são abordadas nesta metodologia.

A Figura 4.1 apresenta a metodologia *DMBuilding*: os retângulos em azul representam as principais fases do projeto. Cada uma dessas fases é composta de vários processos, que são representados por retângulos brancos. As iterações entre as fases e entre os processos de cada fase são representadas pelas setas bi-direcionais.

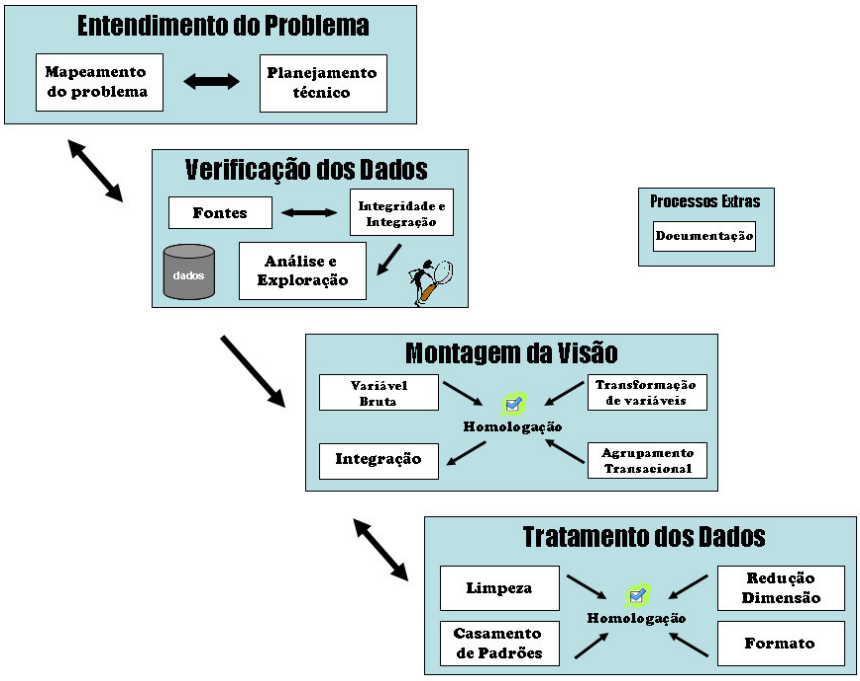


Figura 4.1 – Visão geral da metodologia *DMBuilding*

A metodologia *DMBuilding* está focada no suporte à preparação dos dados e na documentação dos processos realizados para preparar os dados, possibilitando o reuso da informação em projetos futuros e uma melhor estruturação dessa fase que pode demandar uma série de re-trabalhos e retornos a fases anteriores do projeto.

Os processos de todas as fases do projeto são compostos de entradas e saídas e de uma relação dos responsáveis e das atividades a serem desenvolvidas.

A execução de cada processo é totalmente dependente do tipo do problema de mineração de dados e a natureza dos dados. Sendo assim, alguns projetos irão passar por todos os processos da metodologia *DMBuilding* ou somente por aquelas que são relacionadas à sua natureza.

Como essa metodologia está focada na preparação dos dados, as fases de modelagem e avaliação (parte fundamental das metodologias de KDD) não são contempladas, porém tem-se em vista que estas são as próximas fases a serem executadas em qualquer projeto de KDD. Se os resultados obtidos com a fase de modelagem não forem satisfatórios, há a possibilidade do retorno às fases citadas nessa metodologia.

Os responsáveis pelo projeto de montagem da visão são:

- **Analista de negócio:** responsável pelo entendimento do problema como um todo. Tem por responsabilidade levantar, entender e documentar os requisitos do projeto.
- **Líder do projeto:** responsável por agrupar todas as informações relacionadas ao projeto, planejar os passos a serem realizados e garantir que esses passos sejam executados da melhor forma possível.
- **Especialista no domínio:** conhece o domínio no qual o projeto de mineração de dados será aplicado. Tem por responsabilidade dar suporte ao projeto, especialmente em questões relacionadas ao problema.
- **Especialista em Banco de Dados:** responsável pelo fornecimento de mecanismos de análise dos dados e pela construção das variáveis solicitadas para o projeto.
- **Analista de dados (minerador)**: entende as técnicas envolvidas no processo de MD, principalmente as técnicas de tratamento dos dados. Tem conhecimento sobre o funcionamento dos algoritmos e das ferramentas utilizadas no processo, mas não necessariamente conhece o domínio ao qual os dados pertencem.

- **Especialista em TI:** detém todo o conhecimento do funcionamento como um todo da área de informática da instituição a qual será aplicado o resultado do processo de mineração de dados. Seu conhecimento é necessário para o levantamento e entendimento de todos os dados que serão necessários para a realização do projeto.

As próximas seções irão abordar, de forma detalhada, as fases que compõe a metodologia *DMBuilding*.

4.2.1 Entendimento do Problema

Esta fase tem por objetivo formalizar a inicialização do projeto de montagem da visão de dados. É neste momento que são averiguadas informações pertinentes ao problema em questão, atribuídas responsabilidades e definidos os prazos e os custos do projeto.

O projeto de mineração de dados inicia-se com uma definição concisa do problema. É através dessa definição que se pode verificar a possibilidade da aplicação de métodos de mineração de dados para a resolução do problema pretendido. Por este motivo a definição precisa do problema é um dos fatores de sucesso do projeto.

A compreensão do negócio também é um fator decisivo, pois o entendimento de todo o fluxo da informação auxilia na preparação e no tratamento dos dados e evita uma série de re-trabalhos causados por falta de conhecimento sobre o domínio do problema.

Essa fase está dividida em dois processos: Mapeamento do Problema e Planejamento Técnico.

A Figura 4.2 ilustra o fluxo de atividades desta fase.

4.2.1.1 Mapeamento do Problema

O mapeamento do problema é o primeiro processo a ser realizado na metodologia *DMBuilding*. Durante esse processo são elaboradas documentações formais sobre o problema a ser resolvido para que haja um entendimento comum de todos os envolvidos no projeto. As entradas desse processo são o documento de formalização do início do projeto e a informações sobre a situação do negócio da organização.

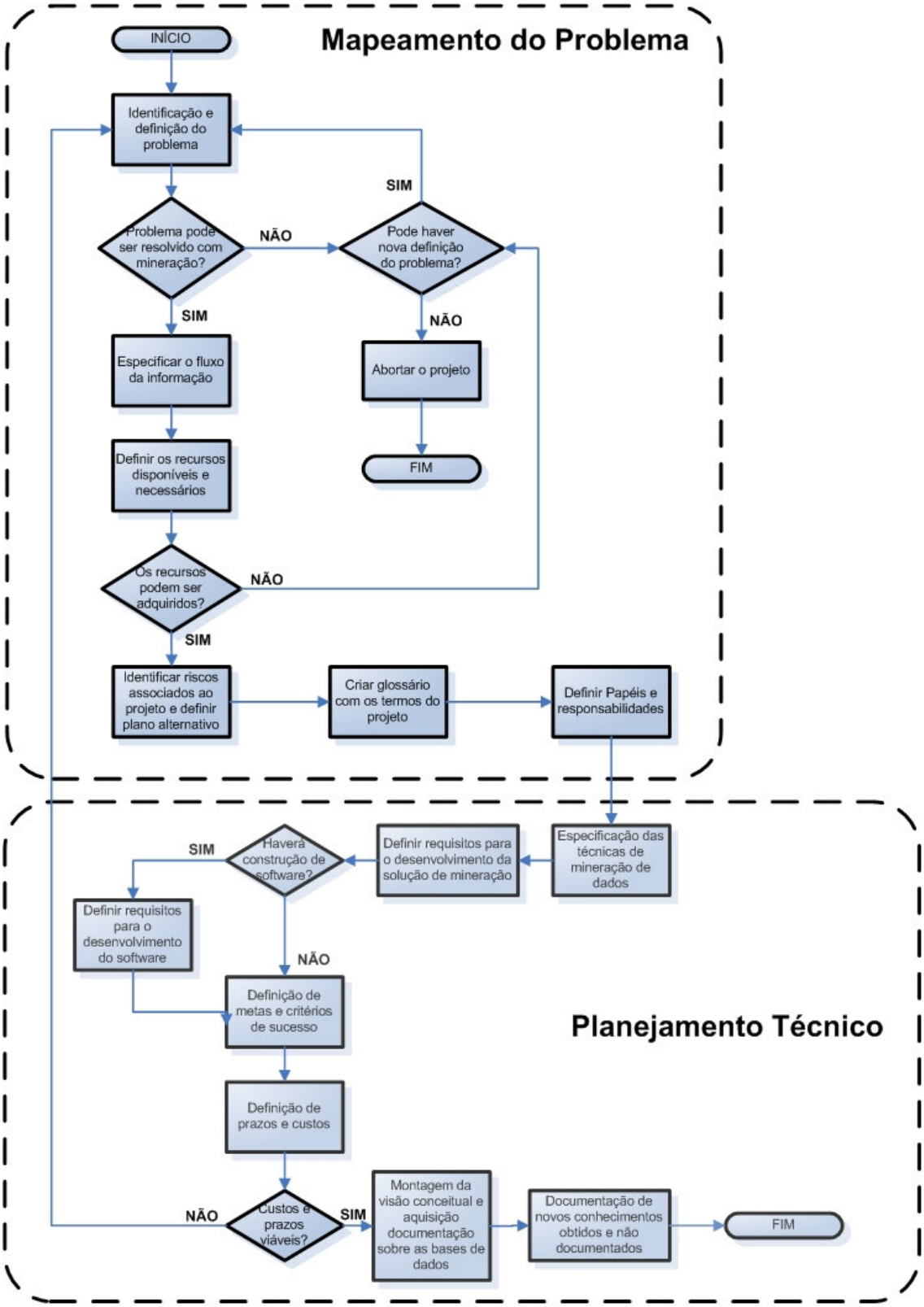


Figura 4.2 – Fluxo de atividades da fase de Entendimento do Problema

Entender o negócio como um todo é um pré-requisito para o descobrimento de conhecimento. Sem esse entendimento, é provável que ocorram diversos problemas na fase de preparação de dados e montagem da visão. Conseqüentemente os dados não serão tratados de forma a produzir resultados confiáveis após a aplicação de qualquer algoritmo [Berry & Linoff 2004].

Os responsáveis por esse processo são o líder do projeto e analista de negócios, que coletam, analisam e documentam todos os objetivos da organização. Contam com o suporte do especialista do domínio, que possui todas as informações referentes ao domínio da aplicação e do especialista em TI, que auxilia nas questões referentes à área de informática.

A primeira atividade a ser realizada é a compreensão do que a organização realmente deseja realizar. É freqüente que existam muitos objetivos concorrentes e restrições que devem ser balanceadas. O objetivo é expor, desde o início, os fatores importantes que podem influenciar o resultado do projeto. Após a compreensão dos objetivos, define-se o problema que a organização deseja resolver. Uma possível conseqüência ao não realizar essa atividade é gastar muito esforço para produzir as respostas certas para as perguntas erradas [Chapman et al. 2000].

Diversas discussões com os responsáveis pela organização devem ser realizadas, a fim de alinhar as informações com todos os envolvidos no projeto. É importante que todas as reuniões e entrevistas sejam documentadas para posteriores consultas.

O próximo passo é verificar se o problema que se pretende resolver pode/deve ser traduzido em um problema que possa ser resolvido através da utilização de técnicas de MD. Nos casos onde o problema pode ser resolvido com MD, o projeto segue adiante. Em alguns casos, ou não é possível se resolver com MD ou existe solução mais prática (como uma simples solução algorítmica). Nos casos onde não é possível o uso de técnicas de MD, deve-se verificar se é possível uma nova definição do problema que seja traduzida em um problema de MD. Se uma nova definição não é possível, o projeto é abortado. Caso haja a possibilidade de tradução, o projeto segue seu fluxo.

Para a definição de um problema, é importante considerar que o projeto deve ter o tamanho ideal para ser considerado praticável e ser relevante para a organização. Em casos de dúvidas quanto à melhor caracterização do problema, uma das alternativas é examinar exemplos de problemas que foram executadas com sucesso na mesma área de atuação da organização.

Após a especificação do problema, deve ser averiguado e documentado o fluxo de informação da organização. Segundo [Petró et al. 2006], o fluxo de informação está

relacionado a atividades ligadas à produção, disseminação e uso da informação, desde a concepção de uma idéia até a sua explicitação e aceitação como parte do conhecimento universal.

Para o problema, o fluxo da informação identifica a forma como os dados são obtidos, relacionados e armazenados. Esse tipo de informação é importante para:

- Definir quais variáveis podem ser utilizadas;
- Entender o motivo de anomalias presentes nos dados, como falta de preenchimento, erro em formato, entre outros;
- Compreender o fluxo do negócio;
- Definir o melhor momento para utilizar os resultados que serão obtidos com o projeto de mineração de dados.

Outra questão importante é a relação dos recursos mínimos para a realização do projeto. Devem ser verificados os recursos atualmente disponíveis (incluindo pessoas, dados, recursos computacionais, softwares, entre outros). Posteriormente é necessário definir quais os recursos extras que devem ser adquiridos. Se os recursos necessários para o desenvolvimento do projeto não podem ser adquiridos, deve-se avaliar se uma nova definição de problema pode ser feita (por exemplo, fazendo a redução do escopo do problema), considerando os objetivos da organização e os recursos disponíveis. Caso não seja possível, o projeto pode ser abortado.

É de suma importância o total envolvimento da equipe de T.I. da organização, pois problemas com máquinas, recursos humanos, documentação, organização, responsabilidade e agilidade podem afetar o rendimento do projeto.

Além da análise dos recursos, é importante identificar os riscos associados ao projeto. Um risco é qualquer que pode causar atrasos ou falhas no projeto. Se existem riscos associados ao projeto, um plano alternativo deve ser criado a fim de descrever as ações que devem ser executadas caso os riscos ocorram.

O analista de negócios também é responsável por criar um glossário com todas as terminologias referentes ao projeto. No glossário devem constar termos referentes ao domínio da organização e à mineração de dados, para que todos os envolvidos no projeto estejam habituados com os termos mais utilizados.

A última atividade desse processo é a definição das pessoas envolvidas no projeto. Um documento relacionando nomes, responsabilidades e formas de contato é gerado nesta

atividade. Na grande maioria dos casos, uma mesma pessoa pode acumular responsabilidades, dependendo do seu papel na organização ou no projeto.

As entradas, as saídas, os responsáveis e as atividades associadas a este processo estão especificadas abaixo.

Entradas:

- Documento de formalização do início do projeto;
- Documentos informativos relacionados ao negócio da organização (mercado, rendimento, área de atuação, etc);

Atividades:

- Identificar e definir o problema a ser resolvido;
- Especificar o fluxo da informação da organização;
- Definir os recursos disponíveis e os necessários;
- Especificar os riscos associados ao projeto e elaborar um plano alternativo;
- Elaborar o glossário;
- Definir os papéis e responsabilidades dos envolvidos no projeto.

Responsáveis:

- Líder do projeto;
- Analista de Negócio;
- Especialista do Domínio;
- Especialista em TI.

Saídas:

- Documento com a descrição do problema: descreve o problema a ser resolvido com mineração de dados;
- Fluxo da informação: mostra o fluxo da informação dentro da organização;
- Recursos disponíveis e necessários: lista os recursos necessários para a realização do projeto, especificando os atualmente disponíveis e os que deverão ser adquiridos;
- Lista de riscos e plano alternativo: contém os riscos associados ao projeto e as alternativas que devem ser tomadas caso os riscos venham a acontecer;

- Glossário do projeto: contém toda a terminologia do projeto, seja relacionada ao negócio ou à mineração de dados;
- Lista de envolvidos no projeto e suas responsabilidades: os nomes, os papéis, as responsabilidades e as formas de contato das pessoas envolvidas no projeto.

4.2.1.2 Planejamento Técnico

Para completar a fase de Entendimento do Problema, é necessário realizar o planejamento técnico do projeto. Este processo envolve a especificação das técnicas de mineração de dados que serão aplicadas, a definição dos prazos e custos, a especificação de metas e dos critérios de sucesso, a identificação dos requisitos para a execução do projeto, a aquisição da documentação sobre as bases de dados e a documentação de novos conhecimentos.

Esse processo recebe todas as saídas do processo de Mapeamento do Problema. Os responsáveis por essa fase são o líder do projeto e o analista de negócios, com o apoio do engenheiro de software. Este só será requisitado caso haja a necessidade de desenvolvimento de software. Se requisitado, o analista de dados também dará apoio a esse processo.

A primeira atividade a ser realizada é a definição da melhor técnica de mineração de dados a ser utilizada para solucionar o problema. A escolha do algoritmo não é uma tarefa fácil, pois não existe um único algoritmo que apresente o melhor desempenho para todos os problemas. As técnicas de mineração de dados são aplicadas com o objetivo de extrair informações estratégicas escondidas em grandes bancos de dados [Goebel & Gruenwald 1999]. Segundo [Harrison 1998], não existe uma técnica que possa resolver todos os problemas de MD. Métodos diferentes servem para diferentes fins. Portanto, é necessário que o analista dos dados (minerador) tenha familiaridade com as diversas técnicas existentes, a fim de auxiliar na escolha de uma delas de acordo com o problema apresentado.

Os seguintes aspectos que devem ser considerados quando selecionando um algoritmo de mineração de dados [Adriaans & Zantinge 1996]:

- Número de exemplos: Alguns algoritmos trabalham melhor com uma quantidade maior de exemplos do que outros;
- Número de atributos: O desempenho de alguns algoritmos, como RNA e Algoritmos Genéticos, deteriora-se consideravelmente à medida que o número de atributos aumenta;
- Tipos de atributos que podem ser manipulados;

- Representação do conhecimento: Como o conhecimento é apresentado para determinado algoritmo, como no caso das RNA que não fornecem explicação de suas respostas;
- Capacidade de aprender de forma incremental: Quando novos dados tornam-se disponíveis, o algoritmo é capaz de revisar suas teorias, sem refazer completamente o processo de aprendizagem. Isto pode ser de grande relevância em aplicações com bases grandes;
- Habilidade de estimar a significância estatística dos resultados: Em alguns algoritmos, como RNA e AG, normalmente é muito difícil avaliar os resultados estatisticamente;
- Desempenho.

Após a definição das técnicas de mineração de dados, é necessário especificar quais os requisitos para o desenvolvimento do projeto.

Em um projeto de MD, o resultado obtido pode ser apresentado de diversas formas (relatório, gráfico, etc). A definição do formato de apresentação dos resultados pode requerer demandas extras, como a implementação de algum software que dê apoio ao processo de tomada de decisão. Caso haja a necessidade de construção de software, os requisitos devem ser definidos, de acordo com alguma metodologia de engenharia de software, como RUP (*Rational Unified Process*) [Fuggetta 2000] [Jacobson et al. 1999], XP (*eXtreme Programming*) [Wojciechowski 2002] e *Agile Software Development* [Watson et al. 2003].

Com os requisitos já definidos, devem ser especificadas as metas relacionadas ao projeto e os critérios de sucesso. Em seguida, os prazos e custos do projeto devem ser analisados. A implantação dos resultados obtidos pode afetar o fluxo atual de a tomada de da organização. Neste caso, é necessário avaliar se essa alteração é possível e se será implementada com facilidade e rapidez, quando os resultados forem obtidos.

A definição do prazo e do custo de qualquer projeto de mineração de dados está ligada ao tamanho e ao escopo do projeto. Um fator importante na definição dos prazos é o comprometimento e a dedicação de todos os envolvidos no projeto, principalmente os responsáveis pela parte de T.I. da organização e do especialista do domínio. Outro fator importante é a qualidade e a complexidade dos dados (averiguada na fase de verificações dos dados). Dependendo da necessidade de correções ou aquisições de dados, pode haver atraso nos prazos do projeto. Os custos estão ligados às demandas necessárias para a realização do projeto.

Caso os prazos e/ou custos estipulados para o projeto não sejam viáveis, é necessário o retorno à atividade de Identificação e Definição do Problema para adequar a definição do problema aos custos e prazos disponíveis.

Por último, um dos passos é a montagem da visão de dados conceitual. O Analista de Negócio é responsável por esta atividade, partindo de seu conhecimento do que este considera ser variáveis importantes para o projeto. Essa visão conceitual é homologada pelo especialista no domínio, que irá verificar a viabilidade da utilização / criação daquelas variáveis, como também sugerir novas variáveis de acordo com o seu entendimento do negócio. Para a realização dessa atividade, é fundamental que a documentação sobre as bases de dados seja a mais completa possível. Diagrama de Entidade e Relacionamento (D.E.R.), Dicionário de dados, meta-dados [Ramakrishnan & Gehrke 2002], relatórios, fichas cadastrais e telas de sistemas são exemplos de documentações importantes a qualquer projeto. É sabido que algumas organizações não possuem uma documentação completa ou não existe documentação disponível para o projeto. Nesses casos, é importante que o analista de negócios documente todas as informações recebidas sobre as bases de dados (com o apoio do especialista em T.I., o especialista em Banco de Dados e do especialista do domínio). Essas documentações serão utilizadas para definir as ações a serem realizadas sobre as bases de dados. Um exemplo disso é a utilização do D.E.R. para visualizar o relacionamento entre as tabelas, verificar problemas inerentes e homologar a estrutura dos dados [Cougo 1999].

Ao chegar ao fim dessas atividades, novos conhecimentos podem ter sido adquiridos. Esses conhecimentos devem ser documentados para futuras consultas.

A fase de entendimento do problema é uma fase iterativa. Em várias situações pode ocorrer o retorno a atividades anteriores, caso algum problema tenha sido encontrado. Essas iterações também devem ser documentadas. Ao final dessa fase, um parecer de viabilidade do projeto deve ser criado pelo analista de negócios.

As entradas, as saídas, os responsáveis e as atividades associadas a este processo estão especificados a seguir:

Entradas:

- Documento com a descrição do problema;
- Fluxo da informação;
- Recursos disponíveis e necessários;
- Lista de riscos e plano alternativo;

- Glossário do projeto;
- Lista de envolvidos no projeto e suas responsabilidades.

Atividades:

- Especificação da(s) técnica(s) de mineração de dados: definir as técnicas de MD que poderiam ser utilizadas na resolução do problema e especificar a mais apropriada para o projeto;
- Definição de requisitos para o desenvolvimento da solução de mineração de dados: definir os requisitos relacionados à execução do projeto. Caso haja necessidade de desenvolvimento de software, os requisitos do software devem ser especificados de acordo com alguma metodologia de engenharia de software;
- Definição das metas e dos critérios de sucesso: formaliza as metas e os critérios de sucesso associados ao projeto;
- Definição de prazos e custos: formaliza os custos e prazos estipulados para a realização do projeto;
- Definição da visão de dados conceitual e aquisição de documentação sobre as bases de dados: Criar uma visão de dados conceitual do projeto. Caso não haja documentação suficiente sobre a base de dados, o máximo de informações devem ser coletadas e documentadas;
- Documentação de novos conhecimentos obtidos e não documentados.

Responsáveis:

- Líder do projeto;
- Analista de negócios;
- Engenheiro de Software;
- Analista de dados;
- Especialista no domínio.

Saídas:

- Documento definindo as técnicas de mineração de dados, os requisitos do projeto, as metas, os critérios de sucesso, os prazos e os custos;
- Visão de dados conceitual, homologada pelo especialista no domínio;
- Documentações sobre as bases de dados;

- Documentação de novos conhecimentos.

4.2.2 Verificação dos Dados

Após a definição dos objetivos do problema, deve-se determinar e analisar os dados que serão usados para o descobrimento do conhecimento.

Uma análise profunda na estrutura e no conteúdo dos dados disponíveis deve ser realizada. Os problemas inerentes aos dados são identificados, listados e, se possível, corrigidos.

Essa fase está dividida em três processos: identificação e seleção da massa de dados, análise de integridade e integração e análise e exploração dos dados.

A fase de Verificação de Dados é muito importante para um projeto de montagem de visão, pois é nessa fase que verifica a qualidade dos dados e decide se os dados estão aptos para resolver o problema proposto.

As saídas da fase de Entendimento do Problema servem de entrada, principalmente as inerentes às bases de dados.

Na fase de Verificação dos Dados existe uma grande interação entre os envolvidos para que não restem dúvidas quanto à natureza dos dados, seus significados e componentes. O responsável por essa fase é o especialista em Banco de Dados. Outros envolvidos são o especialista em T.I. e o especialista no domínio.

A Figura 4.3 define o fluxo de atividades desta fase.

4.2.2.1 Identificação e Seleção das Fontes de Dados

O passo inicial desse processo é a identificação das fontes de dados e de determinados problemas que possam interferir na coleta dos mesmos.

Os dados podem ser coletados de diversas fontes, como por exemplo [Myatt 2007]:

- **Banco de dados operacionais:** contêm dados transacionais de negócio, que são acessados e atualizados constantemente.
- ***Data Warehouses*:** é uma cópia de dados agrupados de várias partes da organização. Os dados são limpos e otimizados para a tomada de decisões. Não é atualizado com a mesma freqüência dos bancos de dados operacionais.

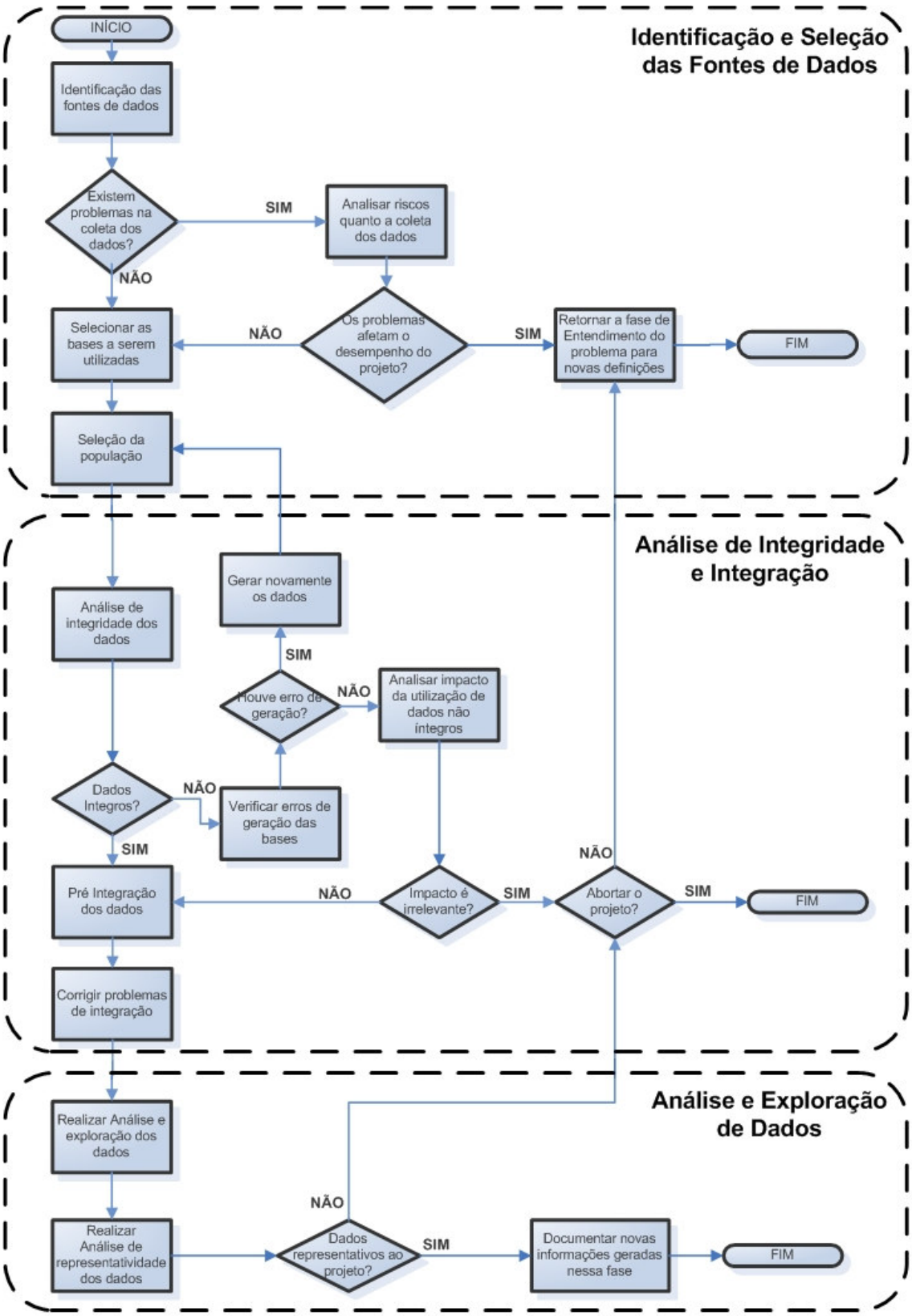


Figura 4.3 – Fluxo de atividades da fase de Verificação dos Dados

- **Banco de dados históricos:** são utilizados para armazenar pesquisas históricas, levantamentos e experimentos.
- **Dados adquiridos:** Em muitos casos, os dados internos da organização não são suficientes para a aplicação de algoritmos de MD. Uma das técnicas utilizadas para resolver esse problema é a aquisição de dados de fontes externas, para que sejam combinados com os dados internos, desde que estes condizem com o problema que está sendo tratado.

Durante esse processo, é importante se atentar a problemas relacionados à coleta dos dados. Eis alguns exemplos de problemas que devem ser considerados [Batista 2003]:

- **Problemas legais e éticos:** Nem todos os dados disponíveis na base podem ser obtidos, pois pode haver barreiras legais ou éticas que impeçam que os mesmos sejam disponibilizados para análise. Dados que identificam o indivíduo (como CPF, endereço, número de cartões de crédito) normalmente possuem essas restrições.
- **Motivos estratégicos:** Por motivos estratégicos da organização, algumas informações não podem ser acessadas. As companhias de cartão de crédito, por exemplo, mantêm em absoluto segredo a proporção de operações fraudulentas [Chain & Stolfo 1998].
- **Razões políticas:** Nem todo projeto de MD conta com o apoio total de todos os responsáveis da organização. Em certos casos, os dados podem pertencer a departamentos cujos responsáveis não apóiam à iniciativa de analisar os dados. Essas pessoas podem impor restrições de acesso aos dados, atrasando ou inviabilizando o projeto.
- **Formato dos dados e conectividade:** Devido à evolução dos sistemas de armazenamento de dados, os formatos dos dados vêm sendo alterados A utilização de diversos tipos de mídias (disquetes, fitas, CD-ROM, entre outros) e de técnicas de codificação de dados (ASCII, EBCDIC, etc.) podem complicar a coleta de dados advindos de diversas fontes. Da mesma maneira, sistemas antigos (legados) e proprietários podem dificultar a conectividade aos dados, uma vez que esses sistemas podem utilizar tecnologias obsoletas para a troca de informações, as quais não estão disponíveis nos novos sistemas computacionais.

- **Bancos de dados e aplicações obsoletas:** Existem sistemas transacionais que foram projetados e desenvolvidos numa época em que técnicas de Engenharia de Software ainda não haviam sido desenvolvidas. Como resultado, existe uma escassez de documentação sobre esses sistemas Essa escassez pode dificultar a localização e a extração dos dados desses sistemas. Além disso, sistemas gerenciadores de banco de dados muito antigos podem não possuir tipos de dados equivalentes nos sistemas atuais, o que dificulta a unificação e representação dos dados.

Se forem detectados problemas na coleta dos dados, é necessário fazer uma análise do risco do projeto quanto à utilização dos dados com problemas. Se os problemas não afetam de forma significativa o desempenho do projeto, o processo pode ser seguido normalmente para as próximas etapas. Caso contrário, o projeto deve ser retornado para a fase de Entendimento do Problema, a fim de que uma nova definição do problema seja realizada que não implique em problemas de coleta.

Após a análise das fontes disponíveis e dos problemas de coleta de dados, é necessário realizar um filtro das bases que podem ser utilizadas para a resolução do problema. Essa atividade é realizada juntamente com o especialista de T.I. e o especialista do domínio. Todas as fontes selecionadas devem ter alguma correlação com o problema (direta ou indireta) e devem estar relacionadas à visão de dados conceitual. No decorrer do projeto, a visão de dados conceitual pode ser alterada, se estiver de acordo com o analista de negócio e especialista no domínio. Caso haja inserção de variáveis, uma nova visão conceitual de dados deve ser armazenada.

De posse de todas as bases selecionadas, é necessário realizar a seleção exata da população a ser tratada no problema. A população está diretamente relacionada ao problema. A seleção da população geralmente é realizada quando a base de dados representa um universo maior que o universo do problema a ser resolvido, havendo assim a necessidade de selecionar parte desse universo. Um exemplo é a seleção de clientes que sejam do tipo "Pessoa Física", pois a base de dados original contém registros tanto de "Pessoa Física" como de "Pessoa Jurídica".

Um aspecto importante a ser considerado nesta atividade é de determinar o período amostral dos dados. É importante garantir que os dados selecionados sejam o mais recentes possíveis [Quadstone 2003].

As entradas, as saídas, os responsáveis e as atividades associadas a este processo estão especificadas a seguir:

Entradas:

- Documento com a descrição do problema;
- Visão de dados conceitual;
- Fluxo da informação;
- Documentações sobre as bases de dados.

Atividades:

- Documentar as origens dos dados;
- Verificar problemas relacionados à coleta e analisar os riscos associados;
- Selecionar as bases de dados a serem utilizadas no projeto;
- Selecionar a população a ser tratada no problema.

Responsáveis:

- Especialista em Banco de dados.
- Especialista em T.I.
- Especialista no domínio.

Saídas:

- Relação das fontes de dados: documento descrevendo a origem dos dados. Esta relação deve especificar o software utilizado para o armazenamento dos dados;
- Relação dos problemas de coleta: documento que descreve todos os problemas relacionados à coleta de dados, suas causas, os riscos associados e as ações que devem ser realizadas caso os riscos ocorram;
- Seleção das bases utilizadas: documento que relata as bases que serão efetivamente utilizadas no projeto e o modo como a seleção dos dados foi realizada;
- Visão de dados conceitual alterada;

4.2.2.2 Análise de Integridade e Integração

A análise de integridade considera aspectos referentes à estrutura e organização das bases disponíveis ao projeto. Essa atividade é fundamental para garantir a utilização das bases de dados com qualidade.

Esta atividade tem como entrada todas as bases de dados selecionadas, como também as documentações geradas no processo anterior, tendo como principal instrumento a documentação dos dados.

O responsável por essa fase é o Especialista em Banco de Dados. Como essa atividade está profundamente ligada à análise detalhada dos dados, é fundamental o suporte do especialista em T.I. e do especialista no domínio.

O primeiro passo a ser realizado é a análise das chaves primárias das tabelas [Ramakrishnan & Gehrke 2002]. Para integrar dados de diversas fontes, é necessário que as tabelas possuam campos que identifiquem de forma única o registro (chamados de chaves primárias). Devido a problemas de geração das massas (advindo da seleção dos dados ou do próprio banco de dados onde os dados foram extraídos), as chaves primárias das tabelas podem apresentar valores duplicados, ausentes ou com preenchimento incorreto. Este é um dos principais problemas relacionados à integridade dos dados e deve ser reportado para possível correção. Caso não possa ser corrigido, devem ser analisadas alternativas para a integração entre as bases (como a utilização de outros campos, podendo gerar dados não íntegros, e a exclusão dos campos repetidos, que pode ocasionar a perda de informação).

Caso as chaves primárias estejam íntegras, é necessário verificar a conexão entre as tabelas. Considerando o exemplo de duas tabelas, *Contrato* e *Parcela,* cada contrato pode ter uma ou mais parcelas e cada parcela só deve pertencer a um contrato. Podem existir problemas como contratos sem parcelas ou parcelas sem o registro do contrato. Além disso é necessário verificar se o período das bases de dados selecionadas são as mesmas estipuladas para o projeto. Todos esses problemas podem ser ocasionados na seleção de dados, pois o período das duas bases pode ser diferente e deve ser reportado para possível correção. Caso não possa ser corrigido, os registros que não possuam correlação com as outras bases podem ser excluídos, havendo a possibilidade de perda de informação.

Em problemas de aprendizado supervisionado, além da verificação da cardinalidade, é imprescindível que seja realizada uma análise considerando o alvo do projeto. Essa análise verifica se esse atributo já está contido nos dados e se este foi calculado de maneira correta (caso as bases relacionadas ao alvo estejam disponíveis). Quando o alvo não está previamente calculado, é necessário verificar a possibilidade de criá-lo de acordo com os dados disponíveis.

Este processo é crítico para determinar a continuação do projeto. Caso seja encontrado algum dos problemas citados acima, a origem dos mesmos deve ser verificada. Se forem erros de geração das bases, os dados devem ser re-gerados e as verificações refeitas. Se os erros

forem intrínsecos à natureza dos dados, deve ser analisado o impacto de utilização destes dados. Dependendo do percentual de problemas encontrados e da impossibilidade de correção, o projeto pode ser abortado. Se possível, uma alternativa é a de retorno à fase de entendimento do problema, para definir uma nova descrição de problema que não contemple erros encontrados neste processo ou que o impacto dos mesmos não seja tão significativo.

Geralmente, essas verificações são feitas utilizando ferramentas de consultas de banco de dados. Os scripts gerados para realizar as verificações devem ser armazenados para posteriores consultas.

A metodologia *DMBuilding* não foca a utilização de nenhum software em específico para a realização das verificações dos dados. É improvável que essa fase seja realizada utilizando somente um software. Cabe ao Especialista em Banco de dados a utilização dos softwares aos quais possua mais conhecimento, ou que melhor se enquadra ao problema em questão.

Com o parecer positivo quanto para utilização dos dados, é necessário realizar a integração dos mesmos. Normalmente a integração é realizada quando os dados advêm de diversas fontes, como bancos de dados diferentes, arquivos com formatos diferentes, ou até mesmo de diversos setores de uma mesma organização. Nem toda organização segue um padrão em seus sistemas. Por tanto, em uma mesma organização, podem existir sistemas operacionais, bancos de dados e até sistemas internos diferentes que foram sendo adquiridos ou desenvolvidos de forma independente ao longo do tempo, não possuindo a integração desejada.

A estrutura e a semântica heterogênica dos dados é um dos grandes desafios da integração de dados. Cuidados na integração de dados de fontes diversas podem ajudar a reduzir ou evitar redundâncias e inconsistências [Kimball & Ross 2002].

Meta-dados podem ser utilizados para auxiliar na integração entre tabelas. Na literatura, o conceito de meta-dados não é um consenso. A definição mais citada é de “dados sobre dados” [Kimball & Ross 2002] [Han & Kamber 2006]. Um dos exemplos da utilização de meta-dados é quando existem dois atributos de tabelas diferentes e com nomes diferentes, porém com o mesmo significado (como, por exemplo, *c_cli* e *codigo_cli*). Com a utilização das informações contidas nos meta-dados, a integração das tabelas é possível. Segundo [Han & Kamber 2006], esse é o problema de identificação de entidades. As informações oriundas dos meta-dados para cada atributo incluem nome; significado; tipo do dado; faixa de valores permitidos; como tratar brancos, zeros ou valores nulos; a existência de alguma dependência entre os dados; entre outros.

Os conflitos de valores dos atributos também devem ser resolvidos. Novamente, se faz necessário o uso de meta-dados. Atributos podem conter conflitos de formato ou escala. No conflito de formato, os atributos possuem o mesmo significado em tabelas diferentes, porém com formatos distintos. Como exemplo, o atributo "*sexo*" pode conter os valores "*F*" ou "*M*" em uma determinada tabela. Entretanto, o mesmo atributo pode assumir os valores "*1*" ou "*2*" (respectivamente representando F ou M) em outra tabela. De posse dos meta-dados, pode-se realizar a padronização de formatos. No conflito de escala os atributos apresentam escalas diferentes. Como exemplo, o atributo "*temperatura*", em uma determinada tabela é representado na escala Celsius, enquanto em outra tabela, a escala é Fahrenheit. Transformações são necessárias para que após a integração os atributos apresentem o mesmo formato ou a mesma escala.

Após a integração, é necessário verificar a existência de redundância ou duplicação nos dados. Segundo [Mayer 1998], redundância é o armazenamento de dados idênticos em um mesmo banco de dados, o que ocasiona o desperdício de espaço de armazenamento, inconsistência e quedas no desempenho do banco de dados.

Toda base de dados pode conter redundância, mas é de extrema importância eliminá-la. A redundância de dados pode ser oriunda de erros em sistemas, de falha na coleta dos dados ou até mesmo do processo de integração.

A redundância pode ser tratada com o auxílio de meta-dados. A partir deles, é possível identificar que dois atributos que aparentemente possuem conteúdos diferentes apresentam o mesmo significado.

A redundância também ocorre quando um atributo pode ser derivado de outros atributos. Nesse caso, deve-se analisar se é mais relevante ao problema a retirada do atributo derivado ou dos atributos que o geraram.

Algumas redundâncias em atributos numéricos podem ser detectadas através de análise de correlação [Han & Kamber 2006], que mede o quão fortemente um atributo está implicando no outro. A correlação entre dois atributos A e B pode ser medida através da Equação 4.1:

$$r_{A,B} = \frac{\sum_{i=1}^{N}(a_i - \overline{A})(b_i - \overline{B})}{N\sigma_A\sigma_B} = \frac{\sum_{i=1}^{N}(a_i b_i) - N\overline{A}\,\overline{B}}{N\sigma_A\sigma_B} \qquad (4.1)$$

onde N é o número de linhas do arquivo, a_i e b_i são respectivamente valores de A e B na tupla i, $\overline{A}$ e $\overline{B}$ são as respectivas médias de A e B, σA e σB são respectivamente o desvio padrão de A e B, e $\sum(a_i b_i)$ é a soma do produto externo de AB (que é o valor de A multiplicado pelo valor de B na linha). O resultante é um numeral entre -1 e 1. Se o resultante da Equação 4.1 é maior do que zero (0), então A e B são positivamente correlacionados, significando que os valores de A crescem a medida que os valores de B crescem. Quanto maior o valor, mais um atributo implica no outro. Este valor alto pode indicar que A (ou B) pode ser removido como redundância. Se o resultante for igual à zero (0), então A e B são independentes e não existe correlação entre eles. Se o resultante for menor que zero, então A e B são negativamente correlacionados, significando que os valores de A diminuem à medida que os valores de B crescem, e vice-versa. É importante ressaltar que correlação não implica em causalidade, ou seja, se A e B são correlacionados, não necessariamente A causa B.

Para atributos categóricos, a correlação entre dois atributos A e B pode ser descoberta através do teste de chi-quadrado (χ^2) [Han & Kamber 2006]. Supondo que A possua c elementos distintos ($a_1, a_2, .. a_c$) e B possua r elementos distintos ($b_1, b_2, .. b_r$). Os dados descritos por A e B podem ser vistas como uma tabela de contingências, com os valores c de A representando as colunas e os valores r de B representando as linhas. Supondo que ($A_i B_j$) denote o evento em que o atributo A assume um valor a_i e o atributo B assume o valor b_j, cada possível combinação de ($A_i B_j$) tem sua própria célula na tabela. O χ^2 é computado através da a seguinte Equação 4.2:

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \qquad (4.2)$$

onde o_{ij} é a freqüência observada do evento ($A_i B_j$) e e_{ij} é a freqüência esperada de ($A_i B_j$), que pode ser computada pela Equação 4.3:

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_i)}{N} \qquad (4.3)$$

onde N é o número de linhas, $count(A = a_i)$ é o número de linhas contendo o valor a_i para o atributo A e $count(B = b_i)$ é o número de linhas contendo o valor b_i para o atributo B.

O chi-quadrado testa estatisticamente a hipótese que A e B são independentes. Se a hipótese for rejeitada, pode-se concluir que A e B são estatisticamente relacionados.

A redução de redundância poderá influenciar na diminuição do tempo de processamento para a maioria dos algoritmos de mineração de dados. Além disso, algumas técnicas de modelagem (especialmente as técnicas baseadas em regressão) apresentam problemas para assimilar dados redundantes, gerando até falhas de execução [Pyle 1999].

Duplicações em nível de tuplas (ou seja, duas ou mais tuplas totalmente idênticas) também devem ser detectadas e eliminadas. O uso de tabelas desnormalizadas é uma das principais causas deste tipo de problema.

A realização do processo de análise de integridade e integração visa à garantia da qualidade dos dados. Ao final desse processo, gera-se uma documentação, contendo todos os erros encontrados. Caso haja interações com outros responsáveis por esse processo, as mesmas também devem ser documentadas. Com o parecer positivo do especialista em banco de dados, o projeto segue para as fases posteriores.

As entradas, as saídas, os responsáveis e as atividades associadas a este processo estão especificadas a seguir:

Entradas:

- Documentações sobre as bases de dados;
- Visão de dados conceitual;
- Bases selecionadas para o projeto;
- Relação dos problemas relacionados à coleta e suas influências;

Atividades:

- Verificar integridade dos dados;
- Identificar erros referentes à geração da massa de dados;
- Analisar impacto da utilização de dados não íntegros;
- Realizar pré-integração dos dados;
- Corrigir erros relacionados à pré-integração;
- Dar o parecer técnico quanto à utilização da massa de dados.

Responsáveis:

- Analista dos dados.
- Especialista em T.I.
- Especialista no domínio.

Saídas:

- Análise de integridade dos dados: documento contendo todas as inconsistências encontradas nas bases de dados, como também os artefatos utilizados;
- Relatório de pré-integração dos dados: contém todas as integrações feitas nos dados, contemplando os problemas e as possíveis soluções;
- Parecer de integridade: um parecer do especialista em banco de dados quanto à integridade dos dados;
- Bases de dados integradas, com possíveis correções de conflitos ou problemas causadas pela fase de integração.

4.2.2.3 Análise e Exploração de Dados

Com os dados íntegros e integrados, o especialista em Banco de Dados parte para a análise da qualidade dos dados com relação aos aspectos de preenchimento e conteúdo. Para tal, são necessários alguns métodos de análise e exploração dos dados.

As entradas para desse processo são todas as saídas do processo anterior.

A primeira análise a ser realizada é referente à granularidade dos dados. Segundo [Inmon 2005], granularidade é o nível de detalhe ou de resumo das unidades de dados. Quanto mais detalhe, mais baixo o nível de granularidade. Quanto menor detalhe, mais alto o nível de granularidade.

No contexto de um projeto de mineração de dados, dois conceitos de granularidade devem ser abordados: do problema e das bases. Com relação à granularidade das bases, quanto menor, mais flexibilidade no processo de criação de novas variáveis, pois os dados atômicos podem ser resumidos de várias maneiras. Além disso, os dados de menor granularidade podem ser estendidos com atributos adicionais, medidas ou dimensões sem romper com processos existentes [Kimball & Ross 2002]. A granularidade do problema é definida pelo nível de detalhes da solução. Esta também define a granularidade final da base de dados, por isso o analista dos dados deve verificar se as bases estão na granularidade esperada para a resolução do problema proposto.

Para aprendizagem supervisionada, existem basicamente dois tipos de variáveis: de entrada (preditivas) e de saída (de alvo) [Witten & Frank 2005]. As variáveis de entrada são aquelas que serão utilizadas para verificar a relação entre suas variações e o comportamento

de outras variáveis, ou seja, correspondem àquilo em função do qual se deseja conseguir realizar previsões e/ou controle. Variáveis de saída são aquelas cujo comportamento se quer verificar em função das oscilações das variáveis preditivas, ou seja, correspondem àquilo que se deseja prever e/ou controlar. Para a aprendizagem não supervisionada, somente o conceito de variáveis de entrada são existentes.

A próxima atividade é a análise dos tipos de dados. Para isso, é interessante verificar informações como a escala dos dados. Diferentes tipos de escala são inerentes a diferentes tipos de informação. Escalas representam formas de medir atributos qualitativos e quantitativos. Os tipos de escala comumente vistos em projetos de mineração de dados são [Pyle 1999]:

- **Nominal:** variáveis com escala nominal possuem um número limitado de valores diferentes, que só podem ser "iguais" ou "diferentes" entre si. Não possui ordem e serve para identificar se determinado valor pertence ou não a uma categoria. Exemplos: estado civil, CEP, cor dos olhos;
- **Ordinal:** impõe uma ordem entre os valores. Cada observação faz a associação do indivíduo com uma determinada classe, sem, no entanto, quantificar a magnitude da diferença face aos outros indivíduos. Exemplos: escala social (pobre, classe média, rico), medida de opinião (ruim, médio, bom);
- **De Intervalo:** possibilita quantificar a distância entre as medições em termos de sua intensidade, ou seja, posicionando as medições com relação a um valor conhecido arbitrariamente (ponto zero). Tal aferição é feita por comparação a partir da diferença entre o valor do ponto zero e um segundo valor conhecido. Portanto, não há um ponto nulo e uma unidade natural. Exemplos: escalas de temperatura (Celsius, Kelvin, Farenheit), em que não se pode assumir um ponto 0 (ponto de nulidade) ou dizer que a temperatura (X) é o dobro da temperatura (Y);
- **De Razão:** é a mais completa e sofisticada das escalas numéricas. Descreve que a mesma diferença entre valores tem o mesmo significado (como num intervalo). É uma quantificação produzida a partir da identificação de um ponto zero, que é fixo e absoluto, representando, de fato, um ponto de nulidade, ausência e/ou mínimo. Nesse tipo de escala, uma unidade de medida é definida em termos da diferença entre o ponto zero e uma intensidade conhecida. Portanto cada observação é aferida segundo a sua distância do ponto zero. Um aspecto importante a ser observado é que um valor de “2” efetivamente indica uma quantidade duas vezes

maior que o valor "1", e assim por diante, o que não necessariamente acontece nas demais escalas.

Para que a fase de verificações dos dados seja completa, é necessária a utilização de métodos de análise descritiva dos dados. A análise descritiva serve para verificar propriedades pertinentes aos dados e possíveis problemas relacionados ao preenchimento, formato e completude.

A estatística descritiva tem um papel importante na fase de verificações de dados, [Han & Kamber 2006]. Seu objetivo básico é sintetizar uma série de valores de mesma natureza, permitindo uma visão global da variação desses valores. Os dados são organizados e descritos de três maneiras: tabelas, gráficos e medidas descritivas.

Vários são os métodos de estatística descritiva. Entre os quais, destacam-se [Myatt 2007]:

- **Intervalo de valores:** representa os possíveis valores que um atributo pode assumir especificando os valores máximo e mínimo. Existem dois tipos de intervalos: os presentes nos dados e os esperados pelo negócio. A comparação entre esses dois tipos de intervalos serve para averiguar anomalias, como valores espúrios (*outliers*);
- **Distribuição de freqüência:** é o conjunto das freqüências relativas observadas para um dado fenômeno estudado. Por uma conseqüência da Lei dos Grandes Números, quanto maior o tamanho da amostra, mais a distribuição de freqüência tende para a distribuição de probabilidade;
- **Medidas de tendência central:** são indicadores que fornecem uma idéia inicial de como os dados se distribuem, informando o valor (ou faixa de valores) da variável que ocorre com maior freqüência. Entre as medidas de tendência central destacam-se:
    - **Média:** é a soma de todos os resultados dividida pelo número total de casos, podendo ser considerada um resumo da distribuição como um todo;
    - **Moda:** é o evento ou categoria de eventos que ocorreu com maior freqüência, indicando o valor ou categoria mais provável;
    - **Mediana:** é o valor da variável a partir do qual metade dos casos se encontra acima dele e metade se encontra abaixo;

- **Medidas de dispersão:** são medidas da variação de um conjunto de dados em torno da média. Este tipo de medida permite identificar até que ponto os resultados se concentram ou não ao redor da tendência central de um conjunto de observações. Amplitude, desvio médio, variância, desvio padrão, erro padrão e coeficiente de variação são exemplos de medidas de dispersão. Quanto maior a dispersão, menor a concentração, e vice-versa;
- **Histogramas:** Representação gráfica da distribuição de freqüência onde o eixo horizontal representa as faixas de valores da variável e o eixo vertical representa a freqüência relativa, conforme exemplo da Figura 4.4.

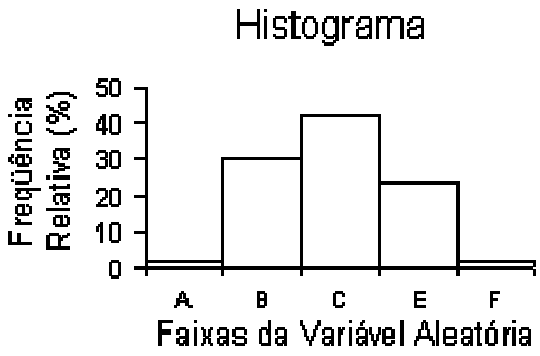


Figura 4.4 – Exemplo de histograma

Nesse processo, ferramentas de visualização dos dados podem ser utilizadas. Estas ferramentas exibem tendências nos dados, *cluters* e diferenças. Alguns dos maiores trabalhos na área de visualização estão focados em novas formas gráficas de representação dos dados, que possibilitem a descoberta de novas tendências e anomalias [Gray et al. 1997]. Existem diversos *software*s (gratuitos ou pagos) voltados para a visualização dos dados. O Microsoft Excell é um dos programas amplamente utilizados. Outros exemplos são *IBM Parallel Visual Explorer*, SAS System, *Advanced Visual Systems* (AVS) [Coutinho 2003].

Outra análise a ser realizada é a verificação da existência de dados transacionais. Dados transacionais referem-se a informações históricas de um determinado indivíduo ou evento [Quadstone 2003]. As bases de dados que contêm essa estrutura transacional não possuem o formato desejado de registro único por indivíduo, o que faz necessário o tratamento desses dados de uma forma especial.

A existência de dados transacionais tende a trazer benefícios à aprendizagem, visto que informações de comportamento são inseridas para cada indivíduo. No entanto, ausência dados transacionais não impede o prosseguimento do projeto, mesmo que para alguns tipos de

projetos (como análise de crédito) estes dados tendem a fazer muita diferença no resultado final.

Se todas as atividades realizadas até o momento não resultarem problemas (ou os problemas identificados foram resolvidos), a próxima atividade a ser realizada é a análise da representatividade dos dados para o problema proposto.

Tendo como base o problema proposto, algumas variáveis são essenciais para o sucesso do projeto. A definição do que é essencial pode ser realizada com a ajuda do especialista do domínio em conjunto com o analista de negócio. A ausência de variáveis importantes pode caracterizar um risco para o projeto.

Também deve ser analisado se a quantidade de dados obtida é suficiente. Não é fácil definir a quantidade mínima para a execução do projeto, pois esta quantidade está diretamente relacionada com o algoritmo de MD e à complexidade dos dados [Berry & Linoff 2004]. Segundo [Nguyen & Chan 2004], o desempenho de uma Rede Neural é pior quando poucos dados são disponíveis ou os dados são insuficientes. Deve ser analisado a possível re-coleta dos dados para resolver esse problema.

Em mineração de dados, quanto mais dados, melhor. No entanto, alguns aspectos devem ser considerados, como por exemplo o tamanho do conjunto de dados e sua densidade. (quantidade de registros que apresentam determinada saída). É desejável que o conjunto de dados selecionado apresente a mesma proporção de exemplos para cada saída. Uma amostra menor, porém balanceada, é preferível a uma grande quantidade de dados com uma proporção muito pequena para uma ou mais saídas [Yu et al. 2006].

O tamanho excessivo das bases de dados também influencia no tempo de treinamento algoritmo de aprendizagem. Teoricamente, quanto maior a base, maior o tempo de aprendizagem [Rezende et al. 2003]. Algumas técnicas de amostragem podem ser utilizadas, para extrair um subconjunto representativo dos dados [Fayyad & Irani 1993] [Han & Fu 1994].

Outro aspecto a ser considerado é o período correto dos dados. Um dos fatores que influenciam a definição do período é a sazonalidade (padrões de comportamento que se repetem durante os períodos da série temporal) [Berry & Linoff 2004]. Portando, deve haver uma quantidade suficiente para capturar os efeitos da sazonalidade.

A definição do período também deve considerar que dados muito antigos tendem a ser menos representativos por causa das mudanças do mercado. Portanto, deve-se se escolher o período mais próximo do momento atual para que os dados representem melhor o que de fato

acontece “hoje”. A quantidade de dados históricos é um fator de risco do projeto que deve ser analisado para tentar definir o seu impacto na solução.

A análise da quantidade de variáveis disponíveis também é importante. Em boa parte das bases de dados, a quantidade de variáveis é grande. Na dúvida do que é útil ou não, muitos usam todas as variáveis disponíveis, sem uma análise do grau de importância dessas variáveis para o problema proposto. Com a ajuda do especialista do domínio é possível definir quais variáveis são mais importantes para o negócio. Após a definição das variáveis mais relevantes, a próxima atividade é uma nova seleção de variáveis, considerando seu preenchimento e formato.

Mesmo que todas as análises realizadas até esse momento tenham sido satisfatórias, os dados ainda podem ser pobres em conteúdo e expressividade. Uma alternativa para amenizar este problema é a aquisição de dados externos. A aquisição de dados externos só deve ser realizada se existem variáveis que permitem a integração com as bases internar. Essa aquisição não é obrigatória, mas tende a trazer benefícios significativos.

As análises realizadas servem de subsídios para a definição de um parecer final com relação à qualidade dos dados e a viabilidade do projeto.

Esse é um dos processos onde existe grande iteração entre os responsáveis do projeto, principalmente com o especialista do domínio. É ele quem dá subsídios para definir se os problemas encontrados são inerentes aos dados, se podem ser corrigidos e as alternativas para resolução dos problemas.

Toda dúvida, inconsistência, problema, contato e ações realizadas nesse processo devem ser bem documentados, pois servirão de subsídio para a avaliação final da viabilidade do projeto, como também servirão de entrada para a próxima fase.

As entradas, as saídas, os responsáveis e as atividades associadas a este processo estão especificadas a seguir:

Entradas:

- Relatório de integridade dos dados;
- Relatório integração dos dados;
- Dados selecionados e integrados;
- Documentações sobre o banco de dados;

Atividades:

- Análise e exploração dos dados: Realização de análise do domínio dos dados, para analisar a sua qualidade quanto ao conteúdo;
- Parecer de representatividade dos dados: parecer final do especialista em Banco de dados quanto à representatividade dos dados para o problema proposto;

Responsáveis:

- Especialista em Banco de dados.
- Especialista em T.I.
- Especialista no domínio.
- Analista de negócios.

Saídas:

- Parecer técnico da viabilidade do projeto no com relação à qualidade dos dados;
- Scripts de validação das bases;
- Relatório de inconsistências;
- Informações dos contatos realizados.

4.2.3 Montagem de Visão

Na fase de Montagem de Visão os dados são trabalhados em sua essência.

Esta fase é importante para extrair informações de variáveis que muitas vezes são descartadas para a fase aprendizagem, devido a uma série de fatores como preenchimento normalmente considerado irrelevante, formatos não apropriados, entre outros.

Nesta fase, em especial, o conhecimento do especialista do domínio é fundamental. Entender o problema como um todo é a chave desta fase, pois somente de posse desse conhecimento é que as variáveis podem ser construídas de forma correta para enriquecer ainda mais as informações contidas nos dados.

Essa fase está dividida em quatro processos: tratamento de variáveis brutas, transformações de variáveis, agrupamento transacional e integração de dados. Os três

primeiros processos englobam ações realizadas diretamente sobre os dados. A ordem desses processos depende da natureza dos dados e do problema. Após a execução desses processos, inicia-se o processo de integração dos dados, que consolida os dados gerados nessa fase para a criação de uma tabela final que irá servir de entrada para a próxima fase. Cada processo está ligado a uma atividade de homologação dos dados, que valida as atividades realizadas.

Os principais responsáveis por essa fase são o especialista em Banco de Dados e o analista de dados. O analista de dados é responsável pela definição conceitual da visão final dos dados e pelo fornecimento de informações de como construir cada variável definida, para que o especialista em Banco de Dados possa aplicar seus conhecimentos e criar fisicamente os dados em uma tabela.

As entradas dessa fase são: toda a documentação dos dados, as bases integradas, os relatórios de integridade e integração, a visão de dados conceitual e o parecer técnico da viabilidade do projeto quanto a qualidade dados.

A Figura 4.5 define o fluxo de atividades dessa fase.

4.2.3.1 Tratamento de Variáveis Brutas

Em qualquer massa de dados, muitas variáveis possuem preenchimentos que muitas vezes são considerados fracos ou até irrelevantes para o problema. Porém, a decisão da utilização de determinada variável deve considerar uma série de fatores referentes ao seu preenchimento, como por exemplo:

- **Criação do alvo:** Determinadas bases de dados não possuem o alvo já pronto (aprendizagem supervisionada). Nesses casos, é necessário criar o alvo de acordo com as especificações do projeto;
- **Colunas com um único valor:** caracterizam variáveis constantes que não agregam conhecimento, pois não contêm informação para diferenciar as linhas. Portanto, tais variáveis devem ser retiradas da visão final dos dados. É importante, porém, verificar se informações como “NULO” estão sendo verificadas na distribuição de freqüência dessas variáveis. Em determinados problemas, o nulo pode agregar algum conhecimento (como por exemplo, quando a presença do valor nulo é igual a um valor válido, como NULO = 0), o que caracterizaria uma variável de preenchimento comum que deve ser mantida na visão final dos dados.

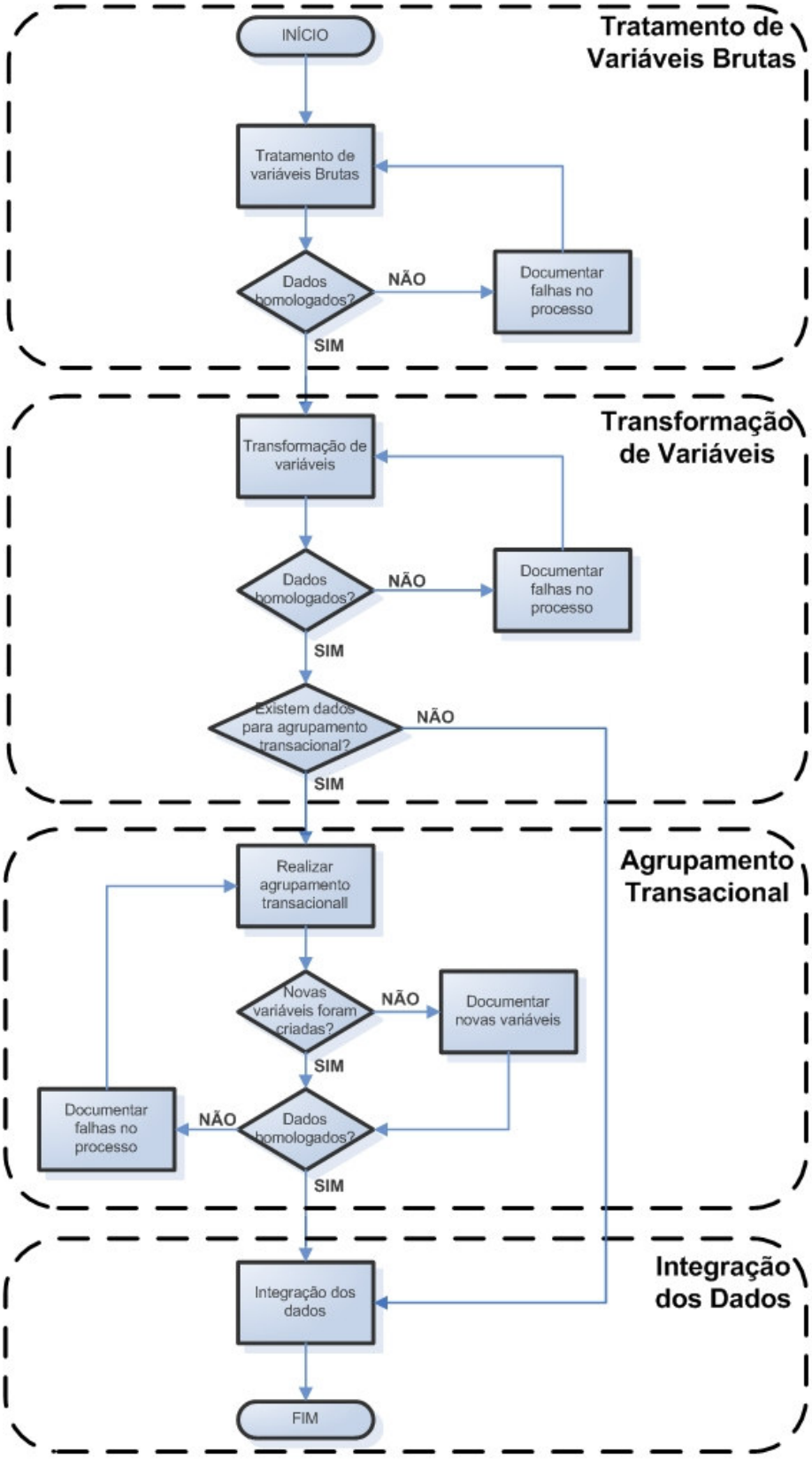


Figura 4.5 – Fluxo de atividades da fase de Montagem de Visão

Variáveis com um único preenchimento podem ser geradas no momento da seleção da população, ou seja, dependendo do subconjunto da população escolhida, algumas variáveis podem apresentar um único valor;

- **Colunas com um valor predominante:** são variáveis que possuem um baixo preenchimento de valores diferentes do valor predominante. Esse caso é freqüente quando as variáveis são raramente preenchidas. A retirada dessas variáveis deve ser analisada. Se a quantidade de linhas que apresentam valores diferentes do valor mais freqüente não for significativa, a variável pode retirada. Se existe uma quantidade significativa, é interessante relacionar os valores com o alvo (em casos de aprendizagem supervisionada), pois os mesmos podem estar fortemente relacionados a uma das classes, o que motivaria a utilização da variável.

  Independente da decisão de utilizar a variável, é fundamental identificar a causa do preenchimento tão baixo para determinados valores. Muitas vezes pode significar eventos raros relacionados ao problema, o que também motiva a sua utilização.

  Segundo [Berry & Linoff 2004], se 95% a 99.99% das linhas estão preenchidas com um único valor, é provável que esta variável seja irrelevante se não for aplicado algum tipo de tratamento;
- **Colunas com valores totalmente distintos:** Variáveis categóricas com todos os valores distintos também figuram um problema no tratamento de variáveis. Variáveis com esse preenchimento geralmente são utilizadas como atributos identificadores. Este tipo de atributo não contribui para a solução do problema, pois não possui valor preditivo e pode causar *overfitting* (ajuste em excesso ao conjunto de treinamento) [Berry & Linoff 2004] [Rezende et al. 2003]. Apesar de não apresentar valor preditivo, informações relevantes podem ser extraídas dessas variáveis através da aplicação de algumas transformações. A variável *número de telefone* é um exemplo usual deste caso. O número em si não possui informação preditiva, porém, se for extraída a parte do DDD, obtém-se informação regional, que pode apresentar valor preditivo;
- **Colunas com muitos valores distintos:** Outro tipo de atributo que geralmente não possui valor preditivo são os que contêm muitos valores distintos. Esse tipo difere do citado anteriormente, pois apesar de apresentar valores bem distintos, esses atributos ainda possuem algumas repetições em seu preenchimento. Esse tipo de atributo, se não tratado, normalmente é excluído. Porém, certas variáveis podem ser divididas em partes menores que possuem algum significado. Um exemplo é a variável *número de*

cartão de crédito. Cartões de crédito geralmente possuem de 13 a 16 dígitos que obedecem a uma estrutura de numeração[1] baseado no formato de número ISO 7812[2]. Esse formato contém um dígito identificador da bandeira (Master, Visa, Amex, etc), seis dígitos identificadores do emissor, um número de conta e um dígito verificador. A informação sobre o emissor e o número da conta podem ser descartadas, mas a informação da bandeira pode ser relevante. Outros exemplos de variáveis que podem ser tratadas dessa maneira são códigos de barras, registros de carro e CEP;

- **Colunas numéricas com conteúdo categórico:** Números geralmente representam quantidades e são bastante úteis para modelagem, pois possuem ordem e permitem a aplicação de funções aritméticas. Porém, existem variáveis que aparentemente são numéricas, mas possuem conteúdo categórico. Nesses casos, a melhor alternativa é tratar essas variáveis como categóricas, para evitar que a ordem a as propriedades aritméticas contribuam para a descoberta de padrões não correspondentes à realidade. Existem alguns métodos para transformar dados numéricos em categóricos. Estes métodos serão abordados no processo "Tratamento de Variáveis";
- **Variáveis relacionadas ao alvo:** Em problemas de aprendizagem supervisionada, uma questão comum são as variáveis preditivas serem diretamente relacionadas com a variável de saída. Se existe casualidade em um modelo, deverá ser da saída sendo influenciada pela variável preditiva, não o inverso. É importante fazer previsões do futuro baseados em dados disponíveis no passado (ou presente). Nem sempre é fácil perceber a correlação *a posteriori* da variável. Modelos com informações *a posteriori* tendem a ser inúteis ou fracos para prever informações de novos dados. A presença de variáveis a posteriori pode ser percebida quando o modelo apresenta um desempenho muito alto no conjunto de treinamento.

Para a realização deste processo, o analista de dados utiliza os resultados das análises realizadas na fase anterior, geralmente tendo como base o histograma e gráficos analíticos das variáveis.

Nesse processo é muito comum o surgimento de dúvidas relacionadas ao preenchimento das variáveis. É muito importante que o especialista do domínio esteja sempre disponível para esclarecê-las. Todo contato deve ser registrado para compor a documentação final do projeto.

[1] http://en.wikipedia.org/wiki/Credit_card_number
[2] http://www.iso.org

Após a análise das variáveis brutas, deve ser elaborado um documento que especifique as variáveis que foram retiradas (e o motivo) e as variáveis que foram criadas (informando a variável de origem). Essa informação é passada para o especialista em Banco de Dados para auxiliar na montagem de visão.

Um processo de homologação dos dados deve ser realizado para garantir que a escolha pela remoção das variáveis é coerente e o conteúdo das novas variáveis condiz com o esperado. Qualquer tipo de erro deve ser documentado, reportado e corrigido. Após a correção dos dados deve haver uma nova homologação. Só passará para o próximo processo quando a homologação não encontrar erros. Neste momento, o analista de dados dá o parecer técnico, autorizando a continuidade do projeto.

As entradas, as saídas, os responsáveis e as atividades associadas a este processo estão especificadas a seguir:

Entradas:

- Relatório de integridade dos dados;
- Relatório de integração dos dados;
- Dados selecionados e integrados;
- Documentações sobre o banco de dados.

Atividades:

- Remoção de variáveis irrelevantes: baseando-se na análise do conteúdo, o analista de dados define as variáveis que devem ser removidas por não possuírem poder preditivo significante;
- Construção de novas variáveis: o analista de dados especifica a criação de variáveis que podem ter poder preditivo e ser utilizadas na construção do modelo;
- Homologação dos dados e parecer técnico: após a remoção e construção de variáveis pelo especialista em Banco de Dados, os mesmos deverm ser homologados para evitar que alguma variável seja removida por engano ou criada de forma incorreta. Qualquer falha deverá ser reportada e corrigida, e o processo de homologação reiniciado até que não haja mais falhas.

Responsáveis:

- Analista de dados
- Especialista em Banco de dados
- Especialista em T.I.
- Especialista no domínio

Saídas:

- Relação das variáveis excluídas (e o motivo de exclusão);
- Relação das variáveis criadas (e suas variáveis de origem);
- Scripts de geração de variáveis;
- Base com variáveis brutas criadas;
- Relatório de homologação dos dados;
- Informações dos contatos realizados.

4.2.3.2 Transformação de Variáveis

As variáveis selecionadas no processo anterior podem apresentar um formato que não é apropriado para o algoritmo de mineração de dados. Algumas transformações devem ser aplicadas sobre estas variáveis a fim de resolver este problema.

Neste processo, o analista de dados é responsável pela análise de cada variável e pela proposição de novas transformações, baseando-se no preenchimento e formato das variáveis. O especialista do domínio e o especialista de T.I. são consultores desse processo.

As transformações mais comumente realizadas são:

- **Tratamento de variáveis do tipo Data/Hora:** A grande maioria dos algoritmos de mineração de dados não trabalham com o tipo "data"/"hora". No entanto, a presença da informação temporal é muito importante para a resolução de determinados problemas [Berry & Linoff 2004]. Várias informações podem ser extraídas de variáveis tipo data/hora, como por exemplo: dia do mês/ano, período do dia, mês, estação, indicação de feriado ou dia comum, tempo desde um determinado período. Um exemplo bastante comum é a criação da variável *idade* a partir da variável d*ata de Nascimento*. A nova variável é obtida através da diferença em anos entre a data de nascimento do indivíduo e uma data de referência específica do problema que está sendo resolvido. Seguindo essa lógica, a grande maioria das variáveis do tipo

data/hora podem ser transformadas, sempre considerando informações sobre o domínio do problema;

- **Categoria "OUTROS":** Em algumas variáveis categóricas, os valores podem estar distribuídos de maneira não uniforme. Isso significa que a distribuição dos valores está concentrada em conjunto restrito de categorias, enquanto diversas outras categorias possuem uma quantidade pequena de exemplos. Nesse caso, uma técnica comumente usada é a criação de uma categoria chamada "Outros", que representa o agrupamento de todas as categorias menos populosas [Pyle 1999] [Witten & Frank 2005].
- **Variáveis indicativas de ausência/presença (flags):** Uma alternativa para tratamento de variáveis com o preenchimento concentrado em um único valor é a criação de um indicativo de presença ou ausência de informação. Esse tipo de abordagem pode ser utilizado tanto para variáveis numéricas quanto para variáveis categóricas. Uma vantagem desse tratamento é a possibilidade de usar variáveis que, por causa do seu preenchimento pobre, normalmente seriam retiradas do modelo;
- **Tratamento de endereço:** Variáveis de endereço normalmente não trazem benefícios se forem utilizadas em sua forma bruta, pois a quantidade de valores distintos é muito grande. Porém, é possível extrair informações geográficas (como Estado, cidade, bairro e CEP) que normalmente apresentam alto poder preditivo;
- **Mapeamento simbólico:** Uma prática comum é a transformação de atributos simbólicos em outros atributos também simbólicos. Essa prática agrupa exemplos de variáveis categóricas em um grupo que as represente. Um exemplo é o agrupamento de produtos. Para um determinado problema, pode ser mais importante mapear os diversos produtos (Ex.: camiseta masculina, saia, carrinho de bebê) para o tipo de departamento aos quais os produtos pertencem (seção masculina, seção feminina, seção infantil, respectivamente). Esse tipo de agrupamento é totalmente dependente do conhecimento do domínio [Bigus 1996]. Pode ocorrer em diversos níveis de granularidade;
- **Combinação de variáveis:** O resultado da combinação de variáveis pode ser mais significativo do que a apresentação destas variáveis em separado. Como exemplo, pode ser criada uma variável que represente a área de um retângulo (Área = C x L) a partir de variáveis de comprimento (C) e largura (L). Esse tipo de combinação é

fortemente baseada no conhecimento do domínio. Portanto o analista do domínio pode fornecer o suporte necessário para essas transformações.

O analista dos dados é responsável pela criação de um documento que especifica as variáveis que devem ser criadas para contemplar as transformações desejadas. Baseando-se neste documento, o especialista em Banco de Dados desenvolve os scripts de geração das novas variáveis.

Após a realização das transformações ou criações de variáveis, deve ser realizada a homologação dos dados para garantir que as atividades deste processo foram executadas corretamente. Qualquer erro detectado deve ser documentado, reportado e corrigido. Após a correção dos dados, deve haver uma nova homologação. Esse ciclo se repete até que a homologação não detecte erros.

As entradas, as saídas, os responsáveis e as atividades associadas a este processo estão especificadas a seguir:

Entradas:

- Documentações sobre o banco de dados;
- Relação das variáveis excluídas e criadas;

Atividades:

- Gerar relatório de transformação das variáveis: relatório com todas as variáveis que a ser criadas, especificando o tipo de transformação necessária;
- Construção de novas variáveis;
- Homologação dos dados e parecer técnico.

Responsáveis:

- Analista de dados
- Especialista em Banco de dados
- Especialista em T.I.
- Especialista no domínio.

Saídas:

- Relação das variáveis criadas (e suas variáveis de origem);
- Relação das variáveis transformadas;

- Scripts de geração de variáveis;
- Base de dados com transformações de variáveis;
- Relatório de homologação dos dados;
- Informações dos contatos realizados.

4.2.3.3 Agrupamento Transacional

Dados transacionais normalmente estão disponíveis nas organizações. Estes dados representam informações detalhadas sobre algum tipo de área de negócio [Burns et al. 2006]. Eis alguns exemplos de dados transacionais em diversas áreas:

- Bancos: créditos e débitos em contas;
- Telecomunicações: ligações telefônicas;
- Comércio: compras, pagamentos;
- Comércio virtual: cliques em links e visitas a páginas;
- Aviação: vôos, compra de passagens;

Dados transacionais normalmente não apresentam o formato que deve ser submetido aos algoritmos de mineração de dados (ou seja, informação de um indivíduo por linha), pois esses dados contêm informações do comportamento indivíduo em diversos momentos do passado.

Esse processo só é realizado se, para o problema em questão for verificado a presença de dados transacionais. Se não houver dados transacionais, o projeto segue para o próximo processo (Integração dos Dados).

Para inserir informações comportamentais, é necessária a aplicação de uma série de tratamentos nos dados transacionais. Esses tratamentos incluem medidas de agrupamento, tendências, indicativos de freqüência e mudança comportamental, podendo ser subdivididos de acordo com o tempo e o tipo [Bain et al. 2002]. Para que esses tratamentos possam ser realizados, é necessário que os dados estejam na mesma granularidade.

Este processo normalmente é trabalhoso, pois não existe uma fórmula para a realização dos agrupamentos. O princípio básico é tentar agrupar informação de diversas maneiras, usando sempre o conhecimento do domínio como guia. Portanto, o especialista no domínio e o analista de negócio são essenciais para a definição das variáveis que podem ser criadas e sumarizadas. Este processo, por ser dependente do domínio, é muito mais arte do que ciência [Quadstone 2003].

O conceito de janela de observação é muito importante para a realização desse processo. O objetivo básico de mineração de dados é desenvolver um modelo que tenha a habilidade de prever um comportamento baseando-se em um comportamento do passado. O período usado na construção desse modelo é chamado de "Janela de Observação". Como pode ser observado na Figura 4.6, a janela está dividida em dois momentos: pré-período (o período em que a informação transacional ou comportamento é capturada) um período posterior (período em que são capturados os resultados que se deseja prever). Para a criação do modelo, é necessário que todos os dados sejam obtidos no pré-período juntamente com a informação do resultado (alvo, que é observado no período posterior).

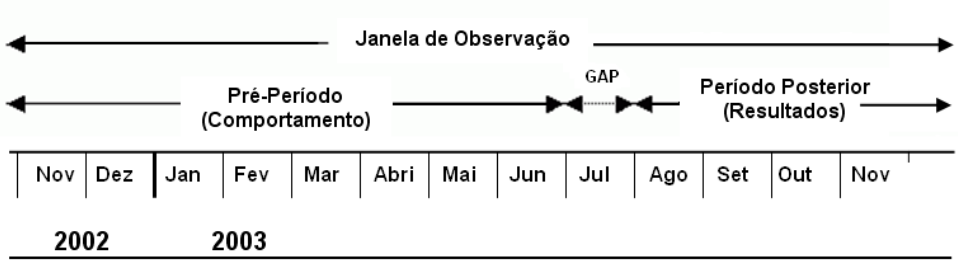


Figura 4.6 – Janela de observação

Considerando a aquisição de dados de acordo com a janela de observação, eis alguns tipos de agrupamentos transacionais comumente realizados:

- **Medidas de agregações:** São medidas mais simples. Medidas como quantidade e soma são utilizadas para agrupar informações de um indivíduo, considerando todo o seu passado. Exemplos: quantidade de transações realizadas, valor total gasto em passagens aéreas.
- **Funções de tendência:** O comportamento de um indivíduo é comumente medido através do uso de funções de tendência. Medidas como mínimo, máximo e médio tendem a resumir as informações referentes a um indivíduo no passado. Exemplos: valor médio, mínimo e máximo de fatura telefônica.
- **Indicativos de freqüência:** Indicativos de freqüência capturam informações importantes que se repetem ao longo do tempo. Também indica tendências. Exemplo: *tipo de produto mais freqüente*.
- **Variação de tempo e tipo:** Todos os agrupamentos citados anteriormente podem ser variados de acordo com o tempo e o tipo. Geralmente, esse tipo de agrupamento é eficiente [Quadstone 2003]. Exemplos: *total de compras realizadas nos últimos três (3) meses*, *total gasto em alimentação*, *total gasto em eletrônicos*.

- **Mudança comportamental (índices):** Mesmo dividindo essas medidas em séries de tempo, em determinadas situações o mais importante é a estabilidade das medidas e a direção de possíveis mudanças. Exemplos de índices que refletem uma mudança comportamental: *Máximo Gastos Mensais / Média de Gastos Mensais*, *Último Gasto Mensal / Média de Gastos Mensais*.

O analista de dados gera um relatório contendo a especificação de todas as variáveis a serem criadas de acordo com o agrupamento transacional de informações. De posse desse relatório, o especialista em Banco de Dados desenvolve os scripts de geração da nova massa de dados, contemplando as variáveis resultantes dos agrupamentos transacionais. O especialista de T.I. também é consultor desse processo.

Após a geração da nova massa, deve ser realizada a homologação dos dados, para garantir que as atividades deste processo foram executadas corretamente. Qualquer erro detectado deve ser documentado, reportado e corrigido. Após a correção dos dados, deve haver uma nova homologação. Esse ciclo se repete até que a homologação não detecte erros.

As entradas, as saídas, os responsáveis e as atividades associadas a este processo estão especificados a seguir:

Entradas:

- Documentações sobre o banco de dados;
- Base de dados que serão utilizadas para o agrupamento transacional;
- Relação das variáveis excluídas, criadas e transformadas dos processos realizados anteriormente;

Atividades:

- Gerar relatório de criação das variáveis resultantes dos agrupamentos transacionais;
- Criação das variáveis;
- Homologação dos dados e parecer técnico.

Responsáveis:

- Analista de dados.
- Especialista em Banco de Dados,
- Especialista em T.I.

- Especialista no domínio.

Saídas:

- Relação das variáveis criadas;
- Scripts de geração de variáveis;
- Relatório de Homologação dos dados;
- Informações dos contatos realizados.

4.2.3.4 Integração dos Dados

Esse último processo tem por finalidade integrar todas as variáveis que foram transformadas ou criadas pelos demais processos anteriores da fase de Montagem de Visão. O resultado dessa integração é uma tabela desnormalizada (linha x coluna), contendo os dados finais. É importante citar que a construção de uma tabela desnormalizada se dá pelo fato de que a maioria dos algoritmos de mineração de dados tem como entrada esse tipo de estrutura.

Algumas das análises citadas no processo 4.2.2.2 podem ser realizadas para garantir que não exista redundância nas variáveis criadas.

O responsável por essa fase é o especialista em Banco de Dados, que irá agrupar todas as variáveis criadas na fase de Montagem de Visão e gerar a tabela desnormalizada.

As entradas, as saídas, os responsáveis e as atividades associadas a este processo estão especificadas a seguir:

Entradas:

- Documentações sobre o banco de dados;
- Relação das variáveis excluídas, criadas e transformadas dos processos realizados anteriormente;
- Todas as bases geradas nos processos anteriores;

Atividades:

- Integração dos dados: o especialista em banco de dados deverá integrar todas as bases geradas nos processos anteriores. Após essa integração pode ser realizada uma análise de redundância entre as variáveis.

Responsável:

- Especialista em Banco de dados.

Saídas:

- Tabela final desnormalizada;
- Relatório de redundância;
- Visão final (conceitual) dos dados;

4.2.4 Tratamento dos Dados

Ao chegar nessa fase, uma massa de dados estará criada, tendo sido homologada e pronta para a sua utilização. Porém, para a continuação do projeto, pode haver a necessidade de algum tratamento de dados ou transformação de formato de apresentação de atributos.

A Fase de Tratamento dos Dados é a fase que foca as técnicas que podem ser usadas para corrigir alguns problemas com relação à qualidade dos dados que ainda são existentes. Após a criação da massa final de dados, é comum verificar problemas referentes à natureza dos dados e verificar qual o formato ideal para a técnica que será utilizada.

Essa fase é dividida em quatro processos: Limpeza, Redução de Dimensionalidade de Variáveis, Casamento de Padrões e Mudança de Formato.

O responsável por essa fase é o analista dos dados, o qual está familiarizado com as técnicas de tratamento de dados, em conjunto com o Especialista de Banco de Dados, sempre com o suporte do Especialista do Domínio.

Apesar das atividades Aplicação do Algoritmo de Mineração de Dados, Avaliação dos Resultados e Definição de Fase de Retorno estarem presentes no fluxograma, as mesmas não serão detalhadas nesta dissertação, pois o escopo da metodologia *DMBuilding* inicia no Entendimento do Problema e finaliza na geração da base de dados que será utilizada na fase de MD. Estas atividades foram inseridas no fluxograma da fase de Tratamento dos dados apenas para contextualizar a fase final desta metodologia com o restante do processo de KDD.

A Figura 4.7 define o fluxo de atividades dessa fase:

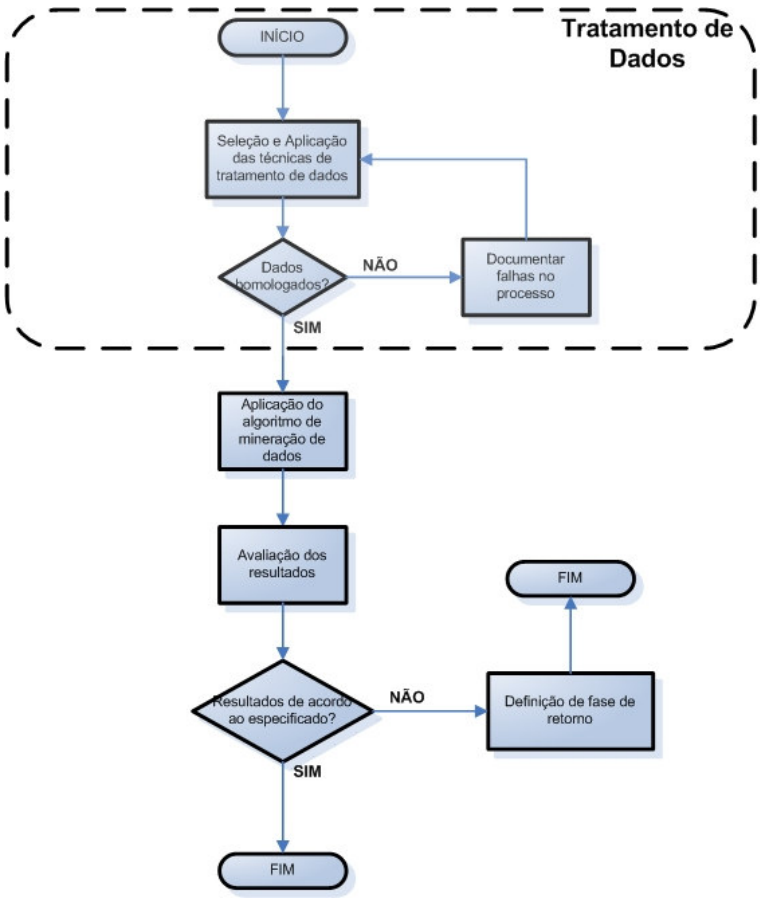


Figura 4.7 – Fluxo de atividades da fase de Tratamento de Dados

4.2.4.1 Limpeza

Mesmo após a aplicação de técnicas para tratamento de variáveis brutas, muitos atributos podem continuar apresentando algum tipo de ruído ou inconsistência, que pode ser resolvida ou amenizada através de técnicas de limpeza de dados. Esses problemas podem ser originados devido à falha no processo de geração, aquisição, integração e transformação dos dados [Monard & Baranauskas 2003].

Os principais problemas encontrados nessa fase de limpeza são: tratamento de valores ausentes (*missing data*) e tratamento de dados espúrios (*outliers*).

Para identificar quais os atributos que possuem um desses dois problemas citados acima, as técnicas de visualização (citadas na seção 4.2.2.2) podem ser utilizadas para auxiliar nesse processo. A utilização de ferramentas de visualização de dados se faz muito útil nessa etapa.

4.2.4.1.1 Valores Ausentes (*Missing Data*)

Valores ausentes (*missing data*) é um dos problemas mais comuns nos dados atualmente, especialmente em bases de dados com muitos atributos. A ausência de informação raramente traz algum benefício, o que traz a necessidade do tratamento.

Algumas técnicas de MD conseguem trabalhar com valores ausentes, sem a necessidade de substituição por algum valor. Em algumas técnicas, os valores ausentes são ignorados e outros a instância (registro) toda é ignorada. Algumas técnicas simplesmente não conseguem tratar atributos que possuam valores ausentes. Por isso, faz-se necessário a aplicação de técnicas que evitem a adição de *Bias* (tendências) ou distorções nos dados [Pyle 1999].

Faz-se muitas vezes necessário o uso de alguma técnica de substituição de valores ausentes. Em todo caso, muita cautela deve ser tomada, porque, mesmo que essa substituição seja interessante para algumas análises e para o desempenho de algumas técnicas de IA, um valor informado manualmente tem suas chances reduzidas de ser um valor confiável (por causa da estimativa) [Dasu & Johnson 2003] [Han & Kamber 2006].

Várias são as técnicas que podem ser utilizadas para o tratamento de *missing*, entre os quais podemos citar:

1. **Ignorar a linha:** essa técnica geralmente é utilizada quando a variável de classe está ausente (assumindo que seja um problema onde a classe é conhecida). Este não é um método eficiente, a menos que o tupla contenha muitos atributos que também estão ausentes. É uma técnica fraca quando o percentual de valores ausentes por atributos varia consideravelmente. Ignorar a linha pode fazer com que os dados sejam diminuídos drasticamente (de 30% a 70%) [Dasu & Johnson 2003] [Han & Kamber 2006].
2. **Preenchendo com valor manual:** Em geral, essa prática é custosa e não é praticável dado uma massa de dados muito grande com vários valores ausentes. Porém, pode ser utilizada se tiver auxílio de algum software para realizar o preenchimento automático [Han & Kamber 2006].
3. **Uso de uma constante:** substituir todos os valores ausentes por uma constante, como por exemplo o rótulo "Desconhecido". Se os valores forem substituídos, por esse rótulo, então a técnica de programação poder pensar sem engano que esses exemplos iguais formam um conceito interessante, porque todos têm um valor em comum.

Apesar de ser um método simples, ele não perfeitamente seguro [Han & Kamber 2006] [Larose 2005].

4. **Usando a média:** Usar o valor da média dos valores de toda a massa de dados para substituir em caso de valores ausentes. É necessário assumir, nesse caso, que os valores ausentes possuem a mesma distribuição dos valores não ausentes [Dasu & Johnson 2003] [Han & Kamber 2006] [Larose 2005] [Populus et al. 1998] [Pyle 1999].
5. **Usando a média por exemplos da mesma classe:** quando se trata de problemas de classificação, pode ser verificado a média dos valores dos exemplos pertencente a mesma classe daquela da tupla e substituir o valor ausente por essa média [Han & Kamber 2006].
6. **Usar o valor mais provável:** Refletem em todos os valores plausíveis para os valores ausentes, ao invés de um único valor fixo. Essa imputação irá resultar em diversas bases de dados, que são analisadas e os resultados combinados. Algumas são as técnicas para imputar os valores mais prováveis, como regressão, ferramentas baseadas pelo uso do formalismo Bayesiano ou indução de árvores de decisão [Dasu & Johnson 2003] [Han & Kamber 2006] [Larose 2005] [Pyle 1999].
7. **Usando conhecimento do domínio:** em alguns problemas, os valores ausentes podem ter algum significado específico, que poderá ser verificado pelo especialista no domínio. Um exemplo disso é quando a ausência de valores no atributo "Quantidade_de_Compras" significa que o cliente não possui compras, ou seja, 0 (zero) compras. Nesse caso, os valores ausentes podem ser substituídos por zero [Boscarioli 2005].
8. **Usando outros atributos:** A utilização de atributos correlacionados com o atributo que tem valor ausente pode ser utilizado. Um exemplo disso é se um CEP está ausente, pode ser utilizado um CEP próximo a localização da rua ou do Bairro do atributo que tem um valor ausente. Essa não é uma técnica simples, pois necessita do conhecimento domínio e na maioria das vezes a criação de algum software para fazer essas checagens [Larose 2005].

A maioria dos métodos propostos na literatura propõe técnicas utilizando regressão, *clustering*¸ algoritmos genéticos, métodos estatísticos, regras de associação, formalismo Bayesiano entre outros. Um importante argumento a favor dessas abordagens é que os atributos freqüentemente possuem correlações entre si, que podem ser utilizadas para criar um modelo preditivo. Seguem alguns exemplos de métodos de predição de valores ausentes:

[Batista & Monard 2003a] [Batista & Monard 2003b] [Grzymala-Busse & Hu 2000] [Lee et al. 1976] [Little & Murphy 1987] [Ragel & Cremilleux 1998] [Schafer & Olsen 1998] [Tseng et al. 2003] [Wei & Tang 2003].

4.2.4.1.2 Identificação de Dados Espúrios (*Outliers*)

Dados espúrios (*outilers*), também conhecidos como dados berrantes ou erros aleatórios, são valores numéricos extremos que estão fora do domínio do que podem ser considerados valores válidos [Boscarioli 2005]. A sua identificação é importante porque eles tendem a representar erros nas entradas dos dados. Mesmo que um *outlier* seja um valor válido e não um erro, alguns métodos estatísticos são sensíveis à presença dos mesmos e acabam gerando resultados instáveis [Larose 2005].

Existem algumas formas de identificar os *outliers*. Uma forma é através da utilização de histogramas para determinado atributo a qual deseja analisar. Valores muito diferentes da distribuição normal tendem a aparecer nos extremos do histograma, o qual facilita essa visualização [Larose 2005].

Para a identificação, algoritmos de *clustering* [Duda et al. 2002] [Jain et al. 2000] [Theodoridis & Koutroumbas 1999] podem ser utilizados, pois como estes algoritmos organizam valores similares em grupos ou *clusters*, intuitivamente, os valores que estão fora do conjunto de *clusters* podem ser considerados *outliers* [Han & Kamber 2006].

É comum que o especialista do domínio forneça suporte para identificar o que seria um *outlier* correto [Boscarioli 2005] [Redpath & Sheard 2005].

Podem ser citadas algumas técnicas para o tratamento de *outliers*, entre as quais:

- Regressão: os dados podem ser suavizados por uma função que prevê o valor de um atributo, através de outro atributo (Regressão Linear) ou de um conjunto de atributos (Regressão Multi Linear) [Han & Kamber 2006].
- *Clipping*: Está técnica é utilizada de duas formas [Abeles et al. 2003]:
    - Primeiramente, é determinado qual a faixa de valores considerados aceitáveis para o atributo (mínimo e máximo). Quando um valor *outlier* é encontrado, seu valor é modificado para o valor mais próximo entre a faixa de valores aceitáveis. Essa técnica se chama *winsorizing*.
    - Ignorando do modelo os valores fora da média. Essa técnica é chamada de *trimming*.

Se for verificado que o valor identificado como *outlier* foi causado por causa de algum erro de coleta, uma alternativa é tentar a correção ou a nova coleta de dados. Caso não seja possível essa correção, uma das alternativas é a utilização das técnicas utilizadas para tratamento de *missing*, citadas na Seção 4.2.4.1.1 [Pyle 1999].

É necessária a análise cuidadosa do tipo de informação que pode ou não ser tratada como *outlier*. Em algumas aplicações, como detecção de fraudes, a falta de informação ou a presença de *outliers* pode ser uma indicação valiosa de padrões interessantes, por isso não deveria ser removida ou preenchida com algum valor [Adriaans & Zantinge 1996] [Han & Kamber 2006].

4.2.4.2 Redução de Dimensionalidade

De posse dos dados limpos, uma atividade muito comum é a Redução de Dimensionalidade de Variáveis (seleção de atributos). Neste momento, grande pode ser o número de variáveis disponíveis para a aplicação de algum algoritmo de MD. Comumente, nem todos os atributos são necessários para o treinamento, e uma representação reduzida dos dados tende a produzir os mesmos resultados analíticos, porém com um tempo de aprendizagem menor [Hand & Mannila 2001]. Em alguns casos, as restrições de espaço em memória ou tempo de processamento, uma grande quantidade de atributos tende a inviabilizar a utilização de algoritmos de MD [Rezende et al. 2003].

A alta dimensionalidade dos dados prejudica o desempenho dos algoritmos, e, portanto, é comum a utilização de técnicas de redução de dimensionalidade para reduzir a complexidade e melhorar o desempenho em mineração de dados [Razente & Traina 2003]. A maioria dos métodos de aprendizagem tentam selecionar os atributos relevantes e ignorar os irrelevantes ou redundantes, mas na prática, o desempenho dos métodos normalmente pode ser melhorada pela seleção de atributos [Witten & Frank 2000].

Existem diversas formas de realizar a redução de dimensionalidade. A forma padrão é através do uso do conhecimento do domínio, técnica essa muito custosa (pois o especialista pode deparar com uma base contendo centenas de variáveis) e tendenciosa a erro (pois apesar do conhecimento, o especialista não pode ser a certeza de quais atributos são realmente importantes para atingir os objetivos) [Han & Kamber 2006].

Para analisar de forma automática a relevância dos atributos, muitos estudos foram propostos. Estes métodos aplicam alguma medida para quantificar a relevância de um determinado atributo com relação a uma classe ou conceito (Ex:. ganho de informação, índice

de Gini e coeficiente de correlação) [Han & Kamber 2006]. Seguem alguns exemplos de métodos de análise de relevância de atributos:

- Existem dois algoritmos de indução bastante conhecidos, tal como árvores de decisão e naïve-bayes. No algoritmo de árvore de decisão C4.5 [Quinlan 1993], quando uma árvore é construída, são selecionados somente os atributos que foram utilizados para gerar a árvore. Os classificadores Naïve-Bayesian [Domingos & Pazzani 1997] usam regras Bayesianas para computar a probabilidade de cada classe dada a instância, assumindo que os atributos são condicionalmente independentes [Soibelman & Kim 2002];
- Análise de Componentes Principais (PCA), que tem por objetivo encontrar uma combinação linear padronizada das variáveis originais de tal forma que a variância seja maximizada. A análise de componentes principais busca algumas combinações que possam sumarizar os dados e ao mesmo tempo minimizar a perda de informações [Haykin 1999];
- Outro critério muito utilizado é a aplicação da Análise Fatorial, que é uma técnica de análise multivariada que objetiva explicar as correlações existentes entre um conjunto grande de variáveis em termos de um conjunto de poucas variáveis aleatórias não observáveis, denominadas fatores. Quanto mais fortes forem as correlações entre algumas variáveis dentre o grupo inicial, mais nítida é a visualização do fator gerado. Variáveis agrupadas num mesmo fator possuem alta correlação, enquanto que variáveis de fatores distintos possuem baixa correlação [Cunico 2005];
- Métodos de aprendizagem baseados em exemplos também podem ser usados para selecionar atributos. Com uma amostra aleatória de exemplos, é feito uma análise verificando exemplos próximos (baseando-se na classe). Se um exemplo próximo pertencer à mesma classe e tiver um valor diferente para um determinado atributo, o atributo tende a ser irrelevante e um peso associado ao atributo deve ser decrementado. Por outro lado, se um exemplo próximo pertencer a uma classe diferente e tiver um valor diferente para um determinado atributo, o atributo tende a ser relevante e seu peso é incrementado. Depois da execução repetida desse processo, somente os atributos com pesos positivos são selecionados [Witten & Frank 2000].

Além da redução de atributos, outra preocupação é com a quantidade de registros da massa de dados. Muitas técnicas, como RNA e técnicas não lineares (como Regressão Logística) não tem um bom desempenho com muitos registros [Dasu & Johnson 2003]. A redução do número de exemplos deve ser feita a fim de manter as características do conjunto de dados original, por meio da geração de amostras representativas dos dados [Rezende et al. 2003]. A abordagem mais utilizada para redução do número de exemplos é a amostragem aleatória, pois este método tende a produzir amostras representativas.

Como os conceitos a serem aprendidos pelos métodos de mineração de dados são dependentes da qualidade dos dados, essa atividade deve ser realizada com extremo cuidado [Monard & Baranauskas 2003].

4.2.4.3 Casamento de Padrões (*String Matching*)

Casamento de Padrões (*string matching*) é uma técnica utilizada para encontrar as ocorrências de uma *string* particular (chamada de padrão) em outra *string* (chamada de texto) [Charras & Lecroq 1997]. Tal técnica deve ser capaz de reconhecer ocorrências sobrepostas.

Um dos algoritmos de casamento de padrão mais utilizados é o algoritmo de Força Bruta. Este algoritmo realiza uma comparação que inicia no começo do texto. Os caracteres do padrão são comparados um a um com os caracteres correspondentes no texto até que uma diferença seja detectada ou uma ocorrência do padrão seja encontrada no texto. Em seguida, desloca-se uma posição do texto para a direita e uma nova comparação é realizada. Este processo é repetido n vezes ($n = t - p + 1$, onde t é igual ao tamanho do texto e p é o tamanho do padrão), o que significa que o texto foi analisado por completo. Este algoritmo é capaz de identificar ocorrências sobrepostas. Considerando o padrão P = "baba" e o texto T = "cbbababaababacaba", como exemplos, esta técnica encontraria três ocorrências do padrão P no texto T iniciando nas posições 3, 5 e 10. Neste caso, existe uma sobreposição entre as ocorrências 3 e 5.

Além do algoritmo de Força Bruta, outros foram propostos com a mesma finalidade, entre os quais se destacam: casamento de padrões com autômato finito, *Knuth-Morris-Pratt, Boyer-Moore, Boyer-Moore-Horspool* e *Karp-Rabin* [Mitra 2003].

Os algoritmos citados até o momento realizam um casamento exato de padrões, ou seja, uma ocorrência do padrão é detectada no texto apenas se o padrão for exatamente igual à parte do texto.

Algoritmos de casamento parcial de padrões também foram propostos. Estes algoritmos encontram partes do texto que são similares a um dado padrão. Por similaridade, entende-se a existência de um número limitado de diferenças entre o padrão e partes específicas do texto. Existem muitas definições de "diferença", entre as quais se destacam:

- ***Hamming***: A distância de *Hamming* deve ser aplicada entre dois textos de tamanhos iguais. Seu valor é igual ao número de posições cujos valores são diferentes. Por exemplo, a distância de *Hamming* entre os textos "cajá" e "caju" é igual a 1 (um), pois eles diferem apenas na última posição [MacKay 2003].
- ***Levenshtein***: Um texto X pode ser transformado em um texto Y através da aplicação de operações de edição (inserção, remoção e substituição) em cada caractere do texto. A distância de *Levenshtein* entre o texto X e Y é igual ao número mínimo de operações de edição necessário para transformar X em Y, ou vice-versa. Por exemplo, a distância de *Levenshtein* entre os textos "cesta" e "festas" é igual a 2 (dois), pois se "c" for substituído por "f" e "s" for inserido ao final do primeiro texto, "cesta" será convertida em "festas" [Navarro 2001].
- **Edição**: É similar à distância de *Levenshtein*. A diferença entre estas medidas é que na distância de edição só é permitido a aplicação de operação de inserção e remoção. Portanto, a distância de edição entre os textos X e Y é igual ao número mínimo de operações de inserção e remoção necessário para transformar o texto X no texto Y, ou vice-versa. Por exemplo, a distância de edição entre os textos "cesta" e "festas" é igual a 3 (três), pois se "c" for removido, "f" for inserido no início e "s" for inserido no final, "cesta" será convertida em "festas" [Mitra 2003].

As técnicas de *string matching* são também utilizadas para tratamento de variáveis categóricas, para realizar a correção automática de erros de digitação. Exemplos como "Recife", "Ricife", "Recif" podem representar o mesmo elemento e com o a utilização de um algoritmo de *string matching*, essa correção é possível.

4.2.4.4 Mudança de Formato

Esta é uma atividade totalmente dependente do algoritmo de mineração de dados que será utilizado para o projeto. Consiste na aplicação de técnicas para transformar os dados em formatos apropriados para a aplicação do algoritmo de mineração de dados. Caso os dados

não estejam num formato adequado, a aprendizagem pode ser prejudicada, resultando em modelos não representativos [Gately 1996].

Como mudança de formato, podemos citar três tipos: mudança de Escala e Normalização, Discretização e Codificação Binária.

Para garantir a qualidade das atividades realizadas, é importante a realização do processo de homologação dos dados. Só passará para a próxima atividade quando a homologação não encontrar erros.

4.2.4.4.1 Escala e Normalização

Um atributo é normalizado usando uma função matemática, que converte o valor de um atributo para uma nova faixa de valores, como 0 a 1 [Myatt 2007]. A normalização é útil para algoritmos de classificação envolvendo RNA, ou algoritmos com medidas de distância (classificação por vizinhos mais próximos - *nearest-neighbor*) [Yu et al. 2006]. A normalização assegura que os valores de um atributo estejam dentro de um determinado intervalo, minimizando, desta forma, os problemas oriundos do uso de unidades e dispersões distintas entre os atributos. Para métodos baseados em distância, quando normalizados, os atributos estão na mesma ordem de magnitude, e por isso é obtida uma medida de distância confiável entre os diferentes atributos. [Han & Kamber 2006].

A normalização pode melhorar a precisão e a eficiência dos algoritmos de mineração que lidam com medidas de distância (Ex.: RNA, classificadores K-Nearest Neighbor ou KNN, e clustering), pois quando os atributos são escalados para mesma ordem de magnitude, uma medida de distância confiável entre os diferentes exemplos é obtida. Caso contrário, os atributos que possuem uma ordem de magnitude maior exercerão mais influência [Adriaans & Zantinge 1996].

Esse tipo de transformação não é recomendada para métodos de representação simbólica (como árvores de decisão e regras de decisão), pois a normalização tende a diminuir a compreensibilidade do modelo gerado por tais algoritmos [Batista 2003].

Existem muitos métodos de normalização, como podemos citar:

- Min-Max; realiza uma normalização linear dos dados originais, através da Equação 4.4:

$$an = MinMax(a) = \frac{a - \min A}{\max A - \min A} \times (new \max A - new \min A) + new \min A$$

(4.4)

Onde ***an*** é o novo valor normalizado, ***a*** é o valor original da variável, ***minA*** e ***maxA*** são os valores mínimos e máximos do atributo, ***newminA*** e ***newmaxA*** são os valores mínimo e máximo do intervalo da normalização. Essa é uma forma muito utilizada. Normalização de *Min-max* preserva o relacionamento entre os dados originais. Poderá constatar um erro de "fora do limite" se no futuro, um caso de entrada para a normalização se enquadrar fora da faixa original de valores para *A* [Han & Kamber 2006].

- Z-Score: normaliza os valores baseado na média de um conjunto de dados. Neste método, os valores para um atributo *A* são normalizados baseados na média e no desvio padrão de *A*. O valor, *v*, de *A* é normalizado para *v'*, através da equação 4.5:

$$v' = \frac{v - \bar{A}}{\sigma_A} \qquad (4.5)$$

Onde $\bar{A}$ e σ_A são, respectivamente, a média e o desvio padrão do atributo A. Este método de normalização é útil quando os valores de mínimo e máximo não são conhecidos, ou quando existem *outliers* que dominam a normalização do *min-max* [Myatt 2007].

- Escala Decimal: normaliza através da mudança do ponto decimal do atributo *A*. O número de pontos decimais movidos depende do máximo absoluto de *A*. Um valor, *v*, do atributo *A* é normalizado para *v'* através da Equação 4.6:

$$v' = \frac{v}{10^j} \qquad (4.6)$$

onde *j* é o menor inteiro tal que $Max(|v'|) < 1$ [Han & Kamber 2006].

A normalização pode alterar os dados originais, especialmente os últimos dois métodos citados. É necessário salvar os parâmetros de normalização (como a média ou o desvio padrão, quando utilizado a normalização *z-score*) para que no futuro os dados possam ser normalizados de maneira uniforme [Han & Kamber 2006] [Myatt 2007].

4.2.4.4.2 Discretização

Através da conversão de dados contínuos em discretos, reduz o número de valores de um atributo contínuo, dividindo a amplitude do atributo em intervalos. Os rótulos dos intervalos substituem os valores.

Essa conversão é necessária quando um determinado atributo contínuo tem de ser considerado como um atributo de intervalo ou de razão. Essa técnica também é chamada de Suavização de dados (*Data Smoothing*). Além disso, certos métodos de mineração de dados só trabalho com dados categóricos, o que faz necessário a conversão de todos os atributos contínuos em valores discretos [Han & Kamber 2006].

Essa técnica geralmente é utilizada quando algoritmos de aprendizagem simbólica são utilizadas, que normalmente só trabalham com dados categóricos [Berka 2002]. Esta técnica melhora o entendimento para alguns algoritmos, porém pode causar perda de informação [Amorim 2004].

Podem ser citadas algumas técnicas de discretização, como:

- ***Binning***: técnica de divisão de um atributo baseada em um número específico de subgrupos. Podem ser utilizados métodos de divisão por mesma freqüência (subgrupos com a mesma quantidade de exemplos) e então substituindo todos os valores dos subgrupos pela média ou mediana dos mesmos. Essa técnica não utiliza a informação do alvo e é uma técnica não supervisionada de discretização. É sensível ao número de subgrupos especificados pelo usuário, assim como pela presença de *outliers* [Witten & Frank 2005];
- **Análise de histogramas** assim como *Binning*, esta também é uma técnica não supervisionada de discretização. Particiona a distribuição dos dados de um atributo em subconjuntos separados. Cada subconjunto representa um par de freqüências. Os subconjuntos podem ser criados pela freqüência ou pela amplitude dos valores [Berka 2002];
- **Discretização baseado na entropia:** é uma técnica supervisionada. Explora a informação da distribuição da classe no cálculo de determinação de pontos de divisão (*spli-points*). Para discretizar um atributo, o método seleciona o valor do atributo que tem a menor entropia como um ponto de divisão, e recursivamente divide os intervalos resultantes até chegar numa discretização hierárquica [Fayyad & Irani 1993].

4.2.4.4.3 Codificação Binária

Para determinados algoritmos, não é possível a apresentação de atributos categóricos (exemplo, RNA). Para que seja possível a utilização desses atributos, estes podem ser representados através de uma codificação binária [Monteiro 1999].

Um dos grandes desafios com a aplicação dessa técnica é garantir que após a codificação, o algoritmo seja capaz de diferenciar e relacionar a magnitude e ordenação dos valores [Amorim 2004].

Exemplos de codificação binária são [Monteiro 1999]:

- 1 de N: Codificação que utilizada N bits, cada um representando um valor discreto do atributo original. Essa codificação é feita acendendo o bit (1) para representar um determinado valor, enquanto os outros bits estão apagados (0), ou seja, N-1 elementos terão o valor 0 e apenas um bit com o valor 1. As vantagens dessa técnica são a simplicidade, facilidade e o fato de algoritmos, como RNA, poderem aprender facilmente a distinguir os valores. O ponto negativo é que quando o atributo com uma grande quantidade de valores distintos, a representação desta codificação tende a ter um custo muito alto. Conseqüentemente, o tempo para obtenção do modelo é aumentado e o poder de generalização pode ser prejudicado;
- Padrão: Cada valor do atributo recebe um valor binário correspondente, crescente de 1 até a dimensão do atributo (número de valores distintos do atributo). A dimensão da representação é igual ao inteiro imediatamente superior a log2(N), onde N é o valor máximo do atributo. Esta representação é bastante simples, mas distancia valores próximos;
- M de N: proposta para resolver os problemas apresentados na codificação padrão (problema da distância) e da 1 de N (problema da dimensão). Nesta codificação temos sempre M bits acesos dos N bits possíveis, desta forma todos os elementos do conjunto terão a mesma representatividade. Uma das codificações mais aplicadas
- Termômetro: É utilizada quando existe uma relação entre os valores do atributo. Neste caso, pode ser necessário aproximar alguns valores e afastar outros, a fim de que o algoritmo de MD faça uso dessas distâncias para achar uma solução melhor, de forma mais rápida. Como exemplo, considerando que

um atributo pode assumir os valores ruim, bom e ótimo, e é desejável que o valor ótimo esteja distante de ruim e próximo de bom, a representação (100, 110, 111) seria adequada. Apesar de funcionar bem, esta técnica exige a aplicação de taxonomias e engenharia de conhecimento, e não é tão aplicada quanto às demais.

A escolha das técnicas de Tratamento de Dados está totalmente ligada ao algoritmo de mineração de dados escolhido. Um exemplo é Redes Neurais Artificiais, que por ter seu processamento baseado em cálculos matemáticos, todos os atributos devem ser numéricos [Haykin 1999]. Dessa forma, na presença de atributos categóricos, pode ser utilizada alguma técnica de codificação binária [Monteiro 1999] para a conversão desses atributos categóricos.

Após a aplicação de qualquer técnica citada no processo 4.2.4, deverá ser realizado uma homologação dos dados, a fim de garantir que as atividades deste processo foram executadas corretamente. Se qualquer erro for detectado, deve ser documentado, reportado e corrigido. Após a correção dos dados, deve haver uma nova homologação. Esse ciclo se repete até que a homologação não detecte erros.

As entradas, as saídas, os responsáveis e as atividades associadas a este processo estão especificadas a seguir:

Entradas:

- Documentações sobre o banco de dados;
- Tabela final desnormalizada;
- Visão final (conceitual) dos dados;

Atividades:

- Aplicação de técnicas de tratamento de dados;

Responsável:

- Analista de dados;
- Especialista em Banco de Dados;
- Especialista no domínio;

Saídas:

- Tabela final apropriada para a aplicação de algum algoritmo de mineração de dados.

4.2.5 Processos Extras

Esta fase tem por finalidade a documentação de qualquer atividade realizada no projeto que não tenha sido contemplada na metodologia *DMBuilding*.

Não existe um fluxograma para esta fase, pois a mesma é composta de um único processo, que é a realização da documentação dos processos extras.

4.2.5.1 Documentação de Processos Extras

O objetivo deste processo é a documentação de quaisquer atividades que necessitem ser realizadas, mas que, no entanto, não estão mapeadas em nenhuma das quatro fases da metodologia *DMBuilding*.

Esta metodologia tem por intuito ser a mais ampla possível, englobando todos os processos realizados em um projeto de Montagem de Visão de Dados. No entanto, existem especificidades inerentes ao problema e à forma como a organização executa os projetos de KDD que não são contemplados pela metodologia *DMBuilding*. Essas especificidades devem ser documentadas.

A documentação dos processos extras são importantes para a evolução da metodologia proposta, pois pode-se averiguar que determinada atividade é tão complexa que dever ser inserida na metodologia *DMBuilding*.

Ao detectar uma especificidade, o líder do projeto analisa a demanda e dispara a solicitação para o responsável, indicando a fase na qual a demanda se encaixa, as entradas disponíveis e o que deve ser feito. Após a execução da atividade, o responsável elabora a documentação, detalhando tudo o que foi feito e o que foi gerado durante a execução da atividade.

Entradas:

- Requisição da atividade, contendo a fase a qual a atividade pertence, o objetivo e os resultados esperados.

Responsáveis:

- Líder do projeto;
- Equipe.

Atividades:

- Realizar e documentar a atividade.

Saídas:

- Documentação da atividade;
- Conjunto de resultados finais: documentação do produto final gerado pela execução da atividade.

4.3 Considerações Finais

Ao comparar a metodologia *DMBuilding* com as metodologias apresentadas no Capítulo 3, alguns pontos positivos podem ser destacados.

A Tabela 4.1 apresenta o quadro comparativo da Seção 3.7, ampliando um campo para contemplar as características da metodologia *DMBuilding*. A notação "N" foi atribuída para as metodologias que não abordam a característica identificada. O valor "M" para as que mencionam, mas não entram em detalhes e "A" para as que abordam em detalhes.

Tabela 4.1 – Comparativo entre a metodologia *DMBuilding* e outras metodologias

Atividade	Fayyad	CRISP-DM	DMEasy	Yu	Quadstone	Han & Kamber	*DMBuilding*
Levantamento / Definição do problema	M	A	A	A	N	N	A
Verificação dos dados	N	M	M	N	N	N	A
Integração dos dados	N	N	N	A	N	A	A
Transformação de variáveis	N	N	M	N	M	PA	A
Agrupamento Transacional	N	N	M	N	A	PA	A
Homologação dos dados	N	N	M	A	N	N	A
Tratamento de variáveis	A	M	M	A	N	A	A
Documentação	N	N	A	N	N	N	A
Processos Extras	N	N	A	N	N	N	A
Abordagem a diversos problemas de MD	N	N	N	N	N	A	PA

N – Não Aborda; **M** – Menciona; **A** – Aborda; **PA –** Parcialmente Aborda

Analisando a Tabela 4.1, pode-se concluir que a metodologia *DMBuilding* apresenta as seguintes vantagens:

1. É a única metodologia que aborda detalhadamente todas as atividades analisadas, focando na integração, na transformação e no tratamento dos dados;
2. Relaciona problemas pertinentes aos dados e técnicas para a resolução dos mesmos, sugestões de transformações e as técnicas de tratamento de dados mais comumente usadas, a fim de enriquecer a base para a aplicação de algum método de mineração de dados;
3. Preocupa-se com a fase de homologação de dados, fase muito importante para garantir a qualidade dos dados;
4. Destaca a importância da documentação de todas as fases, para obter melhor entendimento e controle do projeto;
5. Engloba a documentação de processos extras para evoluir a metodologia.

A metodologia proposta possui a limitação de não ser tão ampla a ponto de englobar o tratamento de dados para todos os problemas de MD. O foco dessa metodologia está nos problemas que são baseados em dados transacionais e cadastrais. Mesmo possuindo esse foco, a metodologia pode ser utilizada em outros problemas de MD, mas com algumas restrições mais específicas dos problemas de MD e que não são claramente abordados nessa metodologia.

Ao ser comparada às metodologias estudadas, a metodologia *DMBuilding* é mais completa, o que pode resultar em um tempo maior dispendido em sua utilização, principalmente devido às documentações envolvidas no projeto. Mesmo demandando um tempo a mais, essas documentações são extremamente necessárias, principalmente para ser realizado o acompanhamento do projeto por qualquer uma das partes envolvidas.

Capítulo 5

Estudo de Caso

5.1 Introdução

Este capítulo apresenta a aplicação da metodologia *DMBuilding* em um problema de análise de risco de crédito de uma instituição financeira especializada em financiamento de produtos. Este é um problema real, de larga escala, que foi escolhido para verificar a aplicação prática de todas as fases desta metodologia.

A tomada de decisões de concessão de crédito baseia-se fundamentalmente na avaliação do risco de inadimplência dos potenciais contratantes de produtos de crédito. Para facilitar essa decisão, técnicas de Inteligência Artificial vêm sendo utilizadas [Lacerda et al. 2003].

Inicialmente, este capítulo apresenta alguns conceitos relacionados à área de análise de risco de crédito. Posteriormente, é apresentada em detalhes a aplicação de toda a metodologia proposta para a base de dados selecionada. O objetivo desta aplicação é demonstrar a viabilidade prática da metodologia *DMBuilding* em problemas do mundo real.

5.2 Análise de Risco de Crédito

O conceito de crédito pode ser analisado sob diversas perspectivas. Para uma instituição financeira, crédito se refere principalmente à atividade de colocar um valor à disposição de um tomador de recursos sob a forma de um empréstimo ou financiamento, mediante compromisso de pagamento em uma data futura [Brito & Neto 2005].

O risco de crédito é a forma de risco mais antiga no mercado financeiro [Figueiredo 2001]. É conseqüência de uma transação financeira contratada entre um fornecedor de fundos (doador do crédito) e um usuário (tomador do crédito). O puro ato de emprestar uma quantia a alguém traz embutida em si a probabilidade da mesma não ser recebida. Portanto, segundo [Caouette et al 2000], se crédito pode ser definido como a expectativa de recebimento de uma soma em dinheiro em um prazo determinado, então risco de crédito é a chance que este recebimento não se concretize.

A maioria dos ambientes de avaliação de crédito utiliza uma enorme quantidade de informações provenientes de diversas fontes durante o processo de tomada de decisões. Informações econômicas e pessoais do solicitante do crédito normalmente estão disponíveis [Amorim 2004].

No processo de concessão de crédito existem basicamente três etapas a serem realizadas [Monteiro 1999]:

- **Análise retrospectiva:** Avalia-se o desempenho histórico do solicitante para identificar fatores na condição atual do solicitante que possam levar ao não cumprimento de seus deveres;
- **Análise de tendências:** Realiza-se uma projeção da condição financeira do solicitante associada à análise de sua capacidade de suportar certo nível de endividamento;
- **Capacidade creditícia:** Baseando-se nas análises anteriores, decide-se pela concessão ou não de crédito e, em caso positivo, o valor permitido. O objetivo é obter a máxima proteção da empresa contra eventuais perdas. Após a concessão do crédito, a empresa acompanha as transações financeiras do cliente e, de acordo com um critério definido, o rotula como adimplente ou inadimplente.

Nesse contexto, observa-se uma maior ênfase das instituições na utilização de modelos que servem de suporte às decisões de concessão de crédito e à gestão das carteiras. RNA têm sido aplicadas com sucesso no domínio de análise de crédito [Widrow et al. 1994] [Mendes Filho et al. 1997] [Sousa & Carvalho 1999] [Vasconcelos et al. 1999] [Amorim et al. 2007]. O sucesso da aplicação de RNA no setor financeiro é conseqüência da capacidade das mesmas em aprender funções complexas e não-lineares. Tal característica é uma vantagem sobre as técnicas estatísticas convencionais (Ex.: análise discriminante) [Lacerda et al. 2003].

Os modelos de classificação de risco buscam avaliar o risco de um tomador ou operação, atribuindo uma medida que representa a expectativa de risco de falha de

pagamento, geralmente expressa na forma de uma classificação de risco (*rating*) ou pontuação (*score*) [Brito & Neto 2005].

Apesar da eficácia comprovada das soluções automatizadas para análise de crédito, estas funcionam como ferramentas de apoio à decisão, ficando, a decisão final por conta do analista [Amorim 2004]. Na verdade, normalmente o analista se preocupa apenas com as propostas mais complexas, nas quais o sistema não chegou a uma decisão dentro de uma margem de confiabilidade [Monteiro 1999].

5.3 Aplicação da Metodologia *DMBuilding*

Esta seção descreve a aplicação de todas as fases da metodologia *DMBuilding* no estudo de caso de análise de risco de crédito.

5.3.1 Entendimento do Problema

Para a visualização das atividades realizadas nesta fase, os processos do fluxograma apresentado no Capítulo 4 (Figura 4.2) foram desmembrados.

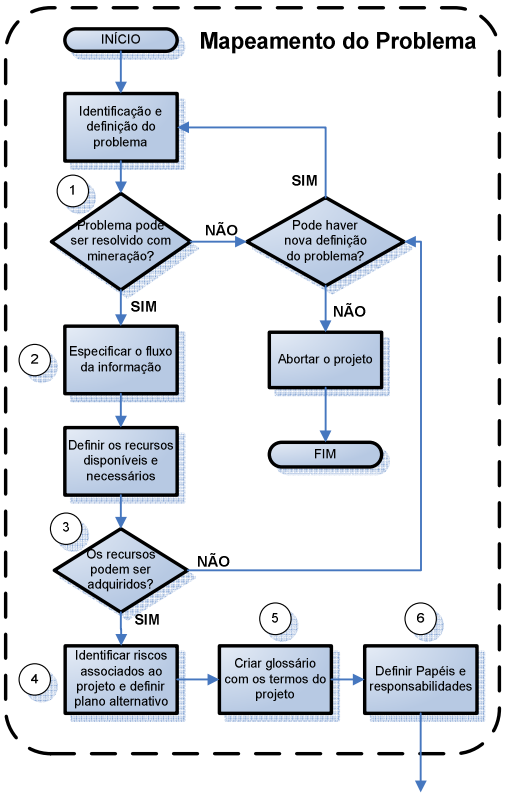


Figura 5.1 – Processo de Mapeamento do Problema

A Figura 5.1 apresenta o processo de Mapeamento do Problema. Os círculos numerados representam as atividades realizadas neste estudo de caso. Cada atividade é descrita a seguir conforme numeração da Figura 5.1.

(1) A Empresa "X" atua no ramo de financiamento de produtos para pessoas físicas e jurídicas. Apesar da experiência na área, a empresa possui dois aspectos a serem aprimorados na carteira de pessoa física:

- A agilidade do processo de tomada de decisão;
- O índice de inadimplência.

Este é um problema típico de Classificação de Padrões, onde objetiva-se a decisão de categorização de um solicitante de crédito entre duas classes existentes: a de BOM e a de MAU cliente. A solução a ser construída, considera a partir da análise de uma base histórica da Empresa "X", será construído (aprendido) um modelo de perfil da definição do BOM e do MAU, para posterior tomada de decisão, quando a solução é colocada em produção. Nesse caso, foi estabelecido como critério pela Empresa "X" que MAU cliente é aquele que encontra-se com uma parcela de contrato vencida e não paga a mais 90 dias. BOM cliente é todo aquele que não é MAU.

(2) Não foi necessário especificar o fluxo da informação, pois a empresa já possuía esta definição (apresentado na Figura 5.2). O fluxo da informação inclui informações de consultas internas e externas, e da análise dos especialistas do domínio.

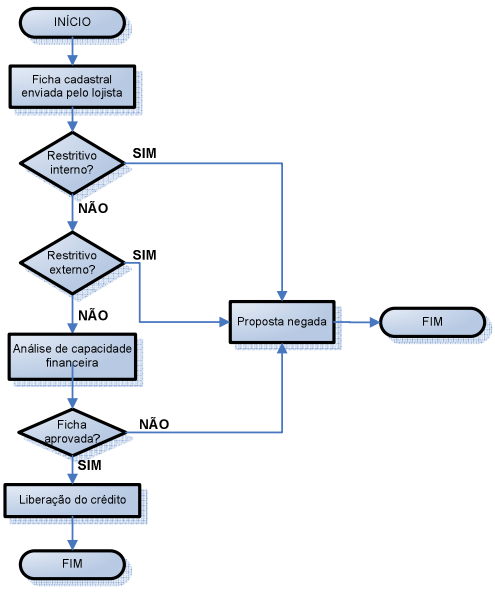


Figura 5.2 – Fluxo de tomada de decisão da Empresa "X"

(3) Alguns recursos foram definidos para garantir a execução do projeto. Cinco pessoas foram alocadas com dedicação exclusiva ao projeto. Além disso, pelo menos dois computadores deveriam estar disponíveis: um para execução do projeto e outro para auxiliar no processo de tomada de decisão. Todos os recursos necessários foram disponibilizados.

(4) Após a definição dos recursos disponíveis e necessários, foram identificados os riscos associados ao projeto e definidas as ações que deveriam ser realizadas (plano alternativo) caso os riscos viessem a ocorrer, como por exemplo:

- **Risco**: Resultados não satisfatórios ao final do projeto;
- **Plano alternativo**: Aplicação de outros métodos de mineração de dados ou a geração de uma nova base de dados.

(5) Para o entendimento de todos os termos utilizados tanto pela equipe da Empresa "X" (relacionados ao negócio) como pela equipe participante do projeto de KDD (relacionados à T.I. e MD), foi construído um glossário que engloba todos esses termos. Um exemplo para esse projeto é a definição do termo "RENDA". Para a empresa em questão, "RENDA" não significa renda (salário) do cliente. Este termo se refere aos "juros sobre a operação".

(6) Após a criação do glossário, definiram-se os papéis e as responsabilidades das pessoas envolvidas no projeto. Para esse projeto, cinco pessoas foram alocadas. Um membro da equipe acumulou papéis, assumindo a responsabilidade de Especialista em Banco de Dados e de Analista de Dados.

Com a definição dos responsáveis, encerra-se o processo de Mapeamento do Problema e inicia-se o processo de Planejamento Técnico.

A Figura 5.3 apresenta o processo de Planejamento Técnico. Os círculos numerados representam as atividades realizadas neste estudo de caso. Cada atividade é descrita a seguir conforme numeração da Figura 5.3.

(1) A primeira atividade deste processo é a especificação da técnica de mineração de dados a ser utilizada. A técnica selecionada para este projeto foi Redes Neurais Artificiais (RNA).

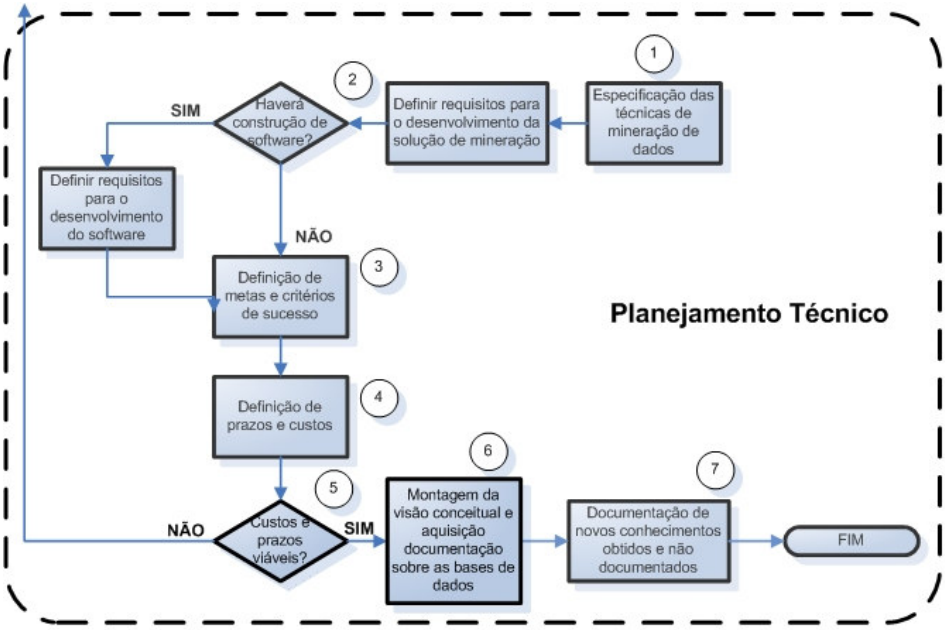


Figura 5.3 – Processo de Planejamento Técnico

(2) Para o estudo de caso, nenhum requisito extra teve de ser definido para a construção da solução, pois já existia um software para a utilização dos resultados da solução de MD na Empresa "X". Este software se chama *NeuralScorer®*[3] da empresa NeuroTech [Neurotech 2002] e será utilizado no momento da análise do crédito.

(3) Para esse projeto, foram definidos os seguintes critérios de sucesso e metas:

- Aprovação superior a 70%;
- Inadimplência inferior a 7% (a inadimplência atual é de 9,5%).

(4) Analisando a complexidade do problema, foi proposto um prazo máximo de 12 semanas para o desenvolvimento do projeto. Ambas as partes concordaram com o prazo e os custos associados ao projeto. A informação de custos do projeto é sigilosa, portanto não pode ser incluída nessa dissertação.

(5) Após apresentado os custos e prazos relacionados ao projeto, a diretoria da Empresa "X" aprovou a viabilidade dos mesmos, sendo possível a continuação do projeto.

(6) Esta atividade foi responsável pela criação da visão de dados conceitual e a aquisição da documentação sobre as bases de dados. O analista de negócio utilizou a documentação de dados disponível para propor a visão de dados conceitual,

[3] ® *NeuralScorer* é propriedade da empresa NeuroTech. Todos os Direitos Reservados.

tendo sido homologada pelo especialista no domínio. A documentação disponível era somente o Dicionário de Dados. Como a Empresa "X" não possuía muita documentação sobre suas bases de dados, ficou a critério do projeto, o especialista de Banco de Dados elaborar posteriormente D.E.R.

(7) Em reuniões realizadas entre as equipes do projeto, foi exposto que toda vez que uma proposta tiver um "AVALISTA", as informações de cadastro do avalista também devem ser adquiridas e analisadas juntamente com as informações cadastrais do próprio cliente. Esse conhecimento foi registrado como conhecimento obtido e não documentado e incrementados na visão de dados conceitual.

Após a realização de todas essas atividades, encerrou-se a fase de Entendimento do Problema. Todas as características do problema foram documentadas para que os envolvidos no projeto possam ter acesso. A próxima fase a ser realizada é a fase de Verificação dos Dados. Após a definição do problema, é necessário adquirir os dados e a verificar sua integridade e qualidade a fim de analisar a viabilidade do projeto de acordo com os dados disponíveis.

5.3.2 Verificação dos Dados

Para a visualização das atividades realizadas nesta fase, os processos do fluxograma apresentado no Capítulo 4 (Figura 4.3) foram desmembrados.

A Figura 5.4 apresenta o processo de Identificação e Seleção das Fontes de Dados. Os círculos numerados representam as atividades realizadas neste estudo de caso. Cada atividade é descrita a seguir conforme numeração da Figura 5.4.

(1) A primeira atividade é a identificação das fontes de dados. Todos os dados do projeto estavam disponíveis em um SGBD da Oracle [Oracle 2007]. Para a realização desse projeto, nenhum problema de coleta foi encontrado.

(2) No banco operacional da empresa existiam mais de 50 tabelas, porém apenas quatro estão relacionadas ao foco do projeto: CLIENTES (contém informações cadastrais dos clientes), PROPOSTAS (contém informações sobre as propostas de financiamento de produtos), CONTRATOS (armazena informações das propostas que viraram contratos – propostas aprovadas) e PARCELAS (armazena todas as parcelas dos contratos). As informações das tabelas podem ser observadas na Tabela 5.1.

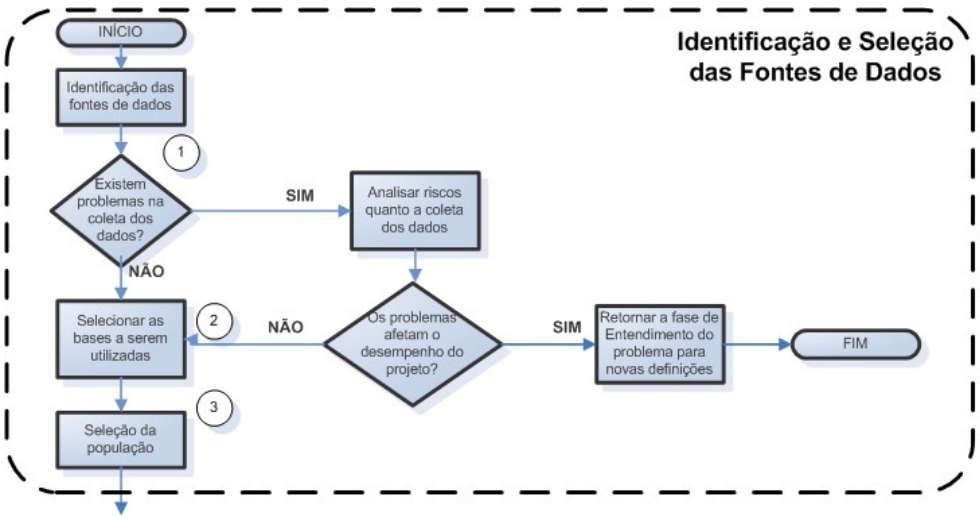


Figura 5.4 – Processo de Identificação e Seleção das Fontes de Dados

(3) Como o foco do projeto é a análise de risco de crédito para pessoa física, foram selecionados apenas os registros que possuíam o atributo "TIPO_PESSOA" com valor igual a "F" (pessoa física). Além desse filtro, outros foram realizados:

- Seleção das propostas aprovadas, pois possibilitam a definição do alvo;
- Seleção das propostas que foram realizadas no período de 02/01/2004 a 02/01/2007.

Ao final de seleção, restaram 38.607 registros. As tabelas foram salvas em arquivos texto (.txt), onde cada linha do arquivo representa um registro e cada coluna representa um atributo.

Tabela 5.1 – Informações das tabelas disponíveis

Tabela	Descrição	Total de registros	Total de atributos	Chave
CLIENTES	Informações dos clientes	214.200	339	CLIENTE
CONTRATOS	Informações dos contratos	54.339	147	CONTRATO
PARCELAS	Informações de parcelas dos contratos	1.659.245	69	CONTRATO, PARCELA
PROPOSTAS	Informações das propostas de crédito	102.998	183	PROPOSTA

Após realizar a seleção da população, iniciou-se o processo de Análise de Integridade e Integração, identificado pela Figura 5.5. Os círculos numerados representam as atividades realizadas neste estudo de caso. Cada atividade é descrita a seguir conforme numeração da Figura 5.5. Durante este processo, o especialista em Banco de Dados utilizou duas ferramentas de análise: o SPSS [SPSS 2007] e o SQL Server 2000 [SQL Server 2000].

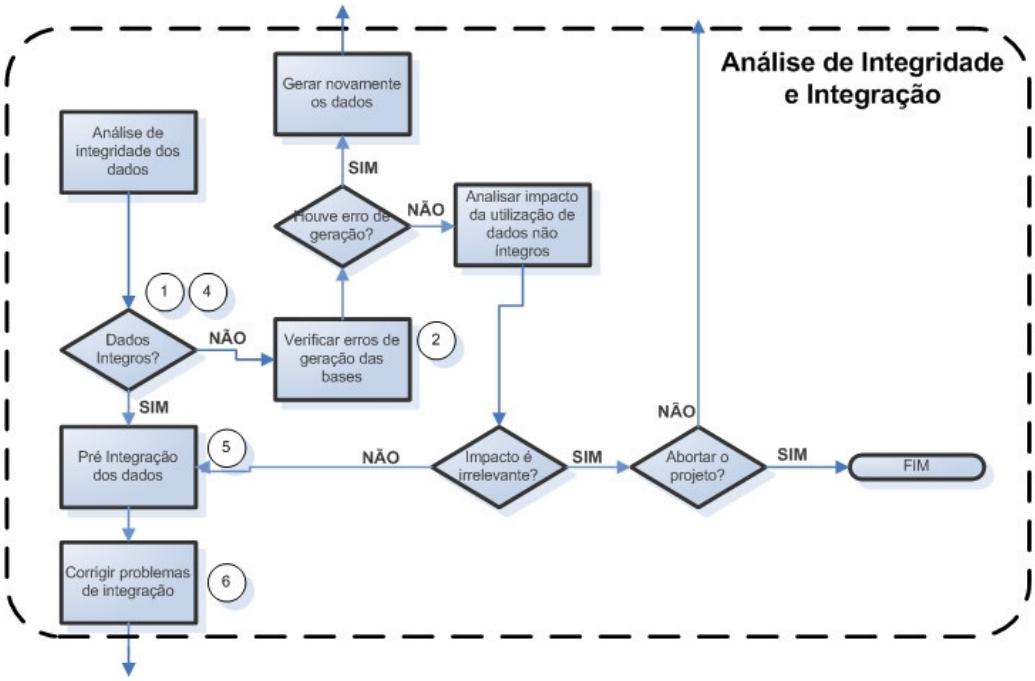


Figura 5.5 – Processo de Análise de Integridade e Integração

(1) Para a realização da análise de integridade, todas as tabelas foram importadas para o SGBD Sql Server 2000. Ao tentar importar a tabela CLIENTES, foi verificado um erro. O arquivo não estava no formato de um cliente por linha. Foi requisitado ao especialista de T.I. da Empresa "X" que investigasse o motivo deste erro.

(2) Após a verificação, foi descoberto que o atributo "OBSERVAÇÃO" continha uma quebra de linha que impossibilitava a importação. Este problema foi causado devido a um erro na geração da base.

(3) O especialista de Banco de Dados da Empresa "X" foi responsável pela nova geração desta base, retirando o atributo com problema. Ao gerar novamente a base de dados, foi necessário o retorno à atividade de seleção da população, referente ao processo anterior (Identificação e Seleção das Fontes de Dados), pois os dados devem ser selecionados de acordo com os critérios de seleção da população. Posteriormente a essa nova geração, a importação ao SQL Server foi realizada com sucesso.

(4) Como houve retorno à fase anterior, a atividade de análise de integridade teve de ser refeita. Para analisar a integridade dos dados, alguns conceitos das tabelas tiveram que ser verificados. Os detalhes para a realização dessas verificações estão contidos na Tabela 5.2, que apresenta informações de relacionamento entre as tabelas.

Tabela 5.2 – Informações de relacionamento entre as tabelas

Tabela_1	Campo da Tabela_1	Tabela_2	Campo da Tabela_2	Relacionamento
CLIENTES	CLIENTE	PROPOSTAS	CLIENTE	1 – N
PROPOSTAS	CONTRATO	CONTRATOS	CONTRATO	0..1 – 1
CLIENTES	CLIENTE	CONTRATOS	CLIENTES	1 – N
CONTRATOS	CONTRATO	PARCELAS	CONTRATO	1 .. N

Como a Empresa "X" não possuía um Diagrama de Entidade e Relacionamento (DER) de suas bases de dados, um DER para as tabelas do projeto foi desenvolvido e pode ser observado na Figura 5.6.

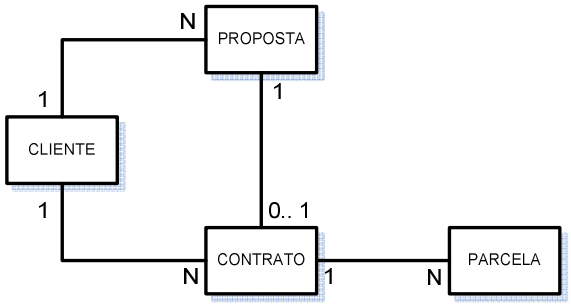


Figura 5.6 – Diagrama de Entidade e Relacionamento (DER) da bases do estudo de caso

As seguintes verificações foram realizadas:

- De posse da informação de quais variáveis compõem a chave primária de cada tabela, foi verificado se essas variáveis apresentavam valores duplicados, ausentes ou com preenchimento incorreto. Nenhum problema foi detectado;
- Com a integridade da chave primária garantida, foram verificadas as conexões entre as tabelas a fim de garantir que os dados pudessem ser desnormalizados de forma correta. Alguns problemas foram encontrados: nem todos os contratos possuíam parcelas relacionadas e algumas parcelas não possuíam informação de contrato associado. Ao analisar a origem do problema, verificou-se que essas duas bases foram selecionadas em períodos diferentes, o que causou o problema de conexão. O problema foi reportado e as duas bases geradas novamente sem apresentar erros de conexão;
- Como se trata de um problema de aprendizado supervisionado (classificação), foi verificada se a variável alvo estava presente na base. Essa variável não foi encontrada, mas pôde ser obtida através da tabela PARCELAS. O cliente foi classificado como MAU se tivesse parcela de contrato vencida e não paga a mais de 90 dias. Caso contrário, foi classificado como BOM;

Todos os *scripts* da linguagem SQL utilizados nas verificações de integridade foram salvos para posteriores consultas.

(5) Após garantir a integridade dos dados, iniciou-se a atividade de pré-integração, que resultou na junção das tabelas PROPOSTA e CLIENTES. Além disso, novas variáveis foram criadas para inserir informações dos avalistas. As informações dos clientes consideradas mais importantes foram replicadas para os avalistas. Os dados dos avalistas só foram preenchidos para os casos em que o campo "AVALISTA" estava preenchido com o CPF do avalista da proposta de crédito. A seleção das variáveis mais relevantes foi feita com o auxílio do especialista do domínio e através da análise do preenchimento das variáveis.

(6) Nenhum problema de pré-integração foi encontrado. Após a análise de integridade e a realização da pré-integração, o especialista em Banco de Dados gerou um relatório especificando as inconsistências encontradas, os artefatos utilizados, as ações realizadas e um parecer de integridade dos dados.

Como o parecer foi positivo, passou-se para o próximo processo: Análise e Exploração dos Dados.

A Figura 5.7 apresenta o processo de Análise e Exploração dos Dados. Os círculos numerados representam as atividades realizadas neste estudo de caso. Cada atividade é descrita a seguir conforme numeração da Figura 5.7.

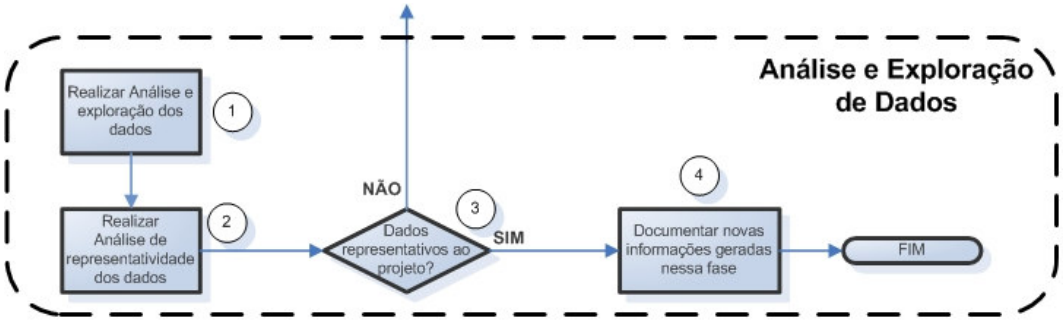


Figura 5.7 – Processo de Análise e Exploração de Dados

(1) Para realizar a análise visual dos dados, utilizou-se o software *Neural Scorer Development*®[4], propriedade intelectual da empresa NeuroTech [Neurotech 2007], gentilmente cedido para a realização deste trabalho.

Foi verificado o preenchimento das variáveis de todas das tabelas, a fim de garantir que o preenchimento era satisfatório e representativo para o projeto.

[4] ® *Neural Scorer Development* é propriedade da empresa NeuroTech. Todos os Direitos Reservados.

Percebeu-se que algumas tabelas continham variáveis com baixo ou nenhum preenchimento. Essas variáveis foram verificadas pelo especialista do domínio de Empresa "X", que confirmou que o baixo preenchimento era correto e normal. Nenhum erro foi detectado nessa análise.

Analisando o negócio da empresa "X", percebeu-se a existência de dados transacionais, pois um cliente que está requisitando um financiamento pode ter um histórico de propostas de financiamento e de compra de produtos na mesma instituição.

(2) Para analisar se os dados são representativos ao problema, o analista de negócio verificou na base de dados a existência de variáveis consideradas essenciais para a realização do projeto. Com o suporte do especialista no domínio, verificou-se a existência de variáveis importantes, como data de nascimento, sexo, endereço, salário, entre outras.

Conforme pode ser verificado anteriormente na Tabela 5.1, a quantidade de dados disponíveis foi considerada suficiente para a realização do projeto, não sendo necessária a realização de amostragem.

Um dos problemas encontrados foi a quantidade de atributos em cada tabela. Como pode ser visto na Tabela 5.1, algumas tabelas possuem mais de 300 atributos (Ex.: CLIENTES). Com o auxílio do especialista do domínio, decidiu-se que seriam removidos os atributos que satisfizessem ao menos um dos seguintes critérios:

a. Atributos 100% não preenchidos;
b. Atributos com conteúdo *a posteriori* (variáveis obtidas após o momento da proposta de crédito);
c. Atributos não relevantes para o problema (segundo análise do especialista do domínio).

Seguindo esses critérios de eliminação de atributos, da tabela CLIENTES, de 339 atributos, 254 foram retirados (restando 74) e, da tabela PROPOSTAS, de 183 atributos, 146 foram retirados (restando 37). Os atributos das tabelas CONTRATOS e PARCELAS não foram eliminados, pois são tratados na fase de Montagem de Visão. As Tabelas 5.3 e 5.4 mostram os atributos selecionados das tabelas CLIENTES e PROPOSTAS, respectivamente.

Para enriquecer a base de dados, foram adquiridas informações externas dos clientes em uma instituição do setor logístico. Este setor logístico possuía um

sistema de escoragem dos clientes consultados. Para o projeto, somente a informação do *"pontuacao_externa"* foi coletada como entrada para o modelo.

Tabela 5.3 – Atributos selecionados da tabela CLIENTES

COD_CLI	TEMPO_EMPREGO_ANOS_CLI	UF_RESIDENCIAL_AVALISTA
TIPO_PESSOA	TEMPO_EMPREGO_MESES_CLI	TEMPO_ANOS_RESIDENCIAL_AVALISTA
DATA_NASCIMENTO_CLI	OUTRAS_RENDAS_CLI	TEMPO_MESES_RESIDENCIAL_AVALISTA
SEXO_CLI	OUTRAS_RENDAS_DESCRICAO_CLI	TIPO_RESIDENCIA_AVALISTA
CPF_CNPJ_CLI	VEICULO_DESCRICAO_CLI	DDD_RESIDENCIAL_AVALISTA
NATURAL_CLI	VEICULO_VALOR_CLI	FONE_RESIDENCIAL_AVALISTA
NACIONALIDADE_CLI	VEICULO_ANO	DDD2_RESIDENCIAL_AVALISTA
UF_NASCIMENTO_CLI	BANCO_CLI	FONE2_RESIDENCIAL_AVALISTA
ESTADO_CIVIL_CLI	AGENCIA_CLI	BAIRRO_COMERCIAL_AVALISTA
CPF_CONJUGE_CLI	CIDADE_REFERENCIA_CLI	CEP_COMERCIAL_AVALISTA
DOCUMENTO_IDENTIFICADOR_CLI	UF_REFERENCIA_CLI	CIDADE_COMERCIAL_AVALISTA
UF_EMISSAO_DOCUMENTO_CLI	FONE_REFERENCIA_CLI	UF_COMERCIAL_AVALISTA
ORGAO_EMISSOR_CLI	FONE2_REFERENCIA_CLI	DDD_COMERCIAL_AVALISTA
DATA_EMISSAO_DOCUMENTO_CLI	FUNCIONARIO	FONE_COMERCIAL_AVALISTA
BAIRRO_RESIDENCIAL_CLI	EMPRESANOME	RAMAL_COMERCIAL_AVALISTA
CEP_RESIDENCIAL_CLI	REFTEMPOCONTA1	ATIVIDADE_EMPRESA_AVALISTA
CIDADE_RESIDENCIAL_CLI	ATIVIDADE_TIPOSETOR	ATIVIDADE_SETOR_AVALISTA
UF_RESIDENCIAL_CLI	CPF_CONJUGE_CLI	ATIVIDADE_SOCIO_AVALISTA
TEMPO_ANOS_RESIDENCIAL_CLI	EMPRESA_CONJUGE_CLI	SALARIO_AVALISTA
TEMPO_MESES_RESIDENCIAL_CLI	FONE_CONJUGE_CLI	TEMPO_EMPREGO_ANOS_AVALISTA
TIPO_RESIDENCIA_CLI	ATIVIDADE_CONJUGE_CLI	OUTRAS_RENDAS_AVALISTA
DDD_RESIDENCIAL_CLI	CARGO_CONJUGE_CLI	OUTRAS_RENDAS_DESCRICAO_AVALISTA
FONE_RESIDENCIAL_CLI	SALARIO_CONJUGE_CLI	VEICULO_DESCRICAO_AVALISTA
DDD2_RESIDENCIAL_CLI	TEMPO_EMPREGO_CONJUGE_CLI	VEICULO_VALOR_AVALISTA
FONE2_RESIDENCIAL_CLI	EMPRESA_CNPJ_CONJUGE_CLI	VEICULO_ANO
ENDERECO_COMERCIAL_CLI	CONJUGE_DATA_NASCIMENTO_CLI	BANCO_AVALISTA
BAIRRO_COMERCIAL_CLI	NASCIMENTO_AVALISTA	AGENCIA_AVALISTA
CEP_COMERCIAL_CLI	SEXO_AVALISTA	CIDADE_REFERENCIA_AVALISTA
CIDADE_COMERCIAL_CLI	NATURALIDADE_AVALISTA	UF_REFERENCIA_AVALISTA
UF_COMERCIAL_CLI	CPF_CNPJ_AVALISTA	FONE_REFERENCIA_AVALISTA
DDD_COMERCIAL_CLI	NACIONALIDADE_AVALISTA	FONE2_REFERENCIA_AVALISTA
FONE_COMERCIAL_CLI	UF_NASCIMENTO_AVALISTA	FUNCIONARIO_AVALISTA
RAMAL_COMERCIAL_CLI	CPF_CONJUGE_AVALISTA	CPF_CONJUGE_AVALISTA
ATIVIDADE_EMPRESA_CLI	BAIRRO_RESIDENCIAL_AVALISTA	SALARIO_CONJUGE_AVALISTA
ATIVIDADE_SETOR_CLI	CELULAR_AVALISTA	TEMPO_EMPREGO_CONJUGE_AVALISTA
ATIVIDADE_SOCIO_CLI	CEP_RESIDENCIAL_AVALISTA	CONJUGE_DATA_NASCIMENTO_AVALISTA
SALARIO_CLI	CIDADE_RESIDENCIAL_AVALISTA	SETOR_ATIVIDADE_AVALISTA

Tabela 5.4 – Atributos selecionados da tabela PROPOSTAS

COD_CLI	FLAG_COMPROVANTE_RESIDENCIA	VALOR_FINANCIAMENTO	BORDERO
COD_PROPOSTA	FLAG_IDENTIDADE	VALOR_COMPRA	AVALISTA2
DATA_PROPOSTA	FLAG_CPF	VALOR_ENTRADA	DATA_APROVACAO
NUM_CONTRATO	FLAG_EXPERIENCIA_CREDITO	TIPO_ENTRADA	DATA_CANCELAMENTO
QTD_PARCELAS	COD_LOJA	VALOR_PRESTACAO	USUARIO
TAXA_MENSAL	COD_AGENTE	DESCRICAO_PRODUTO	TIPO_LIBERACAO
TAXA_RETORNO	COD_AGENCIA	TIPO_FINANCIAMENTO	SEGMENTO
ALIQUOTA_IOF	PRODUTO	SEGMENTO	SITUACAO
PONTUACAO_EXTERNA	TIPO_PROPOSTA	CLASSE	CLASSIFICACAO
AVALISTA	ORIGEM	TERMINAL	FLAG_GARANTIAS
FLAG_DOCUMENTOS	COD_PLANO	HORA	COD_TABELA_FINAN
FLAG_COMPRA_IMEDIATA	TAXA_FINANCIAMENTO	TURNO	ASSESSOR

(3) Após todas essas verificações, como a base de dados estava consistente e apresentava preenchimento satisfatório, o especialista em Banco de Dados emitiu um parecer favorável da viabilidade do projeto com relação à qualidade dos dados.

(4) Nenhuma documentação de novas informações foi gerada.

Após o parecer da viabilidade do projeto com relação à qualidade dos dados, passou-se para a próxima fase: Montagem de Visão.

5.3.3 Montagem da Visão

Para a visualização das atividades realizadas nesta fase, os processos do fluxograma apresentado no Capítulo 4 (Figura 4.5) foram desmembrados.

A Figura 5.8 apresenta o processo de Tratamento de Variáveis Brutas. Os círculos numerados representam as atividades realizadas neste estudo de caso. Cada atividade é descrita a seguir conforme numeração da Figura 5.8.

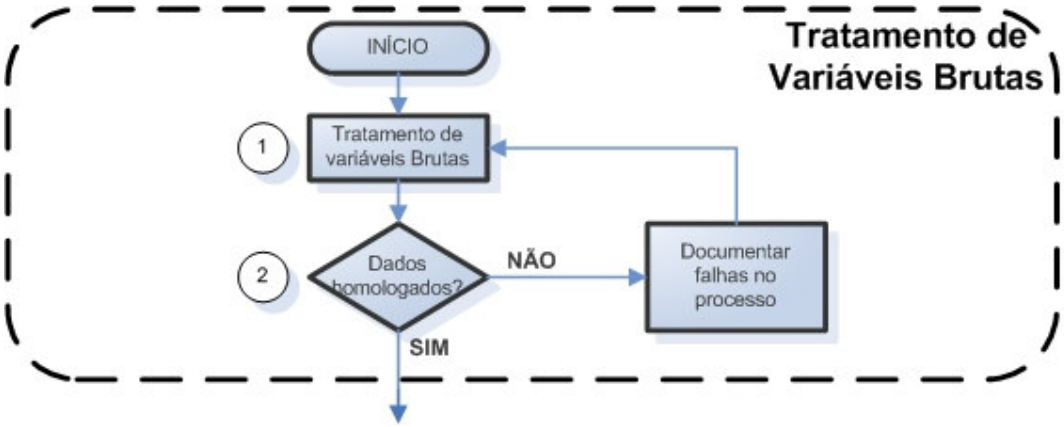


Figura 5.8 – Processo de Tratamento de Variáveis Brutas

(1) A primeira atividade desta fase é o tratamento das variáveis brutas, cujo objetivo é verificar se os atributos selecionados possuem preenchimento apropriado para a fase de aprendizagem.

Os seguintes tratamentos foram utilizados:

- **Colunas com um único valor:** Por se tratar de uma base de dados com muitas variáveis, essa verificação já foi realizada na fase anterior durante a seleção de variáveis;
- **Colunas com um valor predominante:** Atributos que praticamente apresentam um único valor devem ser analisados. Seguem alguns exemplos de

atributos que foram removidos por causa desse problema e o percentual de preenchimento do valore predominante: ALIQUOTA_IOF (99,99%), AVALISTA2 (99,74%), FLAG_EXPERIENCIA_CREDITO (99,21%) e NACIONALIDADE_CLI (99,9%);

- **Colunas com muitos valores distintos:** Variáveis categóricas que possuem quase todos os valores distintos e apresentam alguma estrutura interna em seu conteúdo. As variáveis de CEP fazem parte deste cenário e foram transformadas da seguinte forma: CEP_COMERCIAL_CLI foi subdividida em CEP_COMERCIAL_CLI1 (primeiro dígito do CEP_COMERCIAL_CLI), CEP_COMERCIAL_CLI2 (dois primeiros dígitos do CEP_COMERCIAL_CLI), CEP_COMERCIAL_CLI3 (três primeiros dígitos do CEP_COMERCIAL_CLI) e CEP_COMERCIAL_CLI4 (quatro primeiros dígitos do CEP_COMERCIAL_CLI). A mesma transformação foi realizada a variável RESCEP;
- **Criação do alvo:** Como a variável alvo deste projeto não estava presente nas bases de dados disponíveis, foi gerada a variável "BOM_MAU", que contém a informação sobre a conclusão da proposta de crédito. Se o cliente possuía, no momento da geração da base, alguma parcela não paga a mais de 90 dias, esse cliente era considerado MAU, caso contrário, BOM.

O analista de dados ficou responsável pela elaboração de um documento que especificava as variáveis eliminadas (e o motivo), como também a descrição das variáveis criadas.

(2) Após essa atividade, realizou-se a homologação dos dados para garantir que todas as variáveis foram tratadas de forma correta. Essa atividade foi concluída com sucesso e nenhum erro foi detectado.

Iniciou-se, então, o processo de transformação de variáveis, cujo objetivo é analisar as variáveis que teoricamente deveriam ser removidas devido ao seu conteúdo, mas que podem ser aproveitadas através da aplicação de algumas transformações. A Figura 5.9 apresenta este processo. Os círculos numerados representam as atividades realizadas neste estudo de caso. Cada atividade é descrita a seguir conforme numeração da Figura 5.9.

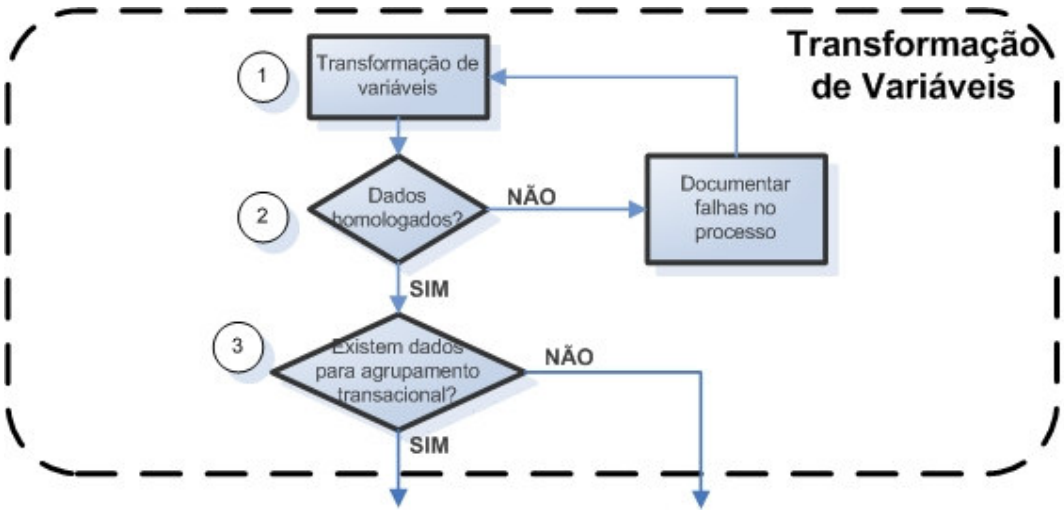


Figura 5.9 – Processo de Transformação de Variáveis

(1) Os seguintes tratamentos foram utilizados:

- **Tratamento de variáveis do tipo Data/Hora:** Esse tratamento foi aplicado à variável DATA_NASCIMENTO_CLI. A idade dos clientes foi calculada pela diferença (em anos) entre as variáveis DATA_NASCIMENTO_CLI e DATA_PROPOSTA, que indica a data da consulta;
- **Categoria "Outros":** Algumas variáveis categóricas possuíam uma distribuição não uniforme, concentrada em um conjunto de categorias. Os valores foram agrupados de acordo com o seguinte critério: categorias com menos de 30 exemplos foram consideradas como sendo da categoria "outros". Seguem alguns exemplos de variáveis que foram afetadas por este tratamento e o percentual de exemplos da categoria "outros": NATURAL_CLI (16,96%), CEP_RESIDENCIAL_CLI3 (1,24%) e CEP_RESIDENCIAL_CLI4 (9,06%). Esse tipo de tratamento foi aplicado em 37 variáveis;
- **Variáveis indicativas de ausência/presença (*flags*):** Das variáveis categóricas restantes, muitas possuíam a distribuição de valores muito concentrada em um único valor ou o seu formato não era apropriado (por exemplo, número de telefone). Nesse caso, foram construídas variáveis de indicativo de ausência/presença de determinado valor. Alguns exemplos em que essa técnica foi aplicada: NOME_CONJUGE_CLI (nome do cônjuge não traz benefício algum, porém a informação se foi ou não cadastrado com cônjuge pode ser relevante) e FONE_RESIDENCIAL_CLI (mesma lógica da variável

NOME_CONJUGE_CLI). Esse tipo de tratamento foi realizado em 22 variáveis.

(2) Uma nova atividade de homologação dos dados foi realizada para garantir que as variáveis foram tratadas de forma correta. Essa atividade foi concluída com sucesso e nenhum erro foi detectado.

(3) Foi verificado que, para esse projeto, existiam dados para a realização da fase de agrupamento transacional, pois as tabelas PARCELAS e PROPOSTAS fornecem informações de todo histórico daqueles clientes que já tiveram algum tipo de negociação com Empresa "X" (seja uma simples proposta de crédito que foi negada ou a efetivação de algum contrato).

Iniciou-se, então, o processo de Agrupamento Transacional. A Figura 5.10 apresenta este processo. Os círculos numerados representam as atividades realizadas neste estudo de caso. Cada atividade é descrita a seguir conforme numeração da Figura 5.10.

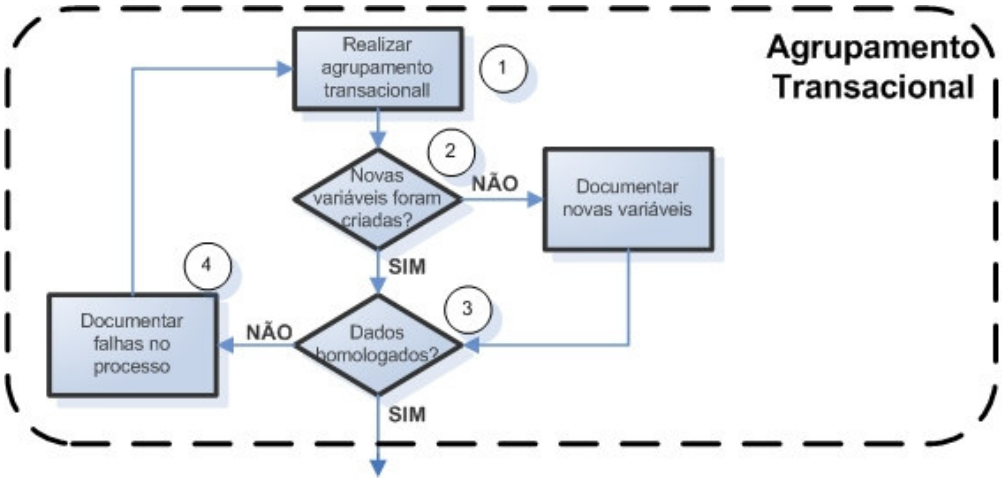


Figura 5.10 – Processo de Agrupamento Transacional

(1) O analista de dados propôs a criação de 22 atributos contendo agrupamento transacional, entre as quais podem ser citados:

- **Medidas de agregação -** Exemplos: QTD_CONTRATOS_ANTERIORES (quantidade de contratos realizados antes da proposta atual), QTD_PARCELAS_PAGAS (quantidade de parcelas pagas referentes a contratos anteriores), VALOR_PARCELAS_PAGAS (valor das parcelas pagas referentes a contratos anteriores) e QTD_NEGADAS (quantidade de propostas de crédito que foram negadas);

- **Funções de tendência -** Exemplos: ATRASO_MEDIO (atraso médio do cliente nos contratos anteriores) e ATRASO_MAXIMO (atraso máximo do cliente nos contratos anteriores);

- **Indicativo de freqüência –** Exemplos: TEMPO_DESDE_PRIMEIRA_NEGADA (tempo desde a primeira proposta negada) e TEMPO_DESDE_ULTIMA_NEGADA (tempo desde a última proposta negada);

- **Variação de tempo e tipo –** Exemplos: QTD_PARCELAS_PAGAS_ATRASO_ATE_90 (Quantidade de parcelas pagas com atraso de até 90 dias), VALOR_PARCELAS_PAGAS_ATRASO_ATE_90 (Valor de parcelas pagas com atraso de até 90 dias), QTD_PARCELAS_NAO_PAGAS_MAIOR_90 (Quantidade de parcelas vencidas e não pagas com atraso superior a 90 dias) e VALOR_PARCELAS_NAO_PAGAS_MAIOR_90 (Valor de parcelas vencidas e não pagas com atraso superior a 90 dias).

(2) Durante a realização dessa atividade, não surgiram propostas de criação de novas variáveis.

(3) Após a geração das novas variáveis, realizou-se a homologação dos dados. Foram encontrados problemas no cálculo de algumas variáveis.

(4) Foram documentados quais os erros encontrados, o motivo do erro e qual a correção para o mesmo. As correções foram realizadas e as variáveis foram novamente geradas (seguindo os passos das atividades 1 a 3). Uma nova fase de homologação foi realizada, porém sem apresentar erros.

Por fim, o processo de Integração dos Dados é realizado. A Figura 5.11 apresenta este processo. Os círculos numerados representam as atividades realizadas neste estudo de caso. Cada atividade é descrita a seguir conforme numeração da Figura 5.11.

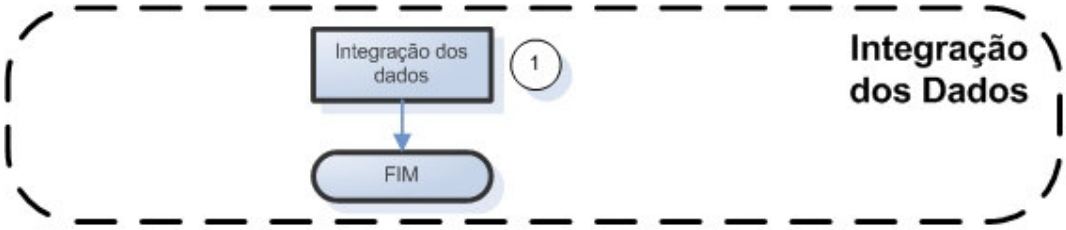


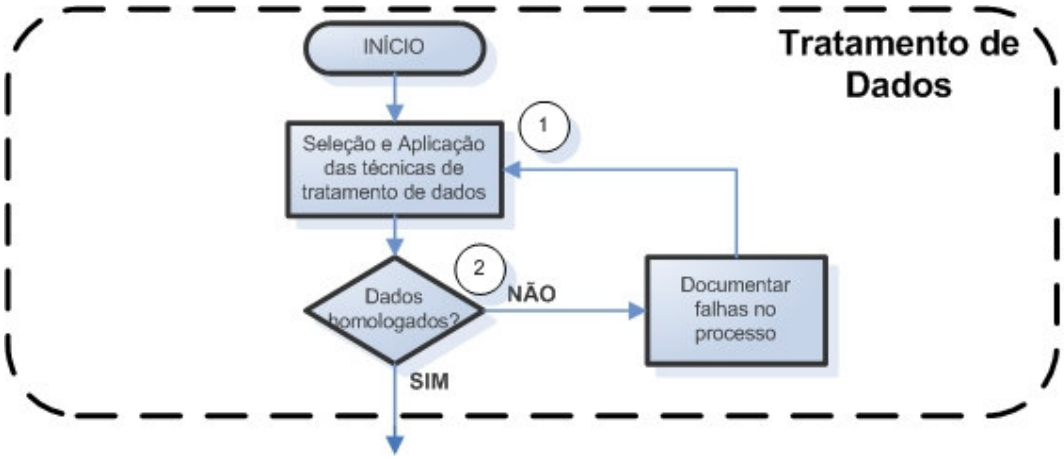
Figura 5.11 – Processo de Integração dos Dados

(1) Foi realizada a integração das bases geradas nos 3 (três) processos anteriores. O objetivo dessa integração é gerar uma tabela desnormalizada em que cada linha representa uma proposta de financiamento. A tabela agrupa informações dos clientes (dados cadastrais e comportamentais), dos avalistas (dados cadastrais) e das propostas. A geração da tabela final é a última atividade da fase de Montagem de Visão.

5.3.4 Tratamento dos Dados

Para a visualização das atividades realizadas nesta fase, os processos do fluxograma apresentado no Capítulo 4 (Figura 4.7) foram desmembrados.

A Figura 5.8 apresenta o processo de Tratamento de Dados. Os círculos numerados representam as atividades realizadas neste estudo de caso. Cada atividade é descrita a seguir conforme numeração da Figura 5.8.



(1) Como a técnica de mineração de dados utilizada neste estudo de caso é RNA, os seguintes tratamentos foram realizados:

- **Limpeza dos dados:** Foi verificada a presença de valores ausentes (*missing*) e valores espúrios (*outliers*), utilizando a ferramenta *Neural Scorer Development*®[5]. Para tratar os valores ausentes, algumas técnicas foram aplicadas. A primeira técnica utilizada foi o preenchimento de um valor específico, tendo em vista o conhecimento do domínio das variáveis e do problema. Alguns atributos (Ex.: SALARIO_CLI, SALARIO_CONJUGE_CLI e ATRASO_MEDIO) possuíam a ausência de valor como indicativo de um

[5] ® *Neural Scorer Development* é propriedade da empresa NeuroTech. Todos os Direitos Reservados.

valor fixo (zero) que foi utilizado no preenchimento. Além dessa técnica, foi aplicada a substituição pela média (Ex.: IDADE_CLI, VALOR_COMPRA e VALOR_ENTRADA). Para tratar *outliers*, utilizou-se a técnica de *winsorizing*, ou seja, foram estipulados valores mínimos e máximos aceitáveis para variável e todo valor que estava fora daquela faixa, era jogado para o extremo mais próximo;

- **Redução de dimensionalidade de variáveis:** As variáveis foram pré-selecionadas com o auxílio do especialista do domínio;
- **Casamento de padrões (S*tring Matching*):** Algumas variáveis categóricas apresentaram problemas de preenchimento manual de valores. Um exemplo é a variável CIDADE_RESIDENCIAL_CLI (cidade residencial). Como é um campo de digitação manual, valores como "São Paulo", "Sao Paulo", "SP" "S.Paulo" representam a mesma informação. Um algoritmo de Força Bruta foi utilizado para realizar o casamento de padrões dessas variáveis;
- **Mudança de formato:** Todos os atributos numéricos foram normalizados utilizando a técnica de Min-Max. Como resultado, todos os valores ficaram dentro do intervalo [0, 1]. Para os atributos categóricos foram usadas duas técnicas de codificação binária. A codificação 1 de N foi aplicada para os atributos cujo conjunto de valores possíveis era pequeno (Ex.: SEXO_CLI, TIPO_FINANCIAMENTO, ORIGEM e ESTADO_CIVIL_CLI). A codificação M de N foi aplicada para os atributos cujo conjunto de valores possíveis era extenso (Ex.: CEP_RESIDENCIAL_CLI3, BAIRRO_RESIDENCIAL_CLI, CIDADE_RESIDENCIAL_CLI e NATURAL_CLI). Os atributos passaram a ser denominados da seguinte forma: *nome_do_atributo* + *n*, onde n varia de 1 (um) até o tamanho do vetor binário resultante da codificação.

(2) Após a geração das novas variáveis, foi realizado o processo de homologação. Esse processo foi realizado e nenhum erro foi encontrado.

5.3.5 Processos Extras

Essa é a fase onde as atividades que estão fora do escopo da metodologia *DMBuilding* devem ser documentadas, pois podem servir de melhoria futura para esta metodologia.

Para o estudo de caso em questão, nenhuma atividade extra foi realizada, pois a metodologia proposta conseguiu englobar todas as atividades necessárias para a realização do projeto.

5.3.6 Avaliação de Desempenho

Esta fase não faz parte do escopo da metodologia *DMBuilding*. Porém, foi realizada com o intuito de apresentar os resultados reais de solução de um problema de classificação de padrões em uma aplicação relevante, a análise de risco de crédito, e incluir o processo de KDD na solução do problema com a montagem da visão produzida a partir da metodologia proposta.

Ao término da fase de tratamento dos dados, a base passou a possuir 335 atributos de entrada e 2 atributos de saída. Antes de aplicar a técnica de mineração de dados, a base foi particionada utilizando o método *Holdout* [Beale & Jackson 1991], onde os dados são aleatoriamente divididos em conjuntos independentes (treinamento e teste). Normalmente, dois terços dos dados são alocados para o conjunto de treinamento e o restante para o conjunto de teste. Essa técnica foi escolhida devido à extensão da base e por ser recomendada pela comunidade científica para investigação experimental de modelos neurais [Prechelt 1994]. Portanto, os padrões foram divididos em três conjuntos independentes: treinamento, com 50% dos dados; validação, com 25% dos dados; e teste, com os 25% restantes. A Tabela 5.5 mostra o total de registros e a distribuição das classes em cada conjunto.

Tabela 5.5 – Particionamento dos dados

Conjunto de dados	Total de registros	BOM	MAU
Treinamento	19.303	17.456	1.847
Validação	9.652	8.728	924
Teste	9.652	8.728	924

Como pode ser observado na Tabela 5.5, existe uma diferença acentuada entre a quantidade de exemplos de classes diferentes. Isso pode prejudicar o processo de aprendizagem [Amorim 2004]. A repetição de exemplos de treinamento das classes menos numerosas e a geração de exemplos ruidosos são algumas técnicas que podem ser usadas para minimizar este tipo de problema [Bigus 1996]. Neste estudo de caso, optou-se por repetir os registros da classe MAU várias vezes nos conjuntos de treinamento até igualar à quantidade de registros da classe BOM [Conde 2000]. Desta forma, o conjunto de treinamento passou a ter 34.912 registros e o de validação 17.456.

Um modelo de RNA chamado MLP (*Multilayer Perceptrons*), treinado com o algoritmo de retropropagação, foi escolhido para etapa de MD [Rumelhart et al. 1986] [Haykin 1999]. Essa técnica foi selecionada devido à excelente capacidade de generalização e à habilidade de realizar aproximação universal de funções [Cunha 2005].

Também chamadas de neurônios artificiais, as redes MLP são constituídas por unidades de processamento simples, que possuem funcionalidades semelhantes àquelas apresentadas pelos neurônios biológicos do cérebro humano. Essas redes possuem três tipos de camada: camada de entrada (responsável pela propagação dos valores de entrada para as camadas seguintes); camadas intermediárias ou escondidas (funcionam como extratoras de características cujos pesos são uma codificação das características presentes nos exemplos de entrada); e camada de saída (recebe os estímulos das camadas intermediárias e fornece a resposta da rede) [Two Crows 1999].

O algoritmo de retropropagação (*backpropagation*) funciona em duas fases: na primeira (fase *forward*), um exemplo é apresentado à camada de entrada, as unidades de processamento computam, camada por camada, funções de ativação até que a resposta seja produzida pela camada de saída; na segunda (fase *backward*), a resposta fornecida pela rede é comparada com a saída desejada para o exemplo atual. Os erros são então propagados a partir da camada de saída até a camada de entrada e os pesos das conexões entre as camadas são ajustados de tal modo que é minimizado o quadrado da diferença entre as saídas geradas e almejadas. O processo é repetido por várias iterações (ou épocas) para todo o conjunto de treinamento até que a saída da rede seja próxima à saída desejada. O treinamento é comumente interrompido quando é alcançado um erro suficientemente baixo, um número máximo de ciclos ou quando o poder de generalização (calculado a partir do conjunto de validação) da rede começa a decrescer [Mitchell 1997]. Os pesos são ajustados através de uma extensão do método gradiente descendente, visando à minimização da soma do erro quadrático, que pode ser descrito pela Equação 5.1:

$$\Delta w_{ij}(n) = - \eta \frac{\partial E}{\partial w_{ij}} \qquad (5.1)$$

onde η é a taxa de aprendizagem e $\partial E/\partial w_{ij}$ é a derivada parcial do erro E em relação ao peso w_{ij}. Como o gradiente especifica a direção e o sentido que produz o maior aumento na taxa de erro e o objetivo é mover o vetor de pesos na direção que minimize esta taxa, utiliza-se o sinal negativo antes da derivada. η assume valores no intervalo [0,1] e controla a magnitude dos ajustes dos pesos. Assim, a taxa de aprendizagem afeta a velocidade de convergência e a

estabilidade da rede durante o treinamento. A fim de evitar grandes oscilações nos valores dos pesos, normalmente η assume valores pequenos.

Foram realizadas várias simulações. Os parâmetros utilizados para a realização dessas simulações estão descritos na Tabela 5.6.

Tabela 5.6 – Parâmetros de treinamento da rede MLP

Parâmetros	Valores
Algoritmo	*Backpropagation*
Taxa de Aprendizagem	0,1; 0,01; 0,001
Camadas Escondidas	1
Neurônios Escondidos	2, 3, 5
Número de Iterações	100, 1.000 e 10.000

Além do critério de parada pelo número máximo de ciclos, foi utilizado o critério por perda de generalização (GL – *Generalization Loss*) igual a 5%. O GL, definido pela Equação 5.2, encerra o treinamento quando o erro de validação aumenta 5% com relação ao menor erro até o ciclo atual [Prechelt 1994]. Este critério diminui as chances de ocorrer *overfitting* e fornece uma visão da capacidade de generalização da rede, pois o seu valor é baseado no conjunto de validação.

$$GL(época\ t) = 100 * \left(\frac{SSE\ atual\ de\ validação}{SSE\ mínimo\ de\ validação} - 1 \right) \qquad (5.2)$$

A escolha da melhor configuração foi baseada no desempenho (erro de classificação) de cada rede no conjunto de validação. A rede com menor erro de classificação no conjunto de validação possuía 3 (três) neurônios na camada intermediária e taxa de aprendizagem de 0,001. A Tabela 5.7 apresenta a matriz de confusão para o conjunto de teste.

Tabela 5.7 – Matriz de confusão do conjunto de teste

	BOM	MAU	Total
Classificado BOM	6261	306	6567
Classificado MAU	2467	618	3085
Total	8728	924	9652

Na Tabela 5.8 é apresentado um resumo dos erros. O erro do Tipo I ocorre quando um BOM cliente é classificado como MAU (erro mais comumente aceitável). O erro do Tipo II ocorre quando um MAU cliente é classificado como BOM (erro considerado mais caro). Por último, é informado o MSE (*Mean Squared Error* ou erro quadrático médio) [Prechelt 1994].

Tabela 5.8 – Erros do Tipo I, II e MSE

Erro Tipo I	79,96%
Erro Tipo II	4,66%
MSE	18,44%

Segundo [Adriaans & Zantinge 1996] e [Monard & Baranauskas 2003], o desequilíbrio na distribuição das classes (9,57% de maus e 90,43% de bons) exige que o desempenho dos modelos apresente um erro máximo inferior a 9,57% (ou taxa de classificação mínima superior a 90,43%). Tal exigência não foi satisfeita (taxa de classificação mínima foi igual a 71,27%) devido aos seguintes fatores: a base de dados utilizada contém apenas informações parciais (informações a respeito das propostas aprovadas e que no futuro os clientes vieram a se tornar bons ou maus) e o problema investigado é de larga escala e envolve dados reais com múltiplos atributos relacionados. Tais características tornam o problema bastante complexo, dificultando a aprendizagem.

Os resultados obtidos foram avaliados, verificando os dois critérios de sucesso do projeto: manter a aprovação superior a 70% e reduzir a inadimplência (inferior a 7%). Em ambos os critérios, os resultados alcançados foram satisfatórios.

5.5 Considerações Finais

Este capítulo apresentou a aplicação da metodologia *DMBuilding* em um estudo de caso de análise de crédito (um problema real e de larga escala), expondo as principais características de um projeto de Montagem de Visão de Dados, assim como as dificuldades envolvidas.

Várias pessoas foram envolvidas na realização do projeto. Para garantir um bom entendimento em todas as atividades, foi necessária uma interação intensa entre os participantes.

Ocorreram alguns retrocessos a atividades anteriores, o que caracterizou o enfoque desta metodologia com processos que podem ocasionar falhas e até o abortamento do projeto.

Após a aplicação da metodologia *DMBuilding* neste estudo de caso, ficou clara a necessidade da utilização de um software que auxilie todo o processo de documentação ou de um software mais amplo, que além da documentação, auxilie em todas as atividades. Para a realização do estudo de caso foi necessária a utilização de diversas ferramentas.

Capítulo 6

Conclusões e Trabalhos Futuros

KDD é uma área que evoluiu, e continua evoluindo, da interseção de pesquisas em campos como banco de dados, aprendizado de máquina, reconhecimento de padrões, análise estatística, inteligência artificial, aquisição de conhecimento para sistemas especialistas, visualização de dados, recuperação de informação e computação de alto desempenho. Esta evolução deve-se principalmente à necessidade cada vez maior das organizações de aplicar métodos computacionais de análise e extração automática (ou semi-automática) de conhecimento a partir de grandes volumes de dados.

Preparação dos Dados é a etapa do processo de KDD que normalmente consome mais tempo e esforço e que melhor contribui para o resultado final do projeto, pois, quanto mais completa e consistente for a preparação, melhor será o resultado da MD. Uma forma de garantir a completude e a consistência dos dados é utilizar uma metodologia que aborde de forma detalhada todas as atividades relacionadas à preparação dos dados.

Muitas metodologias para o desenvolvimento de projetos de KDD foram propostas. Apesar da maioria das metodologias citarem o processo de preparação dos dados, poucas metodologias específicas para montagem de visão de dados têm sido desenvolvidas.

A metodologia proposta desta dissertação, *DMBuilding,* foi inspirada no crescente interesse na área de KDD e na escassez de metodologias específicas para montagem de visão em bases de dados dirigidas a problemas de mineração de dados.

Este capítulo apresenta uma análise do trabalho realizado nesta dissertação de acordo com os objetivos propostos no Capítulo 1, os objetivos alcançados, as contribuições, as limitações, as perspectivas de trabalhos futuros e as considerações finais sobre o trabalho desenvolvido.

6.1 Objetivos Propostos e Alcançados

Os objetivos propostos inicialmente foram:

- Investigar as metodologias para o desenvolvimento de projetos de KDD, enfatizando os aspectos relacionados à preparação dos dados;
- Como resultado da investigação, propor uma metodologia para montagem de visões em bases de dados dirigidas a problemas de MD, com foco em dados cadastrais e comportamentais;
- Aplicar a metodologia proposta em um problema de larga escala no domínio de análise de crédito.

As próximas seções analisam os resultados alcançados em relação a cada um dos objetivos propostos.

6.1.1. Investigação das Metodologias para KDD

Uma parte considerável deste trabalho foi dedicada ao estudo de KDD e de metodologias para o desenvolvimento de projetos de KDD. Durante a execução desta fase, foi realizada uma investigação teórica das principais características, vantagens e deficiências das metodologias propostas, enfatizando os aspectos relacionados à preparação dos dados. Esta investigação foi realizada com o objetivo de identificar as metodologias que seriam usadas como base para o desenvolvimento da metodologia proposta nesta dissertação.

O resultado deste estudo está relatado nos capítulos iniciais da dissertação. O Capítulo 2 descreve detalhadamente todas as etapas do processo de KDD e destaca algumas das principais áreas de aplicação de KDD. O Capítulo 3 descreve algumas metodologias de KDD ou de preparação de dados, suas características e onde essas metodologias são falhas no processo de preparação de dados.

6.1.2. Metodologia Proposta: *DMBuilding*

Este trabalho se propôs, como objetivo principal, desenvolver uma metodologia para montagem de visões em bases de dados dirigidas a problemas de mineração de dados (*DMBuilding*). A metodologia desenvolvida englobou, de forma detalhada todo processo de preparação dos dados, desde o entendimento do problema até a geração da base.

A metodologia *DMBuilding* é fortemente baseada nos fundamentos das principais metodologias de KDD (proposta por Fayyad et al, CRISP-DM e DMEasy) e nas metodologias que são mais focadas na área de preparação dos dados (abordagens de Yu et al. e da empresa Quadstone). As três primeiras metodologias são voltadas para o desenvolvimento de projetos de KDD, englobando todas as etapas do processo. As demais são mais focadas no processo de preparação dos dados.

Levando em consideração as atividades relacionadas ao processo de Montagem da Visão em base de dados, a metodologia *DMBuilding* possui as principais características das metodologias investigadas.

Da abordagem de Fayyad et al., suas características foram incorporadas por ser a primeira metodologia a possuir como foco a solução do problema de KDD como um todo, e não simplesmente dar importância aos resultados obtidos com a aplicação das técnicas de mineração de dados, assim como dar a importância à fase de tratamento de variáveis, a fim de apresentar as variáveis no formato correto para a aplicação de alguma técnica de MD.

Da metodologia CRISP-DM, suas características foram incorporadas por dar importância da fase de levantamento e definição do problema e por fazer menção a necessidade da verificação dos dados para garantir a qualidade dos mesmos.

As principais características da metodologia DMEasy incorporadas à metodologia *DMBuilding* foram a especificação mais detalhada dos processos com relação as outras metodologias, o suporte à documentação do processo e a ênfase a especificação do negócio associado ao problema.

A abordagem de Yu et al., por ser mais focada na fase de preparação dos dados, possui quatro características importantes incorporadas à metodologia proposta, por levar em consideração a aquisição do conhecimento do domínio para o entendimento do problema, a importância da fase de integração de dados, a importância da homologação dos dados que visa à qualidade dos mesmos, a especificação de problemas relacionados aos dados e a apresentação de uma série de técnicas de processamento que podem ser aplicadas para tratamento dos dados.

Por fim, uma característica da metodologia da empresa Quadstone foi incorporada à metodologia *DMBuilding* por tratar a modelagem de variáveis comportamentais, que são de extrema importância para resolução de problemas de KDD.

A metodologia *DMBuilding* aborda as principais características das metodologias investigadas, como também a abordagem com detalhes dos seguintes aspectos não focados em outras metodologias:

- Trata com detalhes a maioria das atividades realizadas na fase de preparação dos dados, como a integração, a transformação e o tratamento dos dados;
- Relaciona o conjunto de problemas relacionados às bases de dados e realiza sugestões de correções e transformações para o enriquecimento da base de dados final;

A metodologia proposta é abordada, em detalhes, no Capítulo 4.

6.1.3. Estudo de Caso Investigado

O estudo de caso desta dissertação, descrito no Capítulo 5, compreendeu uma investigação bastante detalhada da aplicação da metodologia *DMBuilding* em um problema de larga escala no domínio de análise de crédito. O estudo de caso abrangeu praticamente todas as fases, processos e atividades da metodologia.

O domínio aplicado como estudo de caso foi o de análise de crédito ao consumidor, um problema de classificação que define a aprovação ou não de crédito a um determinado solicitante, considerando suas características pessoais e financeiras. Este domínio foi escolhido por se tratar de um problema de larga escala, envolver dados reais com múltiplos atributos relacionados e ser de interesse de diversas instituições e empresas. Tais características permitiram a verificação da viabilidade prática da metodologia.

A base de dados utilizada no estudo de caso era composta de 38.607 registros, dos quais 3.706 (9,6%) são classificados como inadimplentes (MAU) e 34.901 (90,4%) como adimplentes (BOM). Esta base de dados possui os dados fornecidos pelos solicitantes no momento da solicitação de crédito a uma instituição financeira especializada em financiamento de produtos no Brasil. Essas informações são utilizadas pela empresa para decidir pela concessão ou não de crédito ao solicitante. Todos os clientes que obtiveram a aprovação do crédito foram armazenados na base. Com o passar do tempo, alguns desses clientes, que foram considerados bons pagadores pelo sistema decisório da operadora, se tornaram maus pagadores. Portanto, o problema aplicado ao estudo de caso contém informações parciais, pois a base de dados possui apenas informações a respeito dos proponentes aceitos e que vieram a se tornarem adimplentes ou inadimplentes na carteira de clientes da empresa. Tal característica tornou o problema ainda mais complexo.

A maior dificuldade encontrada na execução do estudo de caso foi a obtenção de uma base de dados consistente. As homologações realizadas inicialmente detectaram diversas

inconsistências, sendo necessário que o cliente re-gerasse a base várias vezes, o que resultou no atraso do cronograma do projeto. Um fator que contribuiu para ganho de tempo durante a realização do projeto foi a utilização da ferramenta *Neural Scorer Development*®[6] em diversas atividades da metodologia.

6.2 Contribuições

As principais contribuições desta dissertação são:

- Pesquisa extensiva sobre KDD e metodologias para o desenvolvimento de projetos de KDD, enfatizando os aspectos relacionados à preparação dos dados.

  Durante a execução desta fase, foram realizadas uma investigação teórica das principais características, vantagens e deficiências das metodologias para o desenvolvimento de projetos de KDD, enfatizando os aspectos relacionados à preparação dos dados, e a identificação das metodologias mais promissoras para serem utilizadas como base para a metodologia proposta nesta dissertação. O resultado deste estudo está relatado nos capítulos iniciais da dissertação e pode ser consultado como uma referência nas áreas de KDD e Metodologias para Projetos de KDD;

- Desenvolvimento de uma nova metodologia, *DMBuilding*, para montagem de visões em bases de dados dirigidas a problemas de mineração de dados, com foco em dados cadastrais e comportamentais.

  A principal contribuição desta dissertação é a proposição de uma metodologia para montagem de visões em bases de dados dirigidas a problemas de mineração de dados. A metodologia proposta é fortemente baseada nos fundamentos das principais metodologias de KDD (proposta por Fayyad et al, CRISP-DM e DMEasy) e nas metodologias que são mais focadas na área de preparação dos dados (abordagens Yu et al. e da empresa Quadstone). A proposta desta dissertação foi motivada pelo crescente interesse na área de KDD e pela escassez de metodologias específicas para montagem de visão em bases de dados dirigidas a problemas de mineração de dados.

- Demonstração da aplicabilidade da metodologia em um problema real e de larga escala;

[6] ® *Neural Scorer Development* é propriedade da NeuroTech. Todos os Direitos Reservados.

Diferente de muitos trabalhos que utilizam dados artificiais ou problemas simples para validar suas proposições, a demonstração da aplicabilidade da metodologia proposta foi realizada usando uma base de dados extensa e de alta dimensionalidade. Além disso, o domínio escolhido, o de análise de crédito, representa um problema real que possui múltiplos atributos relacionados. Tais características permitiram a verificação da viabilidade prática da metodologia proposta.

- Desenvolvimento de um padrão de Gestão Empresarial para tratamento da informação.

6.3 Limitações

A principal limitação dessa metodologia é de ter abordado somente as principais técnicas de transformação e tratamento de dados existentes na literatura. Essa limitação se deu, pois existem técnicas que são muito específicas para determinados algoritmos de MD.

A metodologia pode ser considerada genérica e pode ser utilizada em vários problemas de MD, mas o fato de existirem tratamentos específicos de algum algoritmo que não esteja especificado nessa metodologia, pode ser documentado através dos processos extras e servirá como apoio para a constante atualização da metodologia. Além disso, o maior foco é para a aprendizagem supervisionada, necessitando de alguns ajustes para a aprendizagem não supervisionada.

O não desenvolvimento de um software de apoio à documentação também pode ser considerado uma limitação dessa metodologia. Outra proposta seria a construção de um software que abrangesse todas as fases da metodologia, servindo de ferramenta principal para a realização da montagem da visão.

Além dessas limitações, essa metodologia está focada principalmente em problemas baseados em dados transacionais e cadastrais, além de ser focada principalmente em problemas de aprendizagem supervisionada.

6.4 Trabalhos Futuros

Considerando os resultados alcançados nesta dissertação, os seguintes trabalhos futuros podem ser especificados:

- Estudo de novas técnicas de tratamento e transformação de dados já utilizadas na literatura e não citadas em nenhuma das fases da metodologia proposta;
- Desenvolvimento de um software de apoio ao emprego sistemático e interativo da metodologia *DMBuilding*, que tem por finalidade principal à documentação de todo o processo de Montagem da Visão;
- Criação de um software mais amplo que servirá de apoio a todas as fases envolvidas em um projeto de Montagem da Visão, incluindo documentação de todas as atividades do projeto, análise exploratória dos dados, bem como atividades relacionadas à fase de Montagem de Visão e de Tratamento dos Dados, para que não seja necessário o uso de diversas ferramentas para o desenvolvimento desses projetos.

Referências Bibliográficas

[Abeles et al. 2003] ABELES, G., KRISHNAN, R., HORNICK, M., MUKHIN, D., TANG, G., THOMAS, S., VENKAYALA. S., **Oracle® Data Mining - Application Developer's Guide**, Oracle Corporation, E.U.A., 2003

[Adriaans & Zantinge 1996] ADRIAANS, P., ZANTINGE, D. **Data Mining,** Addison-Wesley, 1996.

[Alluri 2005] ALLURI, N.R., **Evaluation of Data Mining Methods to support Data Warehouse Administration and monitoring in SAP Business Warehouse**, Dissertação de Mestrado (Ciência da Computação), Universidade de Ciências Aplicadas - Furtwangen, Alemanha, 2005.

[Almeida 1995] ALMEIDA, F. C., **Desvendando o uso de redes neurais em problemas de administração de empresas**, Revista de Administração de Empresas, p.46-55, São Paulo, 1995.

[Amorim 2004] AMORIM, B. P., **Desenvolvimento de uma plataforma híbrida para descoberta de conhecimento em bases de dados,** Dissertação de Mestrado (Ciência da Computação), Universidade Federal de Pernambuco, Recife, 2004.

[Amorim et al. 2007] AMORIM, B., VASCONCELOS, G. C., BRASIL, L., **Hybrid neural systems for large scale credit risk assessment applications,** Journal of Intelligent and Fuzzy Systems (JIFS), v. 18, p.455-464, 2007.

[Bain et al. 2002] BAIN, T. et al. **Professional SQL Server 2000 Dara Warehousing with Analysis Services**, Wrox Press Ltd, 2002.

[Baragoin et al. 2001] BARAGOIN, C., ANDERSEN, C. M., BAYERL, S., BENT, G., LEE, J., SCHOMMER, C., **Mining your own business in Retail,** Redbooks, IBM, 2001.

[Batista 2003] BATISTA, G. E., **Pré-processamento de Dados em Aprendizado de Máquina Supervisionado,** Tese de Doutorado (Instituto de Ciências Matemáticas e de Computação), ICMC-USP, 2003.

[Batista & Monard 2003a] BATISTA, G. E., MONARD, M. C., **Experimental Comparison of K-Nearest Neighbor and Mean or Mode Imputation Methods with the Internal Strategies Used by C4.5 and CN2 to Treat Missing Data**, Technical Report 186, ICMC USP, 2003.

[Batista & Monard 2003b] BATISTA, G. E., MONARD, M. C., **An Analysis of Four Missing Data Treatment Methods for Supervised Learning**, Applied Artificial Intelligence, vol. 17, p. 519-533, 2003.

[Beale & Jackson 1991] BEAL, R., JACKSON, T. **Neural Computing: An Introduction**, A. Hilger, 1991.

[Berka 2002] BERKA, P., **Discretization And Grouping Operators**, Mining data with the MiningMart system -- Evaluation Report, República Checa, 2002.

[Berry & Linoff 2004] BERRY, M. J. A., LINOFF, G. **Data mining techniques, For Marketing, Sales, and Customer Relationship Management,** John Wiley, 2004.

[Bigus 1996] BIGUS, J. P., **Data Mining with Neural Networks: Solving Business Problems from Application Development to Decision Support,** McGraw-Hill, 1996.

[Boscarioli 2005] BOSCARIOLI, C., **Pré-processamento de Dados para Descoberta de Conhecimento em Banco de Dados: Uma Visão Geral,** Anais do III CONGED - Congresso de Tecnologias para Gestão de Dados e Metadados do Cone Sul. Guarapuava: Unicentro Editora, v. I, p. 101-120, 2005.

[Brito & Neto 2005] BRITO, G. A. S., NETO, A. S., **Modelo de Classificação de Risco de Crédito de Grandes Empresas,** In: 5º Congresso USP de Controladoria e Contabilidade, São Paulo, 2005.

[Burns et al. 2004] BURNS, A.; KUSIAK, A.; LETSCHE, T. **Mining Transformed Data Sets,** Knowledge-Based Intelligent Information and Engineering Systems, vol.3213/2004, p.148-154, 2004.

[Cabena et al. 1997] CABENA, P., HADJINIAN, P., STADLER, R., VERHEES, J., ZANASI, A. **Discovering Data Mining – From Concept to Implementation,** Prentice Hall, 1997.

[Caouette et al. 2000] CAOUETTE, J., ALTMANO, E., NARAYANAN, P., **Gestão do Risco de Crédito**, Qualitymark, Rio de Janeiro, 2000.

[Carvalho 1999] CARVALHO, D. R., **Data Mining através de indução de regras e algoritmos genéticos**, Dissertação de Mestrado (Programa de Pós-Graduação em Informática Aplicada), PUC-PR, 1999.

[Chain & Stolfo 1998] CHAN, P. K., STOLFO, S. J., **The Effects of Training Class Distributions on Performance Using Cost Models**, Intl. Conf. Machine Learning, 1998.

[Chapman et al. 2000] CHAPMAN, P. et al., **CRISP-DM 1.0, Step-by-step data mining guide**, CRISP-DM Consortium, 2000.

[Charras & Lecroq 1997] CHARRAS, C., LECROQ, T. **Exact String Matching Algorithms**, Laboratoire d'Informatique de Ronen, Université de Ronen. France 1997.

[Conde 2000] CONDE, G. A. B. **Análise comparativa de Redes Neuro-difusas para Classificação de Padrões e Extração de Regras**, Dissertação de Mestrado (Ciência da Computação), Universidade Federal de Pernambuco, Recife, 2000.

[Cougo 1999] COUGO, P., **Modelagem Conceitual e Projeto de Banco de Dados**, Ed. Campus, Rio de Janeiro, 1999.

[Coutinho 2003] COUTINHO, F. V., **Data Mining**, DWBrasil, 2003. Disponível em http://www.dwbrasil.com.br/html/dmining.html. Ultimo Acesso: 16/08/2007.

[Cunha 2005] CUNHA, R. C. L. V., **DMEasy: Uma metodologia para Desenvolvimento de Soluções em Mineração de dados,** Dissertação de Mestrado (Ciência da Computação), Universidade Federal do Pernambuco, Recife, 2005.

[Cunico 2005] CUNICO, L. H. B., **Técnicas em data mining aplicadas na predição de satisfação de funcionários de uma rede de lojas do comércio varejista**, Dissertação de Mestrado (Ciências da Computação), Universidade Federal do Pernambuco, Recife, 2005.

[Curoto 2003] CUROTTO, C. L. **Integração de Recursos de Data Mining com Gerenciadores de Bancos de Dados Relacionais**, Tese de Doutorado (Ciência da Computação), Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2003.

[Dasu & Johnson 2003] DASU, T., JOHNSON, T., **Exploratory Data Mining and Data Cleaning,** John Wiley & Sons, 2003.

[Deschaine et al. 2001] DESCHAINE, L.M., MCCORMACK J., PYLE, D., FRANCONE, F. **Genetic Algorithms and Intellingen Agents Team Up: Techniques for Data Assembly, Preprocessing, Modeling and Optimizing Decisions,** PCAI Magazine, 2001.

[Dias 2001] DIAS, M. M., **Um Modelo de Formalização do Processo de Desenvolvimento de Sistemas de Descoberta de Conhecimento em Banco de Dados**, Tese Doutorado (Engenharia da Produção), Universidade Federal de Santa Catarina, Florianópolis, 2001.

[Domingos & Pazzani 1997] DOMINGOS, P., PAZZANI, M., **On the optimality of the simple bayesian classifier under zero-one loss**. Mach. Learn., vol. 29, p. 103–130, 1997.

[Duda et al. 2002] DUDA, O. R., HART, E. P., STORK, G. D., **Pattern Classification**. John Wiley and Sons, New York, 2002.

[Fayyad et al. 1996a] FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **From Data Mining to Knowledge Discovery in Databases**, AI Magazine, vol.17, p. 37-54, 1996.

[Fayyad et al. 1996b] FAYYAD, U., PIATETSKY-SHAPIRO, G., SMYTH, P., **The KDD process for extracting useful knowledge from volumes of data,** Communications of the ACM, vol.39, p.27-34, 1996.

[Fayyad & Irani 1993] FAYYAD, U., IRANI, K., **Multiinterval Discretization of Continuous-Valued Attributes for Classification Learning**, Proc. 13th Int'l Joint Conf. Artificial Intelligence, pp. 1022-1027, 1993.

[Figueiredo 2001] FIGUEIREDO, R. P., **Gestão de Riscos Operacionais em Instituições Financeiras – Uma Abordagem Qualitativa,** Dissertação de Mestrado, Universidade da Amazônia, 2001.

[Freitas 1993] FREITAS, H., **A informação como ferramenta gerencial: um telessistema de informação em marketing para o apoio à decisão**, Ortiz, Porto Alegre, 1993.

[Fuggetta 2000] FUGGETTA, A., **Software process: a roadmap**, Proc. of the Conference on The Future of Software Engineering, ACM Press, p.25-34. 2000.

[Gately 1996] GATELY, E. **Neural Networks for Financial Forecasting**, John Wiley, 1996.

[Goebel & Gruenwald 1999] GOEBEL, M., GRUENWALD, L., **A survey of data mining and knowledge discovery software tools**, ACM SIGKDD, San Diego, v. 1, n. 1, p. 20-33, 1999.

[Gray et al. 1997] GRAY, J., CHAUDHURI, S., BOSWORTH, A., LAYMAN, A., REICHART, D., VENKATRAO, M., PELLOW, F., PIRAHESH, H., **Data cube: A relational aggregation operator generalizing group-by, cross-tab and sub-totals**. Data Mining and Knowledge Discovery, vol.1, p.29-54, 1997.

[Grzymala-Busse & Hu 2000] GRZYMALA-BUSSE, W. J., HU, M., **A Comparison of Several Approaches to Missing Attribute Values in Data Mining**. RSCTC'2000, N. 1, p. 340-347, 2000.

[Gylmour et al. 1997] GYLMOUR, C., M., PREDIBON, D., SMYTH, P., **Statistical themes and lessons for Data Mining,** Data Mining and Knowledge Discovery 1, 11–28, Kluwer Academic Publishers, 1997.

[Han et al. 1997] HAN, J., CHIANG, J. Y., CHEE, S., CHEN, J., CHEN, Q., CHENG, S., GONG, W., KAMBER, M., LIU, G., KOPERSKI, K., LU, Y., STEFANOVIC, N., WINSTONE, L., XIA,

B., ZAIANE, O. R., ZHANG, S., ZHU, H., **DBMiner: A system for data mining in relational databases and data warehouses,** In Proc. CASCON'97: Meeting of Minds, p.249-260, Toronto, Canada, 1997.

[Han & Fu 1994] HAN, J., FU, Y., **Dynamic Generation and Refinement of Concept Hierarchies for Knowledge Discovery in Database**, Proc. AAAI '94 Workshop Knowledge Discovery in Database, p.157-168, 1994.

[Han & Kamber 2006] HAN, J., KAMBER, M. **Data Mining: concepts and techniques,** Morgan Kaufmann, 2006.

[Hand et al. 2001] HAND, D., MANNILA, H., SMYTH, P., **Principles of Data Mining**, The MIT Press, 2001.

[Harrison 1998] HARRISON, T. H., **Intranet data warehouse**, Editora Berkeley Brasil, São Paulo, 1998.

[Haykin 1999] HAYKIN, S., **Neural Networks, A Comprehensive Foundation.** 2nd ed. Prentice Hall, 1999.

[Hofmann & Tierney 2003] HOFMANN, M., TIERNEY, B. **The involvement of human resources in large scale data mining projects,** Proc. of the 1st int. symposium on Information and communication technologies, vol.49, p.103-109, Trinity College Dublin, 2003.

[Inmon 2005] INMON, W. H., **Building the Data Warehouse,** 4 ed., John Willey, 2005.

[Jacobson et al. 1999] JACOBSON, I., BOOCH, G., RUMBAUGH, J., **The unified software development process**, Addison-Wesley, 1999.

[Jain et al. 2000] JAIN, A. K., DUIN, R. P. W., MAO, J. **Statistical pattern recognition: A review,** IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, p.4-37, 2000.

[Jermyn et al. 1999] JERMYN, P., DIXON, M., READ J. B., **Preparing Clean Views of Data for Data Mining,** 12th ERCIM Workshop on Database Research, p.1-15, Amsterdam, 1999.

[John 1997] JOHN, G. H., **Enhancements to the Data Mining Process**, Dissertação de Doutorado (Ciência da Computação), Stanford University, 1997.

[Kimball et al. 1998] KIMBALL, R. et al. **The Data Warehouse,** John Willey, 1998.

[Kimball & Ross 2002] KIMBALL, R. ROSS, M., **The Data Warehouse Toolkit.** J. Willey, 2002.

[Kusiak 2002] KUSIAK, A., **Data Mining and Decision Making**, B.V. Dasarathy (Ed.), Proc. of the SPIE Conf. on Data Mining and Knowledge Discovery: Theory, Tools, and Technology IV, Vol. 4730, SPIE, p.155-165, 2002.

[Lacerda et al. 2003] LACERDA, E. G.; CARVALHO, A. C. P. de L. F. de; LUDERMIR, T. B. Análise de Crédito Utilizando Rede Neurais Artificiais. In: Rezende, S. O. (Coord.) **Sistemas Inteligentes: Fundamentos e Aplicações**, Manole, p. 473-476, Barueri, 2003.

[Larose 2005] LAROSE, D. T., **Discovering Knowledge in Data – An Introduction to Data Mining,** John Willey, 2005.

[Lavraè 1999] LAVRAÈ, N., **Selected techniques for data mining in medicine**. Artificial Intelligence in Medicine, v.16, p.3-23, 1999.

[Lee et al. 1976] LEE, R. C. T., SLAGLE, J. R., MONG, C. T., **Application of Clustering to Estimate Missing Data and Improve Data Integrity**, Proc. Int'l Conf. Software Eng., p. 539-544, 1976.

[Little & Murphy 1987] LITTLE, R. J., MURPHY, P. M., **Statistical Analysis with Missing Data**, John Wiley and Sons, New York, 1987.

[MacKay 2003] MACKAY, D.J.C., **Information Theory, Inference and Learning Algorithms,** Cambridge University Press, 2003.

[Mendes et al. 1997] MENDES, E. F., CARVALHO, A., MATIAS, A., **Credit assessment using evolutionary MLP networks**, Int. Conf. Computional Finance, n.5, 1997, London. REFENES, P. (Ed.) Proceedings of the V Int. Conf. Computational Finance. London: Kluwer Academic Publishers, p.365-371, 1997.

[Michie et al. 1994] MICHIE D., SPIEGELHALTER D. J., TAYLOR, C. C., **Machine Learning, Neural and Statistical Classification.** New York, Ellis Horwood, 1994.

[Mitchell 1997] MITCHELL, T. M. **Machine Learning,** McGraw-Hill, 1997.

[Mitra 2003] MITRA, S., ACHARYA, T., **Data Mining - Multimedia, Soft Computing, and BioInformatics**, John Wiley & Sons, New Jersey, 2003.

[Monard & Baranauskas 2003] MONARD, M. C., BARANAUSKAS, J. A., **Conceitos sobre Aprendizagem de Máquina**, In: REZENDE, S. O. (Coord.) Sistemas Inteligentes: Fundamentos e Aplicações, Manole, Barueri, p.84-114, 2003.

[Monteiro 1999] MONTEIRO, D. S. M. P., **Discovery – Um Ambiente para Descoberta de Conhecimento e Mineração de dados,** Dissertação de Mestrado (Ciência da Computação). Universidade Federal de Pernambuco, Recife, 1999.

[Motta 2007] MOTTA, C. G. L., **Introdução a Técnicas de Data Mining,** Mini-Curso C05, Laboratório Nacional de Computação Científica – LNCC/MCT, Petrópolis, 2007.

[Myatt 2007] MYATT, G. J., **Making Sense of Data: A Practical Guide to Exploratory Data Analysis and Data Mining**, John Wiley & Sons, 2007.

[Navarro 2001] NAVARRO, G., **A guided tour to approximate string matching.** ACM Computing Surveys, vol. 33, p.31–88, 2001.

[NeuroTech 2002] NEUROTECH, **Manual do NeuralScorer Deployment**, NeuroTech, Recife, 2002.

[NeuroTech 2007] NEUROTECH – Site Corporativo. Disponível em: http://www.neurotech.com.br/. Último acesso em: 31/01/08

[Nguyen & Chan 2004] NGUYEN, H. H., CHAN, C. W., **A Comparison of Data Preprocessing Strategies for Neural Network Modeling of Oil Production Prediction,** Proc. Third IEEE Int'l Conf. Cognitive Informatics, p.199-207 , 2004.

[Oliveira 1999] OLIVEIRA, M., **Um método para obtenção de indicadores visando a tomada de decisão na etapa de concepção do processo construtivo: a percepção dos principais intervenientes**, Tese de Doutorado, PPGA/EA, 1999.

[Oracle 2007] Oracle 11g, Siebel, PeopleSoft | Oracle, The World's Largest Enterprise Software Company – Site Corporativo. Disponível em: http://www.oracle.com/global/br/index.html. Último acesso em: 31/01/08

[Paul et al. 2003] PAUL, S., GAUTAM, N., BALINT, R., **Preparing and Mining Data with Microsoft® SQL Server™ 2000 and Analysis Services**, Microsoft Corporation, 2003.

[Petró et al. 2006] PETRÓ, B., MOLOSSI, S., ALTÍSSIMO, T. L., **Fluxo da Informação: recuperação, acesso e uso da informação**, Relatório Técnico, Universidade Federal de Santa Catarina, Florianópolis, 2006.

[PMBOK 2000] **A Guide to the Project Management Body of Knowledge: PMBOK Guide 2000 Edition**, The Project Management Institute Inc, 2000.

[Populus et al. 1998] POPULUS, C. V. K., FOOG, R. J. X., DOWNEY, R. G., **Mean Substitution for Missing Items: Sample Size and the Effectiveness of the Technique**, Technical Report - Kansas State University, vol.13, p.1-20, 1998.

[Prechelt 1994] PRECHELT, L., **Proben1** – A set of benchmarks and benchmarking rules for neural network training algorithms. Relatório Técnico 21/94, Faculdade de Informática, Universidade de Karlsruhe, Alemanha, 1994.

[Pressman 1992] PRESSMAN, R. S., **Software Engeneering: a Practitioner's Approach**, McGraw-Hill, 1992.

[Pyle 1999] PYLE, D., **Data Preparation for Data Mining,** Morgan Kaufmann, 1999.

[Pyle 2003] PYLE, D., **Business Modeling And Data Mining,** Morgan Kaufmann, 2003.

[Quadstone 2003] Portrait Software, **Quadstone Methodology for Customer Behaviour Modelling in Retail,** Internal Report, 2003.

[Quinlan 1993] QUINLAN, R., **C4.5 programs for machine learning**, Morgan Kaufmann, 1993.

[Ragel & Cremilleux 1998] RAGEL, A., CREMILLEUX, B., **Treatment of Missing Values for Association Rules**, Proc. Second Pacific-Asia Conf. Knowledge Discovery and Data Mining, p. 258-270, 1998.

[Ramakrishnan & Gehrke 2002] RAMAKRISHNAN, R., GEHRKE, J., **Database Management Systems**, 3 ed, McGraw-Hill, 2002.

[Razente & Traina 2003] RAZENTE, H. L., TRAINA, C. J., **Análise Visual em Processos de Seleção de Atributos para Mineração em Sistemas de Bases de Dados**, Dissertação de Mestrado (Ciências Matemáticas e de Computação), Universidade de São Paulo - USP, 2004.

[Redpath & Sheard 2005] REDPATH, R., SHEARD, J., **Domain Knowledge to Support Understanding and Treatment of Outliers**, Proc. Int. Conf. on Information and Automation, Sri Lanka, 2005.

[Rezende et al. 2003] REZENDE, S. O. et al. **Mineração de dados**. In: REZENDE, S. O. (Coord.) **Sistemas Inteligentes: Fundamentos e Aplicações**. Manole, Barueri, 2003.

[Romão 2002] ROMÃO, W., **Descoberta de Conhecimento relevante em banco de dados sobre Ciência e Tecnologia**, Tese de Doutorado (Engenharia de Produção), Universidade Federal de Santa Catarina, 2002.

[Rud 2001] RUD, O. P., **Data Mining Cookbook - Modeling Data for Marketing, Risk, and Customer Relationship Management**, John Wiley, 2001.

[Rumelhart et al. 1986] RUMELHART, D. E., HINTON, G. E., WILLIAMS, R. J., **Learning Representations by Backpropagation Errors**. Nature, v.323, p.533-536, 1986.

[Schafer & Olsen 1998] SCHAFER, J. L., OLSEN, M. K., **Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective**, Multivariate Behavioral Research, vol.22, p.545-557, 2003.

[Soibelman & Kim 2002] SOIBELMAN, L., KIM, H., **Data Preparation Process for Construction Knowledge Generation through Knowledge Discovery in Databases,** Journal of Computing in Civil Engineering, American Society of Civil Engineers (ASCE), vol.16, p.39-48, 2002.

[Sousa & Carvalho 1999] SOUSA, H.; CARVALHO, **A. Credit assessment using constructive neural networks**, Proc. of the 3rd Int. Conf. on Computational Intelligence and Multimedia Applications. IEEE, p.40-44, India, 1999.

[SPSS 2007] SPSS, Data Mining, Statistical Analysis Software, Predictive Analysis, Predictive Analytics, Decision Support Systems. Disponível em: http://www.spss.com/. Último acesso em: 31/01/08.

[SQL Server 2000] SQL Server 2000 – Site Corporativo. Disponível em: http://search.microsoft.com/results.aspx?mkt=pt-BR&setlang=pt-BR&q=sql+server+2000. Último acesso em: 31/01/08.

[Srivastava & Chen 1999] SRIVASTAVA, J., CHEN, P. Y. **Warehouse creation – A potential roadblock to data warehousing**, IEEE Transactions on Knowledge and Data Engineering, Vol. 11, No. 1, 1999.

[Theodoridis & Koutroumbas 1999] THEODORIDIS, S. and KOUTROUMBAS, K., **Pattern recognition**, Academic Press. San Diego, 1999.

[Tseng et al. 2003] TSENG, S. M., WANG, K. H., LEE, C. I., **A Preprocessing Method to Deal with Missing Values by Integrating Clustering and Regression Techniques**, Applied Artificial Intelligence, vol. 17, p. 535-544, 2003.

[Two Crows 1999] TWO CROWS CORPORATION, **Introduction to Data Mining and Knowledge Discovery,** 3rd ed, 1999.

[Vasconcelos et al. 1999] VASCONCELOS, G. C.; ADEODATO, P. J.; MONTEIRO, D. S. M., **A neural Network based solution for the credit risk assessment problem**. In: Anais do IV Congresso Brasileiro de Redes Neurais, São José dos Campos, p.269-274, 1999.

[Watson et al. 2003] WATSON, W. H. M. T., D. G. K., B. W., **Standards and agile software development**, Proc. of the 2003 annual research conference of the South African

institute of computer scientists and information technologists on Enablement through technology, p. 178-188, 2003.

[Wei & Tang 2003] WEI, W., TANG,Y., **A Generic Neural Network Approach for Filling Missing Data in Data Mining**, IEEE Int. Conf. On Systems Man and Cybernetics, p. 862-867, 2003.

[Widrow et al. 1994] WIDROW, B.; RUMELHART, D. E.; LEHR, M. A., **Neural networks: Applications in industry, business and science,** Communications of the ACM, vol.37, p.93-105, 1994.

[Witten & Frank 2005] WITTEN, I. H., FRANK, E., **Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations,** Morgan Kaufmann, 2000.

[Wojciechowski 2002] WOJCIECHOWSKI, A., NAWROCKI, J. R. e WLATER, B., **Comparison of CMM Level 2 and eXtreme Programming**, Proc. of the 7th Int. Conference on Software Quality, p. 288-297, Springer-Verlag, 2002.

[Yu et al. 2006] YU, L., WANG, S., LAI, K. K., **An Integrated Data Preparation Scheme for Neural Network Data Analysis**. IEEE Trans. Knowl. Data Eng., vol.18, p.217-230, 2006.

[Zaïne 1999] ZAÏANE, O. R., **Principles of Knowledge Discovery in Databases,** Internal Report, Departamento de Ciência da Computação, Universidade de Alberta, 1999.