UNIVERSIDADE METODISTA DE PIRACICABA

FACULDADE DE CIÊNCIAS EXATAS E DA NATUREZA

MESTRADO EM CIÊNCIA DA COMPUTAÇÃO

MODELAGEM E TESTE DE UMA BASE DE CONHECIMENTO DE INSTRUÇÃO DE MINERAÇÃO DE DADOS RELACIONAIS COM ÊNFASE NA ESCOLHA DA TAREFA DE CLASSIFICAÇÃO

LIDIA MARTINS DA SILVA

ORIENTADORA: PROFª. DRª. ANA ESTELA ANTUNES DA SILVA

PIRACICABA, SP
2010

UNIVERSIDADE METODISTA DE PIRACICABA
FACULDADE DE CIÊNCIAS EXATAS E DA NATUREZA
MESTRADO EM CIÊNCIA DA COMPUTAÇÃO

MODELAGEM E TESTE DE UMA BASE DE CONHECIMENTO DE INSTRUÇÃO DE MINERAÇÃO DE DADOS RELACIONAIS COM ÊNFASE NA ESCOLHA DA TAREFA DE CLASSIFICAÇÃO

LIDIA MARTINS DA SILVA

ORIENTADORA: PROFª. DRª. ANA ESTELA ANTUNES DA SILVA

Dissertação de Mestrado apresentada ao Mestrado em Ciência da Computação, da Faculdade de Ciências Exatas e da Natureza, da Universidade Metodista de Piracicaba – UNIMEP, como parte dos requisitos para obtenção do Título de Mestre em Ciência da Computação.

PIRACICABA, SP
2010

MODELAGEM E TESTE DE UMA BASE DE CONHECIMENTO DE INSTRUÇÃO DE MINERAÇÃO DE DADOS RELACIONAIS COM ÊNFASE NA ESCOLHA DA TAREFA DE CLASSIFICAÇÃO

AUTORA: LÍDIA MARTINS DA SILVA
ORIENTADORA: PROFª. DRª. ANA ESTELA ANTUNES DA SILVA

Dissertação de Mestrado apresentada, em 25 de Fevereiro de 2010, à Banca Examinadora constituída dos Professores:

Profª. Drª. Ana Estela Antunes da Silva
UNIMEP

Profª. Drª. Marina Teresa Pires Vieira
UNIMEP

Prof. Dr. Luiz Camolesi Junior
UNICAMP

PIRACICABA, SP
2010

A DEUS.

À MINHA FAMÍLIA.

AGRADECIMENTOS

A Deus todo Poderoso que me deu saúde e força para prosseguir nesta luta.

Ao meu pai (*in memorian*) e minha mãe que me permitiram ter uma excelente formação de caráter e ética em tudo que faço.

Ao meu esposo Marcos e minha amada filha Amanda, pela compreensão e paciência nos momentos de ausência e pelo incentivo durante esta jornada.

A toda Minha família. Obrigado por vocês existirem. Obrigado por depositarem em mim a confiança em todas as horas. Sei que vocês se orgulham por eu ter atingido mais uma etapa, pois são conhecedores de todas as dificuldades que tive que transpor para conseguir atingir meu objetivo.

À minha orientadora profª. Drª. Ana Estela Antunes da Silva, pela brilhante orientação e principalmente pela paciência, generosidade, pelas críticas e sugestões relevantes feitas durante a orientação.

Aos professores do mestrado: Drª. Marina Teresa Pires Vieira, Pós-Dr. Plínio Roberto de Souza Vilela, Dr. Luiz Eduardo Galvão Martins, Dr. Luiz Camolesi Junior, Dr. Cláudio Kirner, agradeço a partilha do saber e as valiosas contribuições para o meu crescimento acadêmico.

Aos meus amigos do mestrado: Cardoso, Bira, Neto, Luciano, Marco Antônio, Maurício, Andersown e Cleber, pela excelente relação pessoal que criamos e espero que não se perca.

Aos meus colegas de trabalho: José Américo, Bruno, Ozéias, Douglas, Raimyson e Salomão, pela valiosa ajuda nos períodos de ausência.

Aos meus amigos: Alexandre dos Anjos, Aline Domingos, Maria Eloísa, Rodrigo Elias, Aloísio Francisco, Erly, Lourdes, Fernando, Deniz e Kilza, pela força, incentivo e amizade.

Em especial ao meu amigo Emiliano e sua esposa Nitza, que me brindaram com suas amizades e principalmente pela ajuda recebida durante este período.

À Reitora, Luzia Guimarães e prof. Dr. Delarim Martins Gomes, exemplos de força, dignidade, ética e profissionalismo.

Ao Centro Universitário Cândido Rondon, pela ajuda financeira com a qual pude contar durantes meus estudos.

Aos funcionários do departamento de pós-graduação da Unimep, em especial, à Dulce e à Rosa.

Em especial à professora Drª. Marina Teresa Pires Vieira e ao seu orientando Eduardo Fernando Mendes, que tão gentilmente me ofereceu sua dissertação, a qual pude usar como fonte material e de inspiração para o tema de minha dissertação.

Ao colega de grupo de pesquisa André Bindilatti pela oportunidade de compartilhar experiências e ajuda preciosa no desenvolvimento deste trabalho.

Todo agradecimento pode pecar pela ausência de uma ou mais pessoas que tiveram um papel, mesmo que pequeno, na realização deste trabalho, por isso agradeço a todos que direta ou indiretamente contribuíram com a concretização deste sonho.

RESUMO

Este trabalho teve como objetivo unir dois domínios: Representação do Conhecimento e Mineração de Dados, tendo como finalidade a modelagem e teste de uma base de conhecimento de instrução de mineração de dados relacionais com ênfase na tarefa de Classificação. O trabalho tem enfoque instrucional proporcionando uma representação do conhecimento necessário para o entendimento e escolha da aplicação da tarefa de classificação em problemas de mineração de dados. Esta escolha é feita através de perguntas que direcionam o usuário a descobrir se a tarefa de classificação é adequada para ser utilizada em seu domínio de problema. Essas perguntas fundamentam a criação e uso de uma base de conhecimento que representa o conhecimento necessário para a escolha da tarefa de classificação. Para a modelagem da base de conhecimento foram feitas uma rede semântica e regras de produção. No total são 25 regras para representação do conhecimento e onze perguntas específicas sobre a tarefa de classificação com instruções que ajudam os usuários na escolha das respostas às perguntas apresentadas.

Palavras-chave: Representação do Conhecimento, Mineração de Dados, Classificação, Base de Conhecimento.

ABSTRACT

The objective of this work is to integrate two domains: Knowledge Representation and Data Mining. The aim of the work is modeling and testing an instruction knowledge base emphasizing the classification task. The work has an instructional focus which is based on knowledge representation. Such knowledge is necessary for the understanding and choice of the classification task in data mining problems. This choice is obtained through the application of questions which drive the user to discovering whether the classification task is proper to be applied on its problem domain. These questions are the basis for the creation and use of a knowledge base which represents the necessary knowledge for the choice of the classification task. A semantic network and production rules were developed in order to create a model of the knowledge base. This model contains 25 rules which are based on 11 specific questions about the classification task. The questions contains instructions that help users to choose answers to the presented questions.

Keywords: Knowledge Representation, Data Mining, Classification, Knowledge Base.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

AD: ANÁLISE DE DISCURSO

AG: ALGORITMO GENÉTICO

CRISP-DM: *CROSS-INDUSTRY STANDARD PROCESS FOR DATA MINING*

DNA: ÁCIDO DESOXIRRIBO NUCLEICO

DM: *DATA MINING*

IA: INTELIGÊNCIA ARTIFICIAL

ILP: PROGRAMAÇÃO LÓGICA INDUTIVA

KDD: *KNOWLEDGE DISCOVERY IN DATABASES*

RNA: REDE NEURAL ARTIFICIAL

RC: REPRESENTAÇÃO DO CONHECIMENTO

SE: SISTEMAS ESPECIALISTAS

WEKA: *WAIKATO ENVIRONMENT FOR KNOWLEDGE ANALISYS*

1 INTRODUÇÃO

1.1 CONSIDERAÇÕES

A criação de um sistema especialista tem a intenção de capturar o conhecimento e a experiência de especialistas em uma determinada área e emular em um computador o comportamento dos mesmos.

Observa-se como principais vantagens dos sistemas especialistas, o aumento da produtividade, preservação e disseminação do conhecimento existente. Esses sistemas podem ser úteis em variados segmentos, cujas aplicações servem para tomada de decisões.

De acordo com Passos (1989), um sistema especialista é delineado e concebido para atender a uma aplicação determinada e limitada da experiência humana, capaz de manifestar uma decisão, apoiado em conhecimento demonstrado, a partir de uma base de informações, tal qual um especialista de determinada área do conhecimento humano.

Muitas vezes os problemas reais são abordados por uma análise de grupo, em que a tomada de decisão depende de um diagnóstico, da capacitação e treinamento de pessoal. Essa decisão depende de múltiplas avaliações que são funções da competência dos vários peritos em campos específicos. Os sistemas especialistas podem auxiliar na solução desses problemas.

Os sistemas especialistas são desenvolvidos para tratar problemas complexos do mundo real que necessitem da interpretação e da análise de especialistas humanos e que ao mesmo tempo cheguem a conclusões e decisões que o especialista humano chegaria, se estivesse se defrontando com os mesmos problemas. Esses programas computacionais utilizam-se de regras de inferência sobre uma grande base de conhecimentos, sobre sintomas e tratamentos possíveis, para que possam identificar o problema e posteriormente oferecer um tratamento adequado, auxiliando a tomada de decisões (PASSOS, 1989; BARRETO, 2001).

O “coração” de um sistema especialista é sua base de conhecimento. Ela contém conhecimento sob a forma de regras de produção, quadros, redes semânticas, ou seja, de várias formas, contém também a descrição do conhecimento necessário para a resolução do problema abordado na aplicação. A base de conhecimento é o elemento central de um sistema especialista e é responsável por estruturar todo o conhecimento sobre o domínio da aplicação.

Já a tecnologia de mineração de dados é formada por um conjunto de ferramentas que, através do uso de algoritmos específicos, É capaz de explorar um grande conjunto de dados, extraindo, deste, conhecimentos na forma de hipóteses e de regras. Diariamente as empresas acumulam diversos dados em seus bancos de dados, tornando-os verdadeiros tesouros de informação sobre os vários processos e procedimentos das funções da empresa, inclusive com dados e hábitos de seus clientes, suas histórias de sucesso e fracassos. Todos esses dados podem contribuir com a empresa, sugerindo tendências e características relacionadas a ela e seu meio ambiente tanto interno quanto externo, tendo em vista uma célere ação de seus gestores.

1.2 MOTIVAÇÃO

Atualmente existem muitas propostas de sistemas especialistas para a resolução de problemas numa ampla escala de segmentos, tais como medicina, medicina veterinária, agricultura, matemática, química, geologia, economia, direito, educação, ciência da computação, entre outros. Em nossas pesquisas e estudos realizados, o que se observou foi a falta de um sistema especialista sobre mineração de dados e principalmente sobre uma determinada tarefa de mineração de dados.

As ferramentas de mineração de dados atuais não dão apoio ao usuário no entendimento e conseqüente escolha de uma determinada tarefa de mineração de dados; entre as ferramentas existentes pode-se citar aqui a ferramenta WEKA e Oracle Data Mining. O WEKA (Waikato *Environment for*

Knowledge Analisys) é uma ferramenta para mineração de dados com código aberto (*opensource*), que agrega um conjunto de algoritmos de classificação, regras de associação, regressão, pré-processamento e *clustering*, todos implementados em JAVA e foi desenvolvida pela universidade de *Waikato* na Nova Zelândia. Estas ferramentas não têm como enfoque os recursos de instrução sobre as etapas da mineração de dados, apenas de tratamento e visualização de dados.

Uma das ferramentas de mineração de dados que possui um enfoque instrucional é a ferramenta Kira (MENDES, 2009). Este trabalho tem a intenção de colaborar com a ferramenta Kira, no sentido de proporcionar uma representação do conhecimento necessário para o entendimento e escolha da aplicação da tarefa de classificação em problemas de mineração de dados.

Diante da ausência de uma representação do conhecimento sobre mineração de dados, da necessidade de ferramentas de apoio nos processos decisórios de uma empresa, de complementação da ferramenta Kira e a perspectiva de fazer algo novo, surgiu a motivação para fazer este trabalho de modelagem e teste de uma base de conhecimento de instrução para mineração de dados para auxiliar o usuário no momento da escolha de uma tarefa de mineração de dados.

1.3 OBJETIVO DO TRABALHO

O trabalho teve como objetivo modelar e testar uma base de conhecimento para mineração de dados relacionais com enfoque instrucional proporcionando uma representação do conhecimento necessário para o entendimento e escolha da tarefa de classificação em problemas de mineração de dados. A partir da aplicação de técnicas de aquisição do conhecimento, várias perguntas foram elaboradas tendo como objetivo direcionar o usuário a descobrir se a tarefa de classificação é adequada para ser utilizada no seu domínio de problema.

1.4 ESTRUTURA DO TRABALHO

O presente trabalho tem a seguinte estrutura:

No primeiro capítulo são abordados: a introdução, os objetivos do trabalho, metodologia e a estrutura do trabalho.

No segundo capítulo foram abordados os seguintes conceitos: sistemas especialistas, estruturas, classificação, fase de desenvolvimento de sistemas especialistas e base de conhecimento, aquisição de conhecimento, processos de aquisição de conhecimentos, técnicas de aquisição de conhecimento: observação, entrevista, análise de discurso e representação do conhecimento.

Neste trabalho foram aplicadas as técnicas de aquisição do conhecimento, tais como: entrevista com especialista em mineração de dados, realização de estudo de caso utilizando a tarefa de classificação e o algoritmo J48 em um banco de dados real de um Centro Universitário e análise de discurso nos guias da ferramenta Kira, isso pode ser observado nos apêndices A e B deste trabalho.

Optou-se também por utilizar, como forma de representação do conhecimento, redes semânticas e regras de produção, uma vez que são intuitivas aos usuários. A rede semântica procura mostrar, através da linguagem natural e representação visual, o quanto sua descrição se aproxima da realidade, simplificando a forma de representação do problema, em que o ser humano consegue interpretar sem muitas dificuldades.

No terceiro capítulo é apresentada a metodologia CRISP-DM, Mineração de dados, técnicas de mineração de dados, tais como: redes neurais artificiais, algoritmos genéticos e árvore de decisão. É apresentada também a tarefa de classificação e os algoritmos de classificação: ID3, J48 e C4.5. A citação dos três algoritmos deve-se ao fato dos mesmos serem populares e muito utilizados no meio acadêmico e por fazer parte do pacote de ferramentas Weka. O algoritmo J48 foi utilizado em um estudo de caso como processo de aquisição do conhecimento da tarefa de classificação.

O estudo sobre a metodologia CRISP DM deu-se como processo de aquisição de conhecimento, onde o conhecimento adquirido foi utilizado na criação da rede semântica.

O quarto capítulo é composto da apresentação dos resultados obtidos, a conclusão, trabalhos futuros e referências bibliográficas, e por fim são apresentados os apêndices.

1.5 METODOLOGIA

A fim de atingir o objetivo deste trabalho, foram necessários estudos sobre: a estrutura e funcionamento de sistemas especialistas; aquisição de conhecimento, base de conhecimento, regras de produção, mineração de dados; aplicação de análise de discurso nos guias da Ferramenta Kira; entrevistas com especialistas em mineração de dados; fazer análise de discurso para as entrevistas; modelar e testar a base de conhecimento.

Para isso, a metodologia de desenvolvimento seguiu um ciclo de vida clássico com técnicas voltadas ao desenvolvimento de bases de conhecimento. Em uma primeira etapa, foi desenvolvida a fase de aquisição de conhecimento. Essa fase foi realizada com a utilização das seguintes técnicas:

- Entrevistas com especialistas em mineração de dados;

- Estudo sobre a tarefa classificação de mineração de dados;

- Estudo da ferramenta Kira;

- Análise de discurso com especialistas e com a ferramenta Kira.

A segunda fase da metodologia consistiu na modelagem do conhecimento. Essa modelagem é constituída das seguintes técnicas:

- Rede Semântica para representação do conhecimento.

- Regras de Produção para representação do conhecimento.

A terceira fase da metodologia consistiu na criação e teste da base de conhecimento de instrução para mineração de dados utilizando a tarefa de classificação. Para os testes foram criados vários exemplos e os mesmos foram testados manualmente e por meio da ferramenta Chimera (BINDILATTI, 2009).

2 SISTEMAS ESPECIALISTAS

Os sistemas especialistas são programas de computador que usam o conhecimento representado explicitamente para resolver problemas. Esses programas manipulam conhecimento e informação que requerem uma quantidade considerável de conhecimento humano e especializado. Muitas vezes, os problemas reais são abordados por uma análise de grupo, em que a tomada de decisão depende de um diagnóstico, da capacitação e treinamento de pessoal. Essa decisão depende de múltiplas avaliações que são funções da competência dos vários peritos em campos específicos. Os sistemas especialistas podem auxiliar na solução desses problemas.

De acordo com Rezende *et al* (2003), a criação desses sistemas para muitas organizações encontra-se na capacidade dos mesmos em preservar, aproveitar e fazer uso de recursos cada vez mais valiosos: o talento e a experiência dos membros da organização no processo de tomada de decisões. Durante a criação desses sistemas, o conhecimento dos membros da organização necessita ser capturado, organizado e disponibilizado na base de conhecimento, que uma vez construída torna-se permanentemente acessível, facilmente recuperável e pode ser amplamente utilizada por todos, independente de sua capacitação.

Segundo Newell *apud* Rezende *et al* (2003), o desenvolvimento de um SE deve conter a descrição do sistema sob duas perspectivas distintas: a do conhecimento, processável pelo homem e a simbólica, processável pelo computador. Sob a perspectiva do conhecimento, a base restringe-se em descrever o que o sistema deve fazer, enquanto que na perspectiva simbólica; a base deve indicar como o sistema irá proceder. Com esta diferença, o autor enfatizou a importância de separar a análise e modelagem do método de resolução do problema e a atividade de representar este método em um formalismo computacionalmente eficiente.

Para entender esse processo é necessário distinguir dois tipos de operações usadas. A primeira operação é considerada a capacidade de

raciocínio, ou seja, como chegar a certas conclusões interpretando o conhecimento adquirido até o momento. Segundo Newell *apud* Rezende (2003), é cada vez mais reconhecido que o uso restrito dessa capacidade não é suficiente para a resolução adequada de problemas, é importante que se chegue rapidamente a conclusões significativas para resolver um determinado problema e para isso se deve fazer uso da segunda operação, que consiste em guiar o processo de raciocínio. Para a resolução de problemas de forma clara e rápida é indispensável guiar o processo de raciocínio de maneira que apenas conclusões relevantes ao problema em questão sejam consideradas. Essa segunda capacidade é denominada método para resolução de problemas, e o raciocínio é identificado como estratégia de raciocínio ou estratégia de inferência.

Para Beyon-Davis (1991), sistema especialista é um sistema de computação que usa representação de conhecimento ou perícia humana num domínio particular (área de interesse específico) de forma a executar funções semelhantes às de um especialista humano nesse domínio. Para Bittencourt (1998), os sistemas especialistas são projetados para reproduzirem o comportamento de especialistas humanos na resolução de problemas do mundo real, mas o domínio desses problemas é altamente restrito.

Feigenbaum e Barr (1981) consideram SE como sistemas que solucionam problemas no nível de um especialista humano que tenha acumulado um conhecimento exigido na resolução desses problemas.

De acordo com Giarratano e Riley (1998) sistemas especialistas são programas de computador que se utilizam de conhecimento e procedimentos de inferência para resolver problemas bastante complexos que necessitam, para a sua solução, de um conhecimento bastante específico.

Em Barreto (2001), SE são sistemas computacionais que devem apresentar um comportamento semelhante a um especialista em um determinado domínio requerendo conhecimento sobre a habilidade, a experiência e as heurísticas usadas pelo especialista. Seu processo de

desenvolvimento envolve uma profunda interação com o especialista (REZENDE *et al*, 2003).

Segundo Feigenbaum *apud* Harmon (1988), um sistema especialista “é um programa inteligente de computador que usa conhecimentos e procedimentos inferenciais, para resolver problemas que são bastante difíceis, de forma a requererem para sua solução, muita perícia humana” e Flores (2003) define SE como uma forma de sistema baseado no conhecimento especialmente projetado para emular a especialização humana de algum domínio específico.

2.1 ESTRUTURA DE UM SISTEMA ESPECIALISTA

O intuito em usar conhecimento para simular o comportamento dos especialistas humanos é justamente desenvolver programas que possibilitem a utilização dos conhecimentos dos especialistas através de uma máquina que permita o armazenamento e o seqüenciamento de informações e a auto-aprendizagem.

Os sistemas computacionais tradicionais utilizam soluções algorítmicas, sendo aplicáveis a problemas que envolvam a precisão de cálculos numéricos ou que não permitam mudanças no seu comportamento. Caso ocorra uma mudança no negócio, geralmente o programa necessita ser atualizado. Diferentemente, para problemas que não possuam uma solução fechada ou que necessitem de conhecimentos parciais sobre o assunto para gerar uma solução, os sistemas especialistas são apropriados.

De acordo com Barreto (2001), de um modo geral, para construir um sistema especialista é necessário:

- uma fonte do conhecimento – o especialista;

O conhecimento deve ser obtido do especialista, transformado em forma conveniente e armazenado no computador. Este conhecimento

fundamentalmente é de dois tipos: fatos sobre o problema a resolver e regras que mostram como o especialista pensa para chegar a uma conclusão.

Constantemente é necessário dispor de um mecanismo capaz de gerar explicações sobre como o especialista chegou a certa conclusão. Isso é motivado por casos em que o usuário do sistema não concorda totalmente com a sugestão do sistema especialista: ele quer ver qual o raciocínio que foi seguido para se convencer de que o SE tem razão.

Segundo Barreto (2001), as duas primeiras fases, identificação e conceituação são freqüentemente exercidas pelo Engenheiro do Conhecimento o qual emprega várias técnicas de psicologia, freqüentemente usadas intensivamente no início do ciclo de vida do SE, e continua necessária durante todo o ciclo de vida dele para atualizar o conhecimento disponível.

O engenheiro do conhecimento se comunica diretamente com a base de conhecimento, nela colocando a experiência do especialista. Deve ainda monitorar o funcionamento do SE observando-o em funcionamento, de modo a efetuar uma fase de ajustes da base de conhecimentos, indispensável nas fases iniciais de vida do SE.

Na Figura 1 são apresentados os elementos básicos que compõem a arquitetura de um sistema especialista. Verifica-se a existência de uma interface de comunicação com o usuário do sistema, isto é, com a pessoa que irá utilizar o sistema no seu dia-a-dia, e uma interface com o especialista que domina a aplicação. Esta interação é conseguida através do módulo de aquisição de conhecimento, necessário para converter as informações obtidas pelo especialista em uma representação de conhecimento.

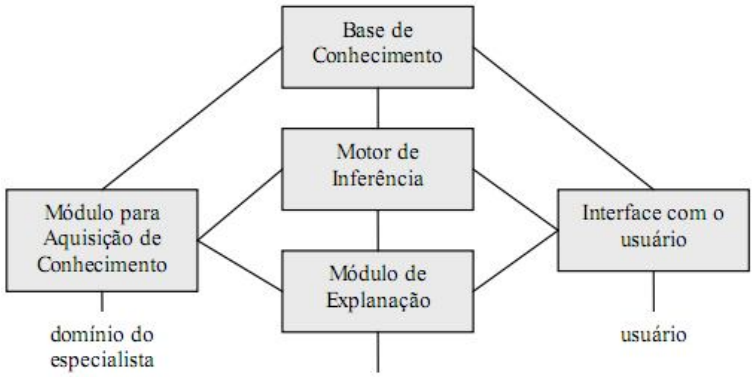

FIGURA 1- ARQUITETURA DE UM SISTEMA ESPECIALISTA
Adaptado de Nikolopoulos (1997)

A arquitetura de um sistema especialista é composta pela base de conhecimento, motor de inferência, módulo para aquisição do conhecimento, módulo de explanação e interface com o usuário.

A base de conhecimento aparece no topo da arquitetura e, segundo Nikolopoulos (1997), é responsável por estruturar todo o conhecimento sobre o domínio da aplicação. Ela é o elemento central de um sistema especialista, pois contêm conhecimento sob a forma de regras de produção, quadros, redes semânticas, ou seja, de várias formas. Uma das mais comuns é por sentenças do tipo “Se – Então”. Ela contém um somatório de fatos, de heurísticas e de crenças.

Para Luger (2005), o motor de inferência aplica o conhecimento para a solução de problemas reais. Ele é essencialmente um intérprete para a base de conhecimento. No sistema de produção, o motor de inferência executa o ciclo de controle reconhecer-agir. Os procedimentos a implementar o ciclo de controle são separados das regras de produção. É importante manter essa separação da base de conhecimento e mecanismo de inferência por várias razões:

- essa separação torna possível representar o conhecimento de uma forma mais natural. SE... ENTÃO ..., por exemplo, estão mais próximas

da forma como os seres humanos descrevem suas habilidades para resolver problemas do que o menor nível de código de computador;

- a base de conhecimento é separada do programa de menor nível de estruturas de controle, os construtores de sistema especialista podem se concentrar na captação do conhecimento e organização de solução de problemas e não nos detalhes de sua implementação em computador;

- idealmente, a separação do conhecimento e controle permite que alterações sejam feitas em uma parte da base de conhecimento, sem criar efeitos colaterais nos outros;

- a separação do conhecimento e os elementos de controle do programa permitem que o mesmo controle e software de interface sejam usados em uma variedade de sistemas (LUGER, 2005).

Para esse autor, o sistema especialista deve acompanhar o caso de dados específicos: os fatos, conclusões e outras informações relevantes para o caso em apreço. Isso inclui os dados apresentados em uma instância do problema, conclusões parciais, as medidas de confiança de conclusões e impasses no processo de busca. Essa informação é separada da base de conhecimentos gerais.

De acordo com Silva (2009), as principais características do motor de inferência disponível em um sistema especialista dizem respeito às seguintes funcionalidades: método de raciocínio, estratégia de busca, resolução de conflito e representação de incerteza.

O método de raciocínio consiste basicamente de dois processos lógicos aplicáveis a regras de produção: encadeamento progressivo ou encadeamento para frente (do inglês, *forward chaining*.) e; encadeamento regressivo ou encadeamento para trás (do inglês, *backward chaining*.). No encadeamento progressivo, também chamado de encadeamento dirigido por dados, a parte esquerda da regra é comparada com a descrição da situação atual, contida na memória de trabalho. As regras que satisfazem a essa

descrição têm sua parte direita executada, o que, em geral, significa a introdução de novos fatos na memória de trabalho.

O mecanismo de inferência que utiliza encadeamento para frente é baseado na busca do sucesso de uma regra através da checagem do Antecedente de uma Regra e depois de seu Conseqüente. A solução é encontrada partindo-se do Antecedente e tentando-se provar o Conseqüente.

A figura 2 e o Algoritmo a seguir mostra a seqüência do Encadeamento para Frente:

1- obtenção dos valores da parte Antecedente de uma Regra;

2- se os valores são verdadeiros então os valores da parte Conseqüente são testados;

3- para verificar o Conseqüente de uma Regra outra Regra é escolhida;

4- a parte Antecedente dessa Regra deve ser testada;

5- os passos de 1 a 4 são repetidos até que o objetivo seja atingido.

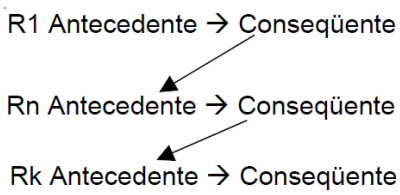

FIGURA 2- ENCADEAMENTO PARA FRENTE

Fonte: Silva (2009)

De acordo com Silva (2009), no encadeamento regressivo, também chamado de encadeamento dirigido por objetivos, o comportamento do sistema é controlado por uma lista de objetivos. Um objetivo pode ser satisfeito diretamente por um elemento da memória de trabalho, ou podem existir regras que permitam inferir algum dos objetivos correntes, isto é, que contenham uma descrição deste objetivo em suas partes direitas. As regras

que satisfazem essa condição têm as instâncias correspondentes às suas partes esquerdas adicionadas à lista de objetivos correntes. Caso uma dessas regras tenha todas as suas condições satisfeitas diretamente pela memória de trabalho, o objetivo em sua parte direita é também adicionado à memória de trabalho.

A figura 3 e o Algoritmo a seguir mostra a seqüência do Encadeamento para Trás:

1. obtenção dos valores da parte Conseqüente de uma Regra;

2. se os valores são verdadeiros então os valores da parte Antecedente são testados;

3. para verificar o Antecedente dessa Regra outra Regra é escolhida;

4. a parte Conseqüente dessa Regra deve ser testada;

5. os passos de 1 a 4 são repetidos até que o objetivo seja atingido.

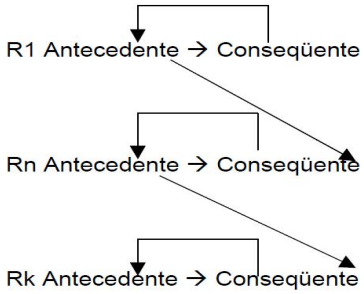

FIGURA 3- ENCADEAMENTO PARA TRÁS

Fonte: Silva (2009)

Conforme Rich & Knight (1994), o módulo para aquisição de conhecimento é o responsável por traduzir o conhecimento conseguido junto a um especialista, em regras. Esse processo deve ser utilizado constantemente de forma a aumentar o refinamento do conhecimento adquirido pelo sistema, o

mais próximo possível do conhecimento do especialista da área, podendo ser automático, semi-automático ou manual.

O módulo de explanação, segundo Giarratano & Riley (1998), expõe ou detalha o raciocínio utilizado pelo sistema para a obtenção do resultado. É essencialmente suscetível em aplicações como a medicina, onde é importante, por exemplo, justificar nitidamente todos os passos utilizados para se chegar a um diagnóstico.

A Interface com o usuário, conforme Nikolopoulos (1997), é o mecanismo pelo qual o usuário e o sistema especialista interagem. Podem ser utilizadas interfaces gráficas GUI (*Graphic User Interface*) e, alternativamente, oferecer conexão com outros sistemas e segundo Nilsson (1998), o uso da linguagem natural também pode ser oferecido para interface com o usuário.

2.2 CLASSIFICAÇÃO DE SISTEMAS ESPECIALISTAS

Os SEs podem ser uma saída quando um problema não puder ser implementado em um algoritmo ou sempre que sua solução ocorra através de um processamento muito demorado, pois os mesmos possuem o seu mecanismo apoiado em processos heurísticos. São utilizados também para preservar e transmitir o conhecimento de um especialista humano em um determinado campo de atuação, não é influenciado por elementos externos, como ocorre com o especialista humano, sendo assim, para as mesmas condições o SE deve fornecer sempre um mesmo conjunto de solução.

Os sistemas especialistas são classificados quanto às características do seu funcionamento. De um modo geral, tais categorias são (SILVA, 2009):

Interpretação: são sistemas que inferem descrições de situações a partir da observação de fatos, fazendo uma análise de dados e procurando determinar as relações e seus significados.

Diagnóstico: são sistemas que detectam falhas oriundas da interpretação de dados. A análise dessas falhas pode conduzir a uma conclusão diferente da simples interpretação de dados. Detectam os problemas mascarados por falhas dos equipamentos e falhas do próprio diagnóstico, que este não detectou por ter falhado. Esses sistemas já têm embutido o sistema de interpretação de dados.

Monitoração: São sistemas que comparam comportamento de sistemas reais com comportamentos esperados. Interpreta as observações de sinais sobre o comportamento monitorado.

Predição: A partir de uma modelagem de dados do passado e do presente, esse sistema permite uma determinada previsão do futuro.

Planejamento: O sistema prepara um programa de iniciativas a serem tomadas para se atingir um determinado objetivo. São estabelecidas etapas e sub-etapas e, em caso de etapas conflitantes, são definidas as prioridades.

Projeto: É um sistema capaz de justificar a alternativa tomada para o projeto final e de fazer uso dessa justificativa para alternativas futuras.

Depuração: Trata-se de sistemas que possuem mecanismos para fornecerem soluções para o mau funcionamento provocado por distorções de dados.

Reparo: Esse sistema desenvolve e executa planos para administrar os reparos verificados na etapa de diagnóstico.

Instrução: São sistemas que avaliam o que os usuários conhecem e suas deficiências e traçam planos para corrigir as deficiências. Por exemplo: ensinar estudantes a rastrear falhas em circuitos elétricos; ensinar estudantes de medicina na área de seleção de anti-micróbios. O sistema de instrução tem um mecanismo para verificar e corrigir o comportamento do aprendizado dos estudantes. Seu funcionamento consiste em ir interagindo com o treinando, em alguns casos apresentando uma pequena explicação e, a

partir daí, ir sugerindo situações para serem analisadas pelo treinando. Dependendo do seu comportamento, vai-se aumentando a complexidade das situações e encaminhando o assunto, de maneira didática, até o nível intelectual do treinamento.

2.3 FASE DO DESENVOLVIMENTO DE SE

De acordo com Luger (2005), na maioria da programação em IA, a construção de sistemas especialistas requer um ciclo de desenvolvimento não-tradicional com base no protótipo inicial e revisão periódica do código. Geralmente, os trabalhos sobre o sistema começam com o engenheiro de conhecimento tentando ganhar alguma familiaridade com o domínio do problema. Isso ajuda na comunicação com o especialista do domínio. Isso é feito em entrevistas iniciais com o perito e pela observação de peritos durante a realização do seu trabalho. Em seguida, o engenheiro de conhecimento e o especialista iniciam o processo de extração do problema do perito de resolução de conhecimento. Isso é feito freqüentemente, dando ao especialista do domínio uma série de problemas e exemplos e como ele explica as técnicas utilizadas na sua solução. Vídeo e/ou fitas de áudio são muitas vezes essenciais nesse processo.

O especialista de domínio é geralmente alguém que tenha trabalhado na área de domínio e compreende as suas soluções e técnicas, como atalhos manipulação de dados imprecisos, avalia soluções parciais, e todas as outras habilidades que marcam uma pessoa como um solucionador de problemas de peritos. O especialista de domínio é o principal responsável para expor essas habilidades para o engenheiro de conhecimento (LUGER, 2005).

Pode-se verificar em Savaris (2002) que o desenvolvimento de um sistema especialista é constituído de 06 (seis) fases:

Identificação: São identificados os participantes do projeto, os recursos envolvidos, as características do problema e os objetivos a atingir.

Conceituação: A fase de conceituação consiste em definir a base conceitual do sistema especialista. O engenheiro de conhecimento e o especialista decidirão quais os recursos básicos necessários para desenvolver o problema (conceitos, relações, mecanismos de controle) e estabelecerão também o grau de refinamento que será usado na representação do conhecimento.

Formalização: Envolve a expressão de conceitos e de relações-chaves, de uma maneira formal, identificando estruturas de suporte para sua representação e armazenamento. Se essas estruturas forem parte integrante de alguma ferramenta existente, poder-se-á utilizá-la para construção do sistema.

Implementação: Essa fase se consume com a edição do conhecimento e a feitura dos programas que o processam, quando não for feita opção por alguma ferramenta já existente.

Teste e Avaliação: O sistema especialista deverá ser testado e avaliado freqüentemente, desde a implementação inicial do sistema. Deverão ser levados em consideração o desempenho e a utilidade.

Revisão: Consiste em revisar o sistema, especialmente para alterar e melhorar aspectos observados na fase de avaliação.

2.4 BASE DE CONHECIMENTO

Segundo Feigenbaum & McCorduck (1983), *"A potência de um sistema especialista deriva do conhecimento que ele possui e não dos formalismos e esquemas específicos que ele emprega"*.

De acordo com os autores Rezende, Pugliesi e Varejão (2003), a base de conhecimento contém a descrição do conhecimento necessário para a resolução do problema abordado na aplicação. Isso inclui argumentos sobre o domínio de conhecimento, regras que descrevem relações nesse domínio e, em muitos casos, heurísticas e métodos de resolução de problemas.

A base de conhecimento não pode ser considerada como uma simples coleção de informações.

Segundo Nilson (1982), a tradicional base de dados com dados, arquivos, registros e seus relacionamentos estáticos são substituídos por uma base de regras e fatos e também heurísticas que correspondem ao conhecimento do especialista, ou dos especialistas do domínio sobre o qual foi construído o sistema e interage com o usuário e com o motor de inferência, permitindo identificar o problema a ser resolvido, as possibilidades de solução e o processo de raciocínio e inferência que levam a conclusões sobre o problema submetido ao sistema adquirindo informações necessárias para a resolução do problema.

Devido à utilização de heurísticas, o usuário é requerido pelo sistema para prestar informações adicionais e, a cada pergunta respondida pelo usuário ou a cada nova informação, reduz-se o espaço de busca a ser percorrido pelo sistema, encurtando-se o caminho entre o problema e sua solução.

Informalmente, uma base de conhecimento é um conjunto de representações de ações e acontecimentos do mundo. Cada representação é chamada de sentença. De acordo com Russel e Norvig (2003), as sentenças são expressas em uma linguagem específica, chamada de representação do conhecimento, que se baseiam em diferentes técnicas de representação, tais como: regras de produção, redes semânticas, frames e lógicas. Também podem usar uma combinação de diferentes técnicas de representação de conhecimento, conhecidos como sistemas híbridos com relação à representação do conhecimento.

Segundo os autores Heisserman, Callahan e Mattikali (2000), bases de conhecimentos podem ser compostas por até dezenas de milhares de sentenças. Estas sentenças apresentam variados graus de generalidade podendo ser desde totalmente específicas do domínio até completamente gerais. As sentenças em sua maioria descrevem relações de causa e efeito no domínio, por exemplo: “se a temperatura do paciente está acima de 37,5 Cº,

então o paciente tem febre". Outras sentenças expõem conhecimento sobre como guiar a busca por uma solução. Este tipo de conhecimento é conhecido como metaconhecimento, ou seja, conhecimento sobre o conhecimento. UM exemplo de metaconhecimento específico do domínio: "se o paciente é alcoólatra, investigue primeiramente doenças hepáticas".

Um sistema especialista deve possuir uma base de conhecimento, formada de fatos, regras e heurísticas sobre domínio, assim como um especialista humano faz, deve ser capaz também de dar sugestões e conselhos aos usuários, como também adquirir novos conhecimentos com essa interatividade. A base de conhecimento é a coleção de informações, representado na forma de regras SE-ENTÃO, as quais supostamente devem agir conforme um especialista humano. Seria, então, a "alma" do sistema especialista.

Os fatos constituem um conjunto de informação que é largamente compartilhado, publicamente disponível e geralmente aceito pelos especialistas em um campo. As heurísticas são regras pouco discutidas, de bom discernimento (regras de raciocínio plausível, regras de boa conjectura), que caracterizam a tomada de decisão em nível de especialista na área (BARRETO, 2001).

Segundo Harmon e King (1988) as heurísticas são regras práticas que "podam" os espaços de busca para dimensões controláveis. Tendem a concentrar a atenção em algumas configurações-chave. Os métodos heurísticos procuram um grau tão grande quanto possível de uma ação a uma situação. Assim, ela engloba estratégias, procedimentos, métodos de aproximação tentativa/erro, sempre à procura da melhor forma de chegar a um determinado fim. Os processos heurísticos exigem muitas vezes menos tempo que os processos algorítmicos, aproximam-se mais da forma como o ser humano raciocina e chega às resoluções dos problemas e garantem soluções eficientes.

2.5 AQUISIÇÃO DO CONHECIMENTO

Aquisição do conhecimento é definida por Buchanan, Barstow e Bechtel (1983) como a transferência e transformação do conhecimento especializado com potencial para a resolução de problemas de alguma fonte de conhecimento para um programa. É uma informação muito genérica, pois denota neutralidade em relação à forma como se obtém o conhecimento, no entanto, ela enfatiza o importante papel do observador, no caso o engenheiro do conhecimento, no processo, tirando assim a imagem de um extrator do conhecimento da cabeça do especialista para a base do sistema.

Rezende *et al.* (2003) fala em uma definição de aquisição do conhecimento mais atual como um processo de modelagem (criação de uma teoria) de problemas e soluções pertinentes a tarefas em um domínio específico. Conhecimento sobre o domínio e sobre o problema, assim como sobre as estratégias de resolução, formam o material observado e interpretado pelo engenheiro do conhecimento para a criação do modelo computacional.

Com a intenção de tornar mais efetivo o processo de aquisição do conhecimento, algumas técnicas têm sido desenvolvidas. Essas técnicas podem ser classificadas como manuais, semi-automáticas e automáticas.

De acordo com Rezende *et al.* (2003), as técnicas manuais representam diretrizes do processo conduzido por engenheiros do conhecimento, podem ser classificadas de acordo com a forma de obtenção do conhecimento em: baseadas em descrições, entrevistas, em acompanhamentos ou em modelos. Nessa técnica, o engenheiro do conhecimento é responsável por adquirir o conhecimento do especialista ou de outras fontes e depois codificá-lo em uma base de conhecimento. As técnicas manuais contribuem na redução da subjetividade e no tempo de execução do processo de aquisição do conhecimento.

As técnicas semi-automáticas objetivam proporcionar aos especialistas ferramentas que possibilitem a criação dos sistemas minimizando a necessidade de um engenheiro de conhecimento; são baseadas em teorias

cognitivas ou em modelos existentes. A utilização dessa técnica reduz o número de agentes humanos envolvidos e, por conseqüência, os seus problemas de comunicação, tornando mais imediato e fácil para o engenheiro ou especialista o processo de obter respostas a respeito do comportamento do sistema e identificar possíveis inadequações da base.

Já o objetivo das técnicas automáticas é minimizar a participação humana minerando conhecimento de extensas fontes de dados ou implantando mecanismos de inferência que permitam aprendizado automático de máquina, tentando induzir regras a partir de exemplos catalogados.

2.5.1 PROCESSO DE AQUISIÇÃO DO CONHECIMENTO

De acordo com Rezende *et al.* (2003), mesmo com algumas regularidades no processo de aquisição do conhecimento, ainda não foi possível estabelecer um consenso sobre um método eficaz. Possivelmente, isso pode indicar que essa área ainda não se encontra em uma fase completamente madura, ou pode ser também resultado da própria natureza situada, multidisciplinar e multiparadigmática dessa área na qual cada problema demanda um processo específico e diferenciado.

O processo do conhecimento é freqüentemente dividido em fases. Os autores Buchanan, Barstow e Bechtel (1983) mencionam um modelo composto por cinco fases: identificação, conceituação, formalização, implementação e teste.

Rezende *et al.* (2003), citam que a fase de identificação é semelhante à fase de análise de requisitos em Engenharia de Software, em que o projetista procura: 1- elementos do domínio que identifique a classe do problema que o SE deverá resolver; 2- os dados sobre os quais o sistema deverá operar; 3- os critérios para classificar as soluções nos contextos de funcionamento do sistema e 4- a maneira como o problema deve ser resolvido.

Na fase de identificação, o engenheiro do conhecimento deve fazer um levantamento bibliográfico sobre o domínio, entrevistar os

especialistas na área para ter uma idéia geral do domínio, da tarefa e do tipo de interação com os dados necessários para resolver problemas do domínio em questão.

Na fase de conceituação são apontados, pelo especialista, os conceitos chaves e as relações entre eles, bem como as características necessárias para descrever o processo de resolução do problema.

Segundo Davis, Shrobe e Szolovits (1993), como em todo processo de modelagem, ao representar o mundo, decide-se o que representar, e também o que não representar. Selecionar esses conceitos significa ao mesmo tempo tomar decisões sobre como e o que ver no mundo. Esse processo de abstração implica em simplificações da realidade tornando-a computacionalmente tratável, produzindo um conjunto de conceitos que se assume existir em determinada área do conhecimento, bem como as relações que poderão existir entre eles.

Na fase de formalização, o engenheiro do conhecimento concentra-se no processo de modelagem computacional do problema, preocupando-se com a natureza do espaço de busca e como a busca nesse espaço deve ser conduzida. Ele deve escolher o formalismo que melhor se adapte para representar o problema/solução, como lógica, frames, redes semânticas e regras de produção.

Durante a fase de implementação, o engenheiro do conhecimento orienta a codificação do modelo de SE, desenvolvido na fase de conceituação, em alguma linguagem de programação. Após a seleção da linguagem, o engenheiro do conhecimento deve fornecer a formalização e a indicação da linguagem para o(s) programador(es), esse(s) terá(ão) a responsabilidade de projetar e codificar a base de conhecimento, levando em conta os aspectos computacionais e contornar eventuais problemas de complexidade de estrutura de dados e algoritmos utilizados.

Após a construção do SE, é necessário verificar se ele atende ao propósito para o qual foi desenvolvido. Nessa fase de teste, o engenheiro do

conhecimento deverá se encarregar de avaliar, juntamente com os especialistas, o desempenho do SE.

2.5.2 TÉCNICAS DE AQUISIÇÃO DO CONHECIMENTO

De acordo com Barreto (2001), a maioria das etapas do ciclo de vida de um SE pode ser considerada como ciclo de vida de um programa qualquer, no entanto a aquisição do conhecimento do especialista pelo engenheiro do conhecimento envolve características peculiares. Cabem aos engenheiros de conhecimento e especialistas do domínio selecionarem as mais adequadas para cada tipo de problema. É necessário, portanto, que eles estejam preparados e sejam versáteis para adaptar as soluções de acordo com a situação, obtendo o máximo de conhecimento possível.

Waterman (1986) cita que as fontes potenciais de conhecimento são os especialistas humanos (principal fonte), livros-texto, bancos de dados, documentos com relatos de experiências e estudos, a experiência pessoal do engenheiro do conhecimento. Quando a fonte de conhecimento é uma pessoa, a atividade é essencialmente um empreendimento social que requer cooperação entre o elicitante (geralmente o engenheiro de conhecimento) e o provedor de conhecimento. Ambas as partes interpretam a situação como sua progressão e ajustam suas respostas para tornarem-se apropriadas.

As principais técnicas citadas por Barreto (2001) são:

- Observação;

- Entrevista com o especialista;

- Análise de Discurso.

2.5.2.1 OBSERVAÇÃO

Nesta técnica o especialista é observado durante seu trabalho, fornecendo uma visão realista de como o especialista trabalha. É considerado

o primeiro passo na construção de uma base de conhecimentos por permitir ao engenheiro do conhecimento se familiarizar com o problema; pode ser direta ou por meio de vídeo.

2.5.2.2 ENTREVISTA

Essa técnica deve ser utilizada quando o engenheiro do conhecimento tiver alguma familiaridade com o assunto; para que possa manter uma conversação com o especialista, e que para isso, é necessário realizar um estudo prévio sobre o domínio.

Em conformidade com Barreto (2001), uma entrevista deve ser planejada e incluir perguntas diretas e indiretas. As perguntas diretas são feitas quando o engenheiro do conhecimento tem uma idéia clara de qual conhecimento é necessário para completar a base de conhecimentos, com respostas explícitas;, já na pergunta indireta o especialista é deixado livre para sugerir novos tipos de conhecimento, sendo aconselhável quando o engenheiro do conhecimento não tem idéia precisa do domínio.

Segundo Rezende *et al*. (2003), a entrevista pode ser não-estruturada e estruturada. A entrevista não-estruturada deve ser feita na fase de identificação, na qual o escopo e o foco da aplicação são determinados; ela poderá ser conduzida informalmente.

Embora a entrevista não tenha sido preparada e estruturada com antecedência, ela pode ser conduzida passo a passo pelo engenheiro, de modo que ele vá aprendendo o que está sendo mencionado.

Já a entrevista estruturada, de acordo com Rezende, é uma abordagem correlata, embora significativamente mais produtiva, referindo-se à identificação dos elementos e relações do domínio, essas podem ser feitas na fase de identificação e conceituação, onde é formulada a descrição do domínio. Essa abordagem é baseada em um processo sistemático orientado a um objetivo que leva a uma comunicação organizada entre o especialista e o engenheiro, ajudando a evitar distorções decorrentes da subjetividade.

2.5.2.3 ANÁLISE DE DISCURSO

Análise de discurso (AD) é uma prática e um campo da lingüística e da comunicação especializado em analisar construções ideológicas presentes em um texto. A análise de discurso compreende a natureza social do discurso, isto é, compreende a historicidade do texto.

É muito utilizada, por exemplo, para analisar textos da mídia e as ideologias que trazem em si; é proposta a partir da filosofia materialista que põe em questão a prática das ciências humanas e a divisão do trabalho intelectual, de forma reflexiva.

Segundo Ferreira (2006) a AD é uma disciplina de conhecimento sobre a linguagem que permite alterar, modificar a experiência e, eventualmente, a ação e o comportamento das pessoas. E isso a faz, em muitos casos, uma disciplina nitidamente de intervenção no meio social, político e histórico.

De acordo com Barreto (2001), a análise de discurso consiste essencialmente em gravar a entrevista com o especialista para depois analisar a conversação e deve ser utilizada como auxílio para as outras técnicas, pois em uma conversação é difícil lembrar tudo que foi dito.

A AD no Brasil, segundo Ferreira (2006), trabalha hoje com materialidades discursivas das mais diversas, que vão desde os discursos institucionalizados até aqueles do cotidiano, podendo com isso abarcar o discurso religioso, indígena, dos movimentos sociais, midiático, pedagógico, etc. Não se detém exclusivamente na linguagem verbal (nas questões da escrita e da oralidade). A imagem, de modo geral, os cartazes, fotografias, charges, pichações e grafites ganham cada vez mais espaço entre os analistas de discurso. Ferreira menciona também o surgimento de novas linguagens que começam a aparecer como objeto de investigação mais recente, relacionadas ao computador e à Internet e que nos forçam a rever noções até então clássicas na teoria como autoria, efeito-sujeito, memória, hiperlíngua, etc. Em todas essas distintas materialidades o acesso se faz pelo fragmento, pelo

resíduo, pelo que sobra e pelo que falta, pelo que escapa ao simbólico, pelo que toca o real da língua e o real do sujeito. Enfim, há uma gama imensa de possibilidades, que atestam a potencialidade e o vigor do aparato téorico-analítico do campo do discurso.

Segundo Reisserman e Mumby *apud* Alvarez (2002), a análise de discurso crítica é baseada em texto e dados associados, o falar que emprega métodos sociolingüísticos

Kuipers e Kassier (1987) apresentam um exemplo em que a transcrição é quebrada em pequenas linhas que correspondem aproximadamente a uma frase na explicação. O formato resultante facilita o ônus da análise posterior.

Fora da transcrição como um todo, são feitas seleções de trechos em que o assunto parece concentrar-se na explicação e apresentar seu conhecimento médico, em vez de expressar uma opinião sobre seus próprios processos mentais.

Em Kuipers e Kassier (1987), a análise de um trecho realiza-se em duas fases:

1 – identificar os objetos e as relações no domínio a que o assunto se refere, como da formulação utilizada para referir-se a eles;

2 – identificar as relações causais que são descritas no segmento.

Neste trabalho, foram aplicadas as seguintes técnicas de aquisição de conhecimento: entrevista com especialista em mineração de dados e a análise de discurso no estudo dos guias da ferramenta Kira (MENDES, 2009), conforme se pode observar nos apêndices A e B deste documento.

2.6 REPRESENTAÇÃO DO CONHECIMENTO

2.6.1 CONCEITOS

De acordo com Liebowitz (1999), a representação do conhecimento (RC) é uma das principais preocupações dos sistemas especialistas e da inteligência artificial.

Os autores Davis, Shrobe e Szolovits *apud* Rezende (2003) definem representação do conhecimento como algo que substitui o objeto ou fenômeno real, de modo a permitir a uma entidade determinar as conseqüências de um ato pelo pensamento ao invés de sua realização. Uma RC pode ser entendida como uma forma sistemática de estruturar e codificar o que se sabe sobre uma determinada aplicação.

Conforme estes autores, ao contrário de uma codificação qualquer, uma RC deve apresentar algumas características, tais como:

- ser compreensível ao ser humano, pois caso seja necessário avaliar o estado de conhecimento do sistema, a RC deve permitir a sua interpretação;

- abstrair-se dos detalhes de como funciona internamente o processador de conhecimento que a interpretará;

- ser robusta, ou seja, permitir sua utilização mesmo que não aborde todas as situações possíveis;

- ser generalizável, ao contrário do conhecimento em si que é individual. Uma representação necessita de vários pontos de vista do mesmo conhecimento, de modo que possa ser atribuída a diversas situações e interpretações.

Para os autores Davis, Shrobe e Szolovits *apud* Rezende (2003), existem várias técnicas de RC e para avaliar as mesmas existem

alguns critérios dos quais os principais são adequação lógica e conveniência notacional, sendo que:

- Adequação lógica observa se o formalismo usado é capaz de expressar o conhecimento que se deseja representar;

- Conveniência notacional verifica as convenções da linguagem de representação. Se essas forem muito complicadas, a tarefa de codificação torna-se extremamente complexa.

A representação do conhecimento de acordo com Ararobóia *apud* Rezende (2003) é um dos problemas cruciais de IA, pois não existe uma teoria geral de RC, no entanto muitas técnicas de RC têm sido estudadas pelos pesquisadores de IA.

A seguir são apresentadas algumas técnicas para representação do conhecimento: Representação Lógica, Rede Semântica, Regras de Produção e Frames.

2.6.2 REPRESENTAÇÃO LÓGICA

Segundo Kowalski & Hogger *apud* Rezende (2003), desde o final dos anos 70 tem havido grande interesse no uso de métodos que derivam da lógica matemática, na pesquisa em Inteligência Artificial, em contraste ao uso de métodos mais intuitivos e heurísticos.

A lógica matemática é uma linguagem formal. Diferente de linguagens naturais, nas quais as regras gramaticais são imprecisas, nas linguagens formais sempre se pode dizer se uma seqüência de símbolo está de acordo com as regras para a construção de expressões da linguagem. A lógica matemática possui várias regras sintáticas de dedução, ou seja, formas de realizar inferências dedutivas exclusivamente a partir de formato sintático das expressões da linguagem, sem se basear em quaisquer idéias extras ou intuitivas. Dedução automática refere-se ao comportamento de qualquer programa de computador que realiza inferências dedutivas a partir das leis da

lógica matemática. A figura 4 mostra um exemplo de representação do conhecimento por meio de lógica.

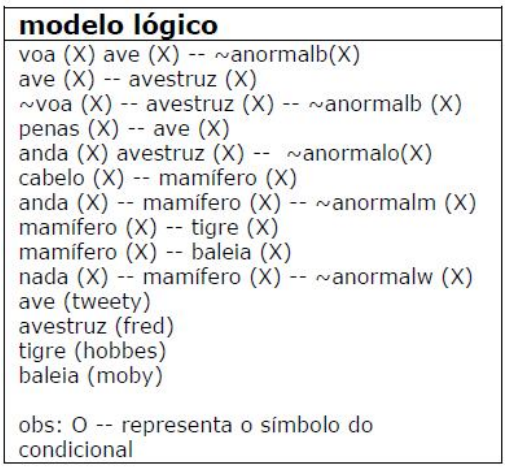

FIGURA 4- REPRESENTAÇÃO LÓGICA

Fonte: Rover, 2009

2.6.3 REDES SEMÂNTICAS

Rede semântica é uma forma de representação de conhecimento muito usada nas pesquisas de inteligência artificial relacionadas com processamento de linguagem natural (entendimento da língua portuguesa falada e escrita).

De acordo com Passos (1989), a representação do conhecimento, através de redes semânticas é uma tentativa de simular o modelo psicológico de memória associativa humana, modelando o conhecimento como um conjunto de pontos chamados nós ou nodos, conectados por ligações chamados arcos que descrevem as relações entre os nós.

Para Rezende *et al* (2003), uma rede semântica é um grafo rotulado e direcionado formado por um conjunto de nós representando os objetos (indivíduos, coisas, eventos, conceitos, situações em domínio) e por um conjunto de arcos representando as relações entre os objetos. Um arco é rotulado com o nome da relação que ele representa. Vários arcos podem ter o

mesmo rótulo, entretanto cada objeto é representado por apenas um nó. Os arcos em geral dependem da espécie de conhecimento que está sendo representado; por exemplo, **é-um:** as relações entre os objetos estão em uma taxonomia hierárquica; **é-parte:** as relações entre os objetos obedecem a um tipo de composição, ou seja, um objeto é componente de outro, não havendo nenhum tipo de herança.

Segundo Luger (2005), uma rede semântica representa o conhecimento como um grafo, com os nós correspondentes a fatos ou conceitos e os arcos as relações e associações entre conceitos. Ambos os nós e as ligações são geralmente rotulados, por exemplo, uma rede semântica que define as propriedades de neve e gelo. Essa rede poderia ser utilizada (com as regras de inferência adequadas) para responder a uma série de perguntas sobre neve, gelo e bonecos de neve. Essas inferências são feitas seguindo os links de conceitos relacionados.

O termo "rede semântica" engloba uma família de grafo baseado em representações. Esses diferem principalmente nos nomes que são permitidos para nós e ligações e as deduções que podem ser realizadas, no entanto um conjunto comum de pressupostos e preocupações é compartilhado por uma rede de linguagens de representação, que são ilustrados por uma discussão sobre a história da representação da rede.

Grande parte da pesquisa em representações da rede tem sido feita no campo da compreensão da linguagem natural. No caso geral, a compreensão da linguagem exige a compreensão do senso comum, as formas pelas quais os objetos físicos se comportam, as interações que ocorrem entre seres humanos, e as maneiras pelas quais as instituições humanas são organizadas. Um programa em linguagem natural deve compreender as intenções, crenças, do raciocínio hipotético, planos e objetivos. Devido a essas exigências, a compreensão da linguagem sempre foi uma força motriz para a investigação em representação do conhecimento (LUGER, 2005).

Uma característica chave da representação de rede semântica é que relevantes associações podem ser feitas explicitamente. Fatos

importantes sobre um objeto ou conceito podem ser deduzidos dos nós aos quais eles estão ligados diretamente, sem uma pesquisa no contexto.

Sowa (2002) define uma rede semântica como sendo uma notação gráfica composta por nodos interconectados, podendo ser utilizadas para representação de conhecimento, ou como ferramenta de suporte para sistemas automatizados de inferências sobre o conhecimento.

Uma das propriedades mais importantes dessas relações é a transitividade, pois permite uma declaração concisa de propriedades nos objetos mais gerais. Mecanismos de inferência podem então ser utilizados para derivar essas propriedades para os objetos mais específicos, denominado de Herança de Propriedades (REZENDE *et al*, 2003).

Para esses autores, uma das razões, senão a principal, das redes semânticas serem bem aceitas e atrativas na comunidade de RC é a possibilidade de visualização gráfica das estruturas de conhecimento e suas relações, conforme observa-se na figura a seguir.

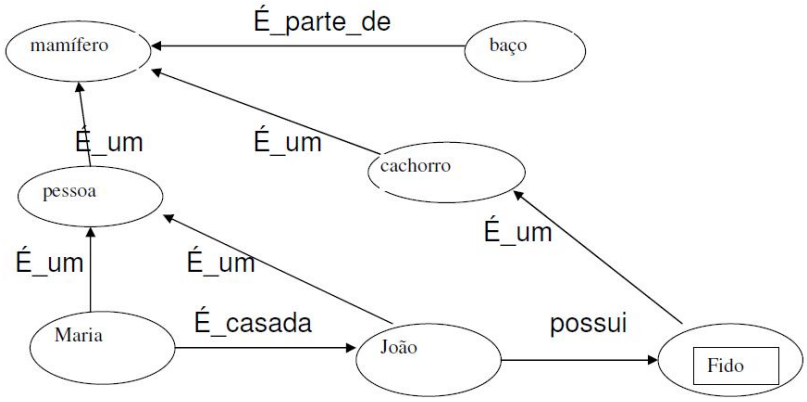

FIGURA 5- REDE SEMÂNTICA

Fonte: Silva (2009).

2.6.4 REGRAS DE PRODUÇÃO

Regras de produção são pares de expressões consistindo em uma condição e uma ação. Uma vez que as condições da parte antecedente da regra são verdadeiras a ação da parte conseqüente é executada.

As Regras de Produção constituem uma forma de representação do conhecimento utilizado em sistemas baseados em conhecimentos. Segundo Rezende *et al* (2003), esses sistemas se inspiraram na idéia que o processo de tomada de decisão humano poderia ser modelado por meio de regras do tipo "*SE condições ENTÃO conclusões e ações*", portanto as regras podem expressar relacionamentos lógicos e equivalências de definições para simular o raciocínio humano.

Em Barreto (2001), é colocado que a representação do conhecimento usando regras de produção usa regras de forma semelhante às regras da representação de conhecimento em lógica, entretanto essas regras são usadas de modo diferente. Com efeito, uma regra de produção indica uma ação a ser feita para poder resolver um problema. Ela age em um estado da solução do problema e procura modificar esse estado de modo a se aproximar da solução desejada.

As regras de produção são regras no formato SE - ENTÃO, permitindo-se o uso dos conectivos lógicos (E, OU, NÃO, e outros desejados), além do tratamento de incertezas.

As regras de produção são formadas de duas partes: a primeira, chamada de antecedente, ou premissa, ou condição, ou parte *IF* (SE), a segunda, chamada de conseqüente, ou conclusão, ou ação, ou parte *THEN* (ENTÃO).

Conforme Passos (1989), representar o conhecimento, por regras de produção ou simplesmente regras, é uma maneira bastante utilizada nos diversos sistemas especialistas existentes no mercado mundial. Nesse esquema, os conhecimentos são representados através de pares condição – ação.

As regras são estruturas do tipo:

Se <condição> então <ação>, sendo que:

- <condição> estabelece um teste, cujo resultado depende do estado atual da base de conhecimento. Tipicamente o teste verifica a presença ou não de certas informações na base.

- <ação> altera o estado atual da base de conhecimento, adicionando, modificando ou removendo unidades de conhecimento presentes na base. Uma ação pode acarretar também efeitos externos à base como a escrita de uma mensagem no vídeo, por exemplo.

Segundo Waterman (1986), as regras de produção são apropriadas para representar conhecimentos oriundos de recomendações, diretrizes, estratégias e quando o domínio do conhecimento é resultante de proposições empíricas que foram desenvolvidas ao longo do tempo através da experiência de especialistas na resolução de problemas.

Neste trabalho, optou-se por utilizar, como forma de representação do conhecimento, redes semânticas e regras de produção, uma vez que são intuitivas aos usuários. A rede semântica procura mostrar, através da linguagem natural e representação visual, o quanto sua descrição se aproxima da realidade, simplificando a forma de representação do problema, em que o ser humano consegue interpretar sem muitas dificuldades.

Nas regras de produção, cada regra representa um pedaço do conhecimento independente. A representação do conhecimento é de forma modular, uniforme e natural (WALTERMAN,1986; REZENDE *et al*, 2003). Devido à sua modularidade, a manutenção é relativamente simples. As regras de produção são fáceis de compreender e de modificar, além de que, novas regras podem ser facilmente inseridas na base de conhecimento.

A figura 6 apresenta um exemplo de regras de produção.

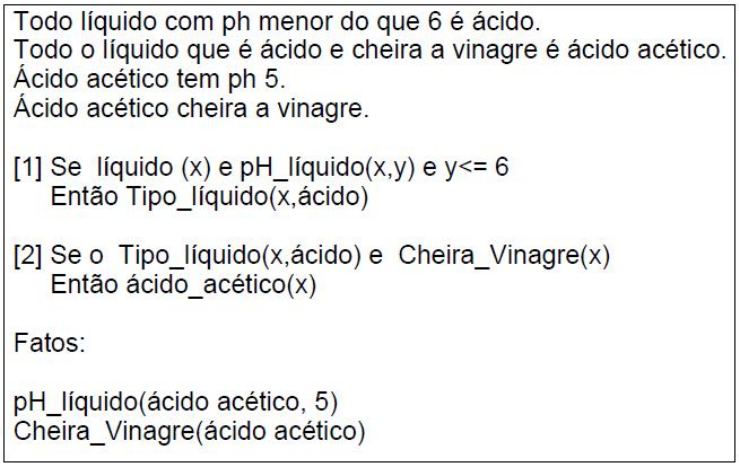

FIGURA 6 - EXEMPLO DE REGRAS DE PRODUÇÃO

Fonte: Silva (2009)

2.6.5 FRAMES

O modelo de frames para a representação do conhecimento foi introduzido inicialmente em 1975 por Marvin Minsky. Conforme Minsky (1985), Frame é um termo usado para designar um agrupamento de conhecimentos relevantes a uma coisa, um indivíduo, uma situação ou um conceito. O frame possui um nome que identifica o conceito por ele definido e consiste de um conjunto de atributos, chamados *slots* (RICH & KNIGHT, 1993).

Segundo Minsky *apud* Luger (2002), um frame pode ser visto como uma estrutura de dados estática usada para representar bem situações estereotipadas. Estruturas *framelike* parecem organizar o nosso próprio conhecimento do mundo e ajustar a cada nova situação, chamando a informação estruturada por experiências passadas.

De acordo com Luger (2005), essa representação apóia a organização de conhecimento em unidades mais complexas que refletem a organização dos objetos no domínio.

Os frames integram conhecimento declarativo sobre objetos e eventos e conhecimento procedimental sobre como recuperar informações ou

calcular valores. Os atributos também apresentam propriedades, que dizem respeito ao tipo de valores e às restrições de número que podem ser associadas a cada atributo. Essas propriedades são chamadas facetas. As facetas contêm informações que descrevem os *slots*. Essas informações definem explicitamente valores que o *slot* pode assumir, ou podem indicar a maneira de calcular ou deduzir o seu valor (procedimentos). Exemplo de facetas: tipo, domínio, valor default, etc. Os valores dos *slots* podem ser explicitamente definidos ou implicitamente herdados de um dos seus ancestrais (MINSKY, 1985). Umas das principais características desse modelo de representação é a Herança de Propriedades, na qual uma classe mais especializada pode herdar todas as propriedades da classe mais geral.

Assim como nas redes semânticas, uma das características nos frames é a possibilidade de que sejam criados novos subtipos de objetos que herdem todas as propriedades da classe original. Essa herança é bastante usada tanto para a representação do conhecimento como para a utilização de mecanismos de inferência.

Sistemas de herança nos permitem armazenar informações do mais alto nível de abstração, o que reduz o tamanho das bases de conhecimento e ajuda a evitar inconsistências e atualização, obrigando-nos a definir os traços essenciais apenas uma vez, ao invés de exigir a sua afirmação para cada indivíduo. Herança também nos ajuda a manter a consistência da base de conhecimento ao adicionar novas classes e indivíduos. (LUGER, 2005).

Segundo Minsky (1985), as associações entre frames determinam a sua estrutura hierárquica. Cada associação liga um frame-pai ao seu filho. O frame-filho pode ser entendido como uma especialização do frame-pai, ou o frame-pai como uma generalização do frame-filho. Um frame-filho pode herdar valores (default ou correntes) de qualquer um dos seus frames-pais, que por sua vez herdam de seus pais, e assim por diante, permitindo a distribuição da informação sem duplicação.

A hierarquia de frames permite que dados sejam armazenados de maneira abstrata e aninhada com propriedades comuns que são automaticamente herdadas através da hierarquia. Isso evita a duplicação desnecessária de informações, simplifica o código e proporciona um sistema de fácil leitura e manutenção (MINSKY, 1985). Um exemplo de frame pode ser observado na figura 7.

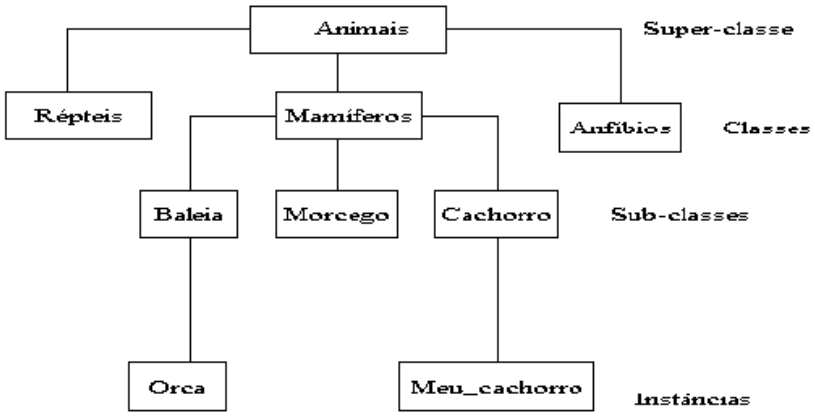

FIGURA 7- EXEMPLO REPRESENTAÇÃO DO CONHECIMENTO ATRAVÉS DE FRAME

Fonte: Rover, 2009.

3 O DOMINIO DA TAREFA DE CLASSIFICAÇÃO

Com a finalidade de construir uma base conhecimento com um enfoque instrucional na escolha da tarefa de classificação em mineração de dados, o problema do domínio foi estudado de acordo com as características que poderiam conduzir o usuário a escolher ou a rejeitar a tarefa de classificação, dependendo do tipo de problema apresentado por ele.

Sendo assim, foram estudos conceitos tais como: Metodologia CRISP-DM, mineração de dados, classificação, etc.

3.1 METODOLOGIA CRISP-DM

Com o avanço da tecnologia e o crescente interesse do mercado pela mineração de dados, houve a necessidade de se criar uma nova metodologia padrão para ajudar no processo de mineração de dados.

Diante dessa necessidade, as empresas NCR, Daimler-Chrysler AH, SPSS Inc. e OHRA se reuniram nos meados da década de 90 e criaram um processo padrão de mineração de dados denominado *CRISP-DM - Cross-Industry Standard Process for Data Mining*.

Segundo Chapman *et. al.* (2000) o CRISP-DM é uma metodologia padrão não-proprietário e de livre distribuição, amplamente utilizado por membros de indústrias para definir, desenvolver e implementar um projeto de mineração de dados.

3.1.1 CONCEITOS

Segundo Olson e Dellen (2008), CRISP-DM pode ser definida como Metodologia padrão não proprietária que identifica as diferentes fases na implantação de um projeto de mineração de dados, permitindo rapidez na realização do processo, melhor controle em nível gerencial e confiabilidade amplamente utilizada por membros da indústria. Em Chapman *et. al.* (2000), a metodologia CRISP-DM é apresentada em termos de um modelo hierárquico

de processo, que consiste em conjuntos de tarefas em quatro níveis de abstração (do geral para específico): fases, tarefas genéricas, tarefas especializadas e instâncias de Processos.

Observando-se a figura 8, nota-se que no nível superior, o processo de mineração de dados é organizado em uma série de fases, cada fase constitui o segundo nível em várias tarefas genéricas. O segundo nível é chamado de Genérico, isso porque se destina a ser geral o suficiente para cobrir todas as possíveis ocorrências da mineração de dados. O terceiro nível é chamado de Tarefas Específicas, descreve como as ações das tarefas genéricas devem ser realizadas para determinadas situações especificas. No quarto nível, tem-se a Instância do Processo que é um registro das ações, decisões e os resultados de uma mineração de dados.

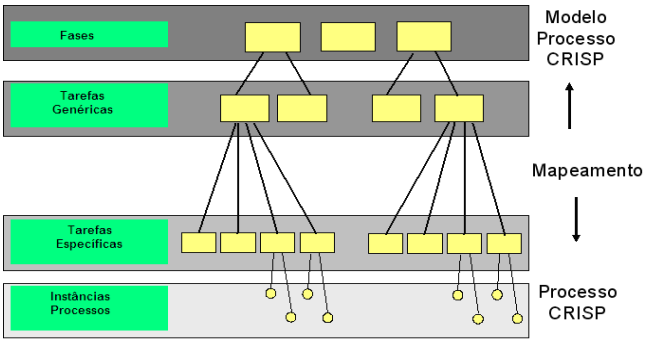

FIGURA 8: QUATRO NÍVEIS DE METODOLOGIA
Adaptado de Chapman *et al*. (2000)

O atual modelo de processo de extração de dados fornece uma visão geral do ciclo de vida de um projeto de mineração de dados. Ele contém as fases correspondentes de um projeto; seus respectivos trabalhos e as relações entre estas tarefas.

O ciclo de vida de um projeto de mineração de dados, conforme Chapman *et al*. (2000), consiste em seis fases. O resultado de cada fase determina qual fase, ou tarefa específica de uma fase, é a próxima a ser realizada. Ela depende do resultado de cada fase, ou tarefa específica de uma fase, que tem de ser realizada. As setas ao lado direito da figura 8 indicam as

mais importantes e freqüentes dependências entre as fases. Segundo Olson & Delen (2008), os analistas experientes podem não precisar aplicar cada fase para cada estudo; CRISP-DM fornece uma estrutura útil para a mineração de dados.

O círculo exterior, na figura 9, simboliza a natureza cíclica da mineração de dados em si, ou seja, um processo de mineração de dados continua após uma solução ter sido implantada. As lições aprendidas durante o processo pode desencadear novas lições, muitas vezes, focalizado mais em questões empresariais. Processos de mineração de dados posteriormente irão se beneficiar das experiências anteriores.

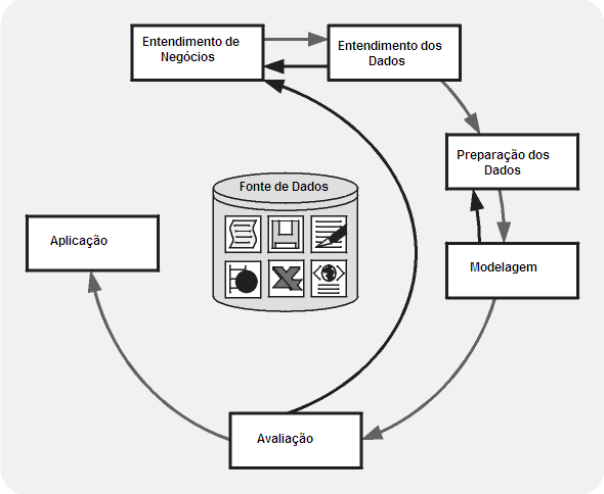

FIGURA 9: PROCESSO DE CRISP-DM
Adaptado de Olson & Delen (2008)

A seguir uma súmula das fases do modelo CRISP–DM (Chapman *et al.*, 2000; Olson & Delen, 2008):

Entendimento de Negócios: A fase inicial do processo visa o entendimento dos objetivos do projeto e dos requisitos sob o ponto de vista do negócio. Baseado no conhecimento adquirido, o problema de mineração de dados é definido e um plano preliminar é projetado para ativar os objetivos.

Entendimento dos Dados: Inicia com uma coleção de dados e procede com atividades que visam: buscar familiaridade com os dados;

identificar problemas de qualidade de dados. Descobrir os primeiros discernimentos nos dados ou detectar subconjuntos interessantes para formar hipóteses da informação escondida.

Preparação dos Dados: A fase de preparação dos dados abrange todas as atividades para a construção final de dados (dados que serão integrados na ferramenta de modelagem a partir de dados brutos iniciais). Essa preparação visa seleção de tabelas, registros e atributos, a transformação e limpeza dos dados para as ferramentas de modelagem, podendo ser realizada várias vezes, e sem uma determinada ordem.

Modelagem: Nessa fase, várias técnicas de modelagem são selecionadas e aplicadas, e seus parâmetros são calibrados para obtenção de valores melhores. Normalmente para um mesmo tido de problema podem se aplicar várias técnicas de mineração de dados.

É citado por Larose (2005), que para execução da fase de modelagem, é indispensável o conhecimento sobre as técnicas de mineração de dados e o formato de cada uma delas pelo analista de dados. Em vários casos, algumas técnicas de mineração de dados deverão ser tomadas para abordar um determinado problema, demandando do analista de dados uma grande ciência sobre mineração de dados.

Avaliação: Nesta fase do projeto, tem-se construído um modelo (ou modelos) com alta qualidade, a partir de uma perspectiva da análise de dados. Antes de proceder o final de implantação do modelo, é importante avaliá-lo mais profundamente e rever as etapas executadas para sua construção atingindo os objetivos de negócio estabelecidos na primeira fase da metodologia.

Aplicação: Criação do modelo, geralmente, não é o fim do projeto. Mesmo que a finalidade do modelo seja a de aumentar o conhecimento dos dados, os conhecimentos adquiridos terão de ser organizados e apresentados de uma forma que o cliente possa utilizá-lo.

A seqüência das fases descritas não é rígida, o avanço ou retorno entre as diferentes fases pode ser necessário (CHAPMAN *et al*., 2000; OLSON & DELEN, 2008).

A metodologia CRISP-DM foi utilizada como aquisição do conhecimento para criação da rede semântica representando o conhecimento sobre a tarefa de classificação.

3.2 MINERAÇÃO DE DADOS

O processo de descoberta de conhecimento em bases de dados (*KDD - Knowledge Discovery in Databases*) foi proposto em 1989, por *Piatetsky-Shapiro*, para enfatizar que conhecimento é o produto final de uma descoberta impulsionada em dados (FAYYAD; PIATETSKY-SHAPIRO; SMITH, 1996 a*pud* MENDES, 2009).

O conjunto de atividades do processo de KDD é composto de sete etapas, segundo Han e Kamber (2006): limpeza dos dados, integração dos dados, seleção dos dados, transformação dos dados, mineração dos dados, avaliação dos padrões, apresentação do conhecimento, sendo que a etapa de mineração de dados é a mais importante desse processo.

3.2.1 CONCEITOS

Nesses últimos, anos tem-se verificado um crescimento da quantidade de dados armazenados em meios eletrônicos. Esses dados, produzidos e armazenados em larga escala, são inviáveis de serem lidos ou analisados por especialistas através de métodos tradicionais (Piatetsky-Shapiro, 1991), tais como planilhas de cálculos e relatórios informativos operacionais, nas quais o especialista testa sua hipótese contra a base de dados. Por outro lado, sabe-se que grande quantidades de dados equivalem a um maior potencial de informação, no entanto as informações contidas nos dados não estão caracterizadas explicitamente, uma vez que sendo dados operacionais, não interessam quando são estudados individualmente. Diante

desse cenário, surge a necessidade de se explorar esses dados para extrair informação – conhecimento implícito, e utilizá-lo no âmbito do problema.

Argumenta-se que a necessidade de sistemas para dar suporte à decisão tem-se desenvolvido ao longo dos anos cada vez mais dentro de uma granularidade de informações mais refinada, da seguinte maneira: nos anos 60 as exigências e necessidades estavam ao nível de mercado; nos anos 70, ao nível de nichos, grupos de interesse; nos anos 80, no nível de segmento de mercado; e nos anos 90, no nível de clientes. Esse último nível, naturalmente, requer o uso de mais dados para extrair conhecimento (KELLY, 1995). A exploração do valor desses dados, ou seja, a informação neles contida implicitamente depende de técnicas como regras de associação, classificação, agrupamento, entre outras, capazes de gerenciar tarefas complexas.

O uso de tecnologia e suas ferramentas permitem a mineração desses dados com a intenção de gerar um real valor do dado, transformando-o em informação e conhecimento; no caso de uma empresa, o auxílio é desde a tomada de decisão até a descoberta de fraudes ou perfis de consumidores (Fayyad *et al*. 1996). No entanto, para que estas informações sejam extraídas corretamente, é necessária a utilização de ferramentas que propiciem a descoberta ou mineração de padrões.

A mineração de dados pode ser vista também como um conjunto de técnicas automáticas de exploração de grandes massas de dados de forma a descobrir novos padrões e relações que, devido ao volume de dados, não seriam facilmente descobertas a olho nu pelo ser humano (Carvalho, 2005). De fato, muitas são as técnicas utilizadas, porém a mineração de dados ainda é mais uma arte do que uma ciência. O sentimento do especialista não pode ser dispensado, mesmo que as mais sofisticadas técnicas sejam utilizadas.

Elmasri & Navathe (2005) definem mineração de dados como “garimpagem” de dados, descoberta de novas informações em função de padrões ou regras em grandes quantidades de dados e que para ser realmente

útil precisa ser realizada eficientemente em grandes arquivos e banco de dados.

O Gartner Group[1] *apud* Elmasri & Navathe (2005), aponta em relatório, a mineração de dados como uma das tecnologias mais promissoras para um futuro próximo.

Para Fayyad; Piatetsky-Shapiro e Smith (1996), os objetivos da mineração de dados são definidos pelo usuário do sistema. A mineração de dados busca alcançar dois objetivos diferentes. A verificação cujo sistema limita-se a verificar as hipóteses do usuário. Descoberta, onde o sistema, automaticamente, busca por padrões. Esta, por sua vez, é subdividida em predição, onde o sistema busca por padrões para tentar prever o futuro; e a descrição, onde o sistema busca padrões para apresentar ao usuário a tentativa de proporcionar uma forma mais compreensível. Segundo Han e Kamber (2006), as tarefas descritivas têm por objetivo encontrar uma propriedade geral dos dados de um determinado banco de dados. A tarefa preditiva tem por objetivo executar inferências sobre os dados atuais a fim de fazer previsões.

Para os autores Han e Kamber (2006), para definir se uma determinada tarefa de mineração de dados gerou regras úteis, utilizam-se medidas conhecidas como medidas de interesse, para estabelecer o nível de interesse de uma determinada regra gerada. As regras geradas pelas tarefas precisam estar dentro do valor mínimo estabelecido para as medidas de interesse. Caso uma regra seja encontrada dentro dos padrões mínimos fornecidos pelo analista de dados, é considerada como uma regra de interesse.

De acordo com Berry e Linoff (2004), a mineração de dados é a exploração e a análise de grandes quantidades de dados, a fim de descobrir padrões e regras significativas. Segundo esses autores, o objetivo da mineração de dados é permitir que uma empresa melhore o seu marketing,

[1] O relatório do Gartner Group é um dos exemplos entre as muitas publicações sobre tecnologia em que os executivos confiam para tomar suas decisões relacionadas à tecnologia.

vendas e operações de apoio ao cliente através de uma melhor compreensão dos mesmos.

A mineração de dados pode ser alcançada através da aplicação de processos de: associação, classificação, agrupamento, predição, padrões seqüenciais e seqüências similares de tempo (OLSON e DELEN, 2008).

Na associação, a relação de um determinado item em uma transação de dados sobre outros itens na mesma transação é utilizada para prever os padrões, por exemplo, se um cliente compra um microcomputador portátil (X), então ele também adquire um mouse (Y) em 60% dos casos.

Em classificação, os métodos são destinados à aprendizagem de diferentes funções que mapeiam cada item dos dados selecionados em um conjunto predefinido de classes. Dado o conjunto de classes pré-definidas, uma série de atributos e um conjunto de treinamento, os métodos de classificação podem automaticamente prever a classe de outros dados de um conjunto de aprendizagem. Dois problemas-chave da investigação relacionada com os resultados da classificação são a avaliação da classificação e o poder de previsão. Técnicas matemáticas que são muitas vezes utilizadas para a construção de métodos de classificação são árvores de decisão binária e redes neurais. Usando árvores de decisão binária, um modelo de indução de árvore com um formato "sim-não" pode ser construído para separar os dados em classes diferentes de acordo com seus atributos (OLSON e DELEN, 2008). Usando redes neurais é possível classificar os dados de entrada em diversas saídas dependendo da arquitetura da rede neural (HAYKIN, 1994).

Problemas de classificação têm por objetivo identificar as características que indicam o grupo a que cada caso pertence. Esse padrão pode ser utilizado tanto para compreender os dados existentes e predizer como novas instâncias irão se comportar.

A mineração de dados cria modelos de classificação através da análise dos dados já classificados (casos) e indutivamente encontra um modelo

preditivo. Esses casos já existentes podem ser provenientes de um banco de dados históricos, por exemplo: pessoas que já tenham sido submetidas a um tratamento médico particular. Eles podem ser provenientes de um experimento no qual uma amostra do banco de dados inteiro é testada no mundo real e os resultados utilizados para criar um classificador. Às vezes, um especialista classifica uma amostra do banco de dados e essa classificação é então utilizada para criar o modelo que será aplicado ao banco de dados inteiro (*TWO CROWS CORPORATION*, 1999).

De acordo com Olson e Delen (2008), a análise de agrupamento de dados utiliza técnicas automáticas para colocar dados em grupos. A tarefa de agrupamento é uma tarefa que requer aprendizado não supervisionado, portanto não requer um conjunto de aprendizagem. Ele compartilha um terreno comum com a classificação metodológica. Em outras palavras, a maioria dos modelos matemáticos mencionados anteriormente em relação à classificação pode ser aplicada à análise de agrupamento também.

Análise de Predição está relacionada às técnicas de regressão. A idéia-chave de análise de predição é descobrir a relação entre as variáveis dependentes e independentes, a relação entre as variáveis independentes (um contra o outro, um contra o resto, e assim por diante). Por exemplo, se as vendas são uma variável independente, então o lucro pode ser uma variável dependente. Ao usar dados históricos de vendas e lucro, ou técnicas de regressão não linear pode produzir uma curva de regressão ajustada, que pode ser utilizada para a previsão de lucro no futuro (OLSON e DELEN, 2008).

Análise do padrão seqüencial visa encontrar padrões semelhantes em dados de transações durante um período de negócios. Esses padrões podem ser utilizados por analistas de negócios para identificar as relações entre os dados. Os modelos matemáticos de padrões seqüenciais são as regras da lógica, a lógica fuzzy. Como uma extensão dos padrões seqüenciais, seqüências similares de tempo são aplicadas para descobrir seqüências similares a uma seqüência conhecida em ambos os períodos de atividades empresariais passadas e atuais (OLSON e DELEN, 2008).

Neste trabalho será modelado o conhecimento necessário à escolha da tarefa de classificação.

3.2.2 TÉCNICAS DE MINERAÇÃO DE DADOS

De acordo com Harrison (1998), não há uma técnica que resolva todos os problemas de mineração de dados. Diferentes métodos servem para diferentes propósitos, cada método oferece suas vantagens e suas desvantagens. A familiaridade com as técnicas é necessária para facilitar a escolha de uma delas de acordo com os problemas apresentados.

As técnicas de mineração de dados normalmente usadas são Redes Neurais Artificiais, Algoritmos Genéticos, Árvores de decisão, etc. (HARRISON, 1998).

3.2.2.1 REDES NEURAIS ARTIFICIAIS

Conforme Barreto (2001), as redes neurais artificiais são sistemas compostos por vários neurônios de modo que as propriedades de sistemas complexos são usadas. Esses neurônios estão ligados por conexões, chamadas conexões sinápticas.

Redes neurais artificiais é um conceito da computação que visa trabalhar no processamento de dados de maneira semelhante ao cérebro humano. O cérebro é tido como um processador altamente complexo e que realiza processamentos de maneira paralela. Para isso, ele organiza sua estrutura, ou seja, os neurônios, de forma que eles realizem o processamento necessário. Isso é feito numa velocidade extremamente alta e não existe qualquer computador no mundo capaz de realizar o que o cérebro humano faz (ALECRIM, 2004).

As redes neurais oferecem uma arquitetura computacional robusta e distribuída, composta de significativas funções de aprendizado e capacidade de representação de relacionamentos não-lineares e multivariáveis (KLOSGEN *et. al.* 2002).

De acordo com Barreto (2001), essa tecnologia oferece um avançado poder de mineração, no entanto, é de difícil compreensão. Essas redes tentam construir representações internas de modelos ou padrões achados nos dados, mas essas representações não são apresentadas para o usuário, o processo de descoberta de padrões é tratado pelos programas de mineração de dados dentro de um processo chamado “caixa preta”.

3.2.2.2 ALGORITMOS GENÉTICOS

De acordo com Barreto (2001), os Algoritmos Genéticos (AGs) constituem um paradigma de aprendizado de máquina em que seu funcionamento encontra inspirações em um dos mecanismos básicos da evolução na natureza, chamado “seleção dura”.

Essencialmente, no paradigma dos AGs, proposto inicialmente por Holland em 1975, se cria no computador uma população de indivíduos representados por cromossomos, tal como na molécula de DNA do núcleo celular, chamados de população. Soluções de uma população são utilizadas para formar uma nova população. Isso é motivado pela esperança de que a nova população será melhor do que a primeira. Soluções que são selecionadas para formar novas gerações de soluções são selecionadas de acordo com sua adequação - quanto melhores mais chances de reprodução terão (BARRETO, 2001). Esse processo é repetido até que alguma condição é satisfeita (por exemplo, o número de populações ou o aperfeiçoamento da melhor solução).

3.2.2.3 ÁRVORES DE DECISÃO

As árvores de decisão são consideradas representações simples do conhecimento e, um meio eficiente de construir classificadores que predizem classes baseadas nos valores de atributos de um conjunto de dados.

De acordo com Mittchel (1997), as árvores de decisão classificam instâncias partindo da raiz da árvore para algum nodo folha que fornece a classe da instância. Cada nodo da árvore especifica o teste de algum

atributo da instância, e cada arco alternativo que desce daquele nodo corresponde a um dos possíveis valores desse atributo. Uma instância é classificada começando no nodo raiz da árvore e testa o atributo relacionado a esse nodo e segue o arco que corresponde ao valor do atributo na instância em questão. Este processo é repetido então para a sub-árvore até chegar a um nodo folha.

A figura 10 apresenta uma árvore de decisão típica. Esta árvore de decisão classifica os dias, conforme eles são satisfatórios ou não, para jogar tênis.

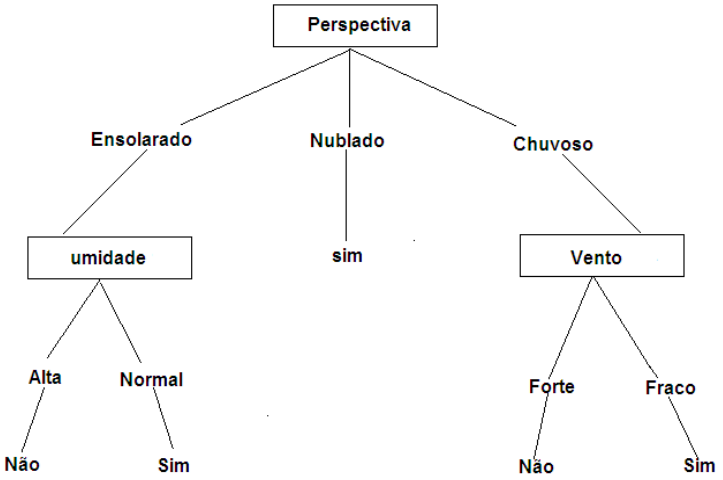

FIGURA 10: UMA ÁRVORE DE DECISÃO PARA O CONCEITO DE JOGAR TÊNIS.
Adaptado de Mittchel (1997).

Por exemplo, a instância (Perspectiva = Ensolarado, Umidade = Alta, Vento = Forte) seguirá o caminho mais à esquerda da árvore de decisão e será classificada então como uma instância negativa (a árvore prediz que jogar tênis = não).

Em geral, árvores de decisão representam uma disjunção de conjunções dos valores de atributo das instâncias. Cada caminho, da raiz da árvore para uma folha, corresponde a uma conjunção de testes de atributo, e a própria árvore uma disjunção destas conjunções, por exemplo, a árvore de decisão mostrada, corresponde à expressão:

(Perspectiva = Ensolarado ^ Umidade = Normal)

v (Perspectiva = Nublado)

v (Perspectiva = Chuvoso ^ Vento = Fraco)

O aprendizado utilizando árvore de decisão geralmente é mais indicada para problemas com as seguintes características:

- instâncias são representadas através de pares atributo-valor. Instâncias são descritas por um conjunto fixo de atributos (por exemplo: Temperatura) e seus respectivos valores (por exemplo: Quente);

- a função tem valores discretos. A árvore de decisão classifica com valores lógicos (Verdadeiro: sim ou Falso: não) para cada exemplo. Métodos de árvore de decisão podem ser facilmente estendidos para funções com mais de dois valores possíveis;

- permitem descrições disjuntas. Como notado acima, árvores de decisão naturalmente representam expressões disjuntas;

- os dados de treinamento podem conter erros. As árvores de decisão são robustas a erros, tanto erros nas classificações dos exemplos de treinamento, quanto erros nos valores dos atributos que descrevem esses exemplos;

- os dados de treinamento podem conter valores de atributo indefinidos. Podem ser usados métodos de árvore de decisão até mesmo quando alguns exemplos de treinamento têm valores desconhecidos (por exemplo, se a umidade do dia é conhecida somente em alguns dos exemplos de treinamento).

Muitos problemas práticos possuem essas características. Aprendizado utilizando árvore de decisão foi aplicado então a problemas como classificar os pacientes pela doença, causa de mau funcionamento de equipamentos e a probabilidade de candidatos a empréstimo ficarem inadimplentes. Tais problemas, nos quais a tarefa é classificar exemplos em

possíveis categorias discretas, são freqüentemente chamados de problemas de classificação (MITTCHEL, 1997).

São várias as tarefas de mineração de dados, porém, neste trabalho, estaremos abordando apenas a tarefa de classificação, pois a mesma é um dos nossos focos principais.

3.3 TAREFA DE CLASSIFICAÇÃO

As bases de dados são ricas em informações escondidas que podem ser usadas para tomada de decisões comerciais mais seguras. Classificação e predição são duas formas de análise de dados que podem ser usadas para extrair modelos que descrevem classes importantes de dados ou para prever as tendências futuras de dados, por exemplo, um modelo de classificação pode ser construído para categorizar pedido de empréstimo em banco, classificar pedidos de créditos como de baixo, médio e alto risco; esclarecer pedidos de seguros fraudulentos; identificar a forma de tratamento na qual um paciente está mais propício a responder, baseando-se em classes de pacientes que respondem bem a determinado tipo de tratamento médico e um modelo de previsão pode ser construído para prever os gastos de clientes potenciais em equipamentos de informática dada a renda e ocupação.

A classificação é uma das técnicas mais utilizadas de mineração de dados, simplesmente porque é uma das mais realizadas tarefas humanas no auxílio à compreensão do ambiente em que se vive, sendo, também, a tarefa mais estudada em KDD. As tarefas de classificação têm inúmeras aplicações, incluindo detecção de fraudes, alvos comerciais, fabricação e diagnóstico médico (HAN E KAMBER, 2006).

Para Berry (2004), a classificação baseia-se em análise das características de um objeto apresentado e sua atribuição a um conjunto definido de classes ou grupos. Já Elmasri e Navathe (2005) definem classificação como sendo o processo de encontrar um modelo que descreva classes diferentes de dados e as classes são predeterminadas.

Segundo Han e Kamber (2006), classificação é uma forma de análise dos dados que pode ser usada para extrair modelos que descrevem importantes classes de dados ou predizer tendências futuras. Essa análise pode ajudar a proporcionar uma melhor compreensão dos dados em geral. A tarefa de classificação ou predição tem por objetivo construir um modelo que será utilizado para classificar dados, visando categorizá-los em classes.

De acordo com os autores, o processo de classificação ou predição é dividido em duas etapas. A primeira etapa, mostrada na figura 11 a seguir, é chamada de fase de treinamento ou aprendizagem e tem por objetivo construir um classificador que descreve, previamente, um conjunto de dados em classes ou conceitos. O algoritmo de classificação constrói um classificador, analisando um conjunto de dados de treinamento. Cada instância do conjunto de dados de treinamento deve pertencer a uma classe previamente definida por um atributo chamado atributo classe.

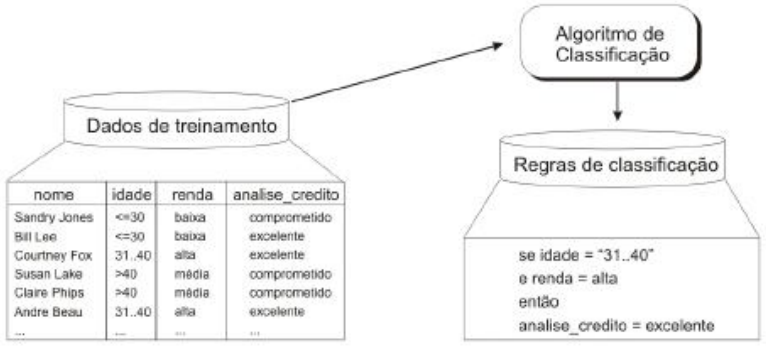

FIGURA 11 - PROCESSO DE CLASSIFICAÇÃO DE DADOS- 1ª ETAPA.

Fonte: Adaptado de Han e Kamber (2006).

Na segunda etapa, figura 12, de acordo com Han e Kamber (2006), o modelo é usado para classificar. Em primeiro lugar, a precisão preditiva do classificador é avaliada. A precisão de um classificador é avaliada, calculando o percentual de instâncias corretamente classificadas, de um determinado conjunto de teste. Se esse teste de precisão produzir resultados aceitáveis, o modelo poderá ser utilizado na classificação de novas instâncias.

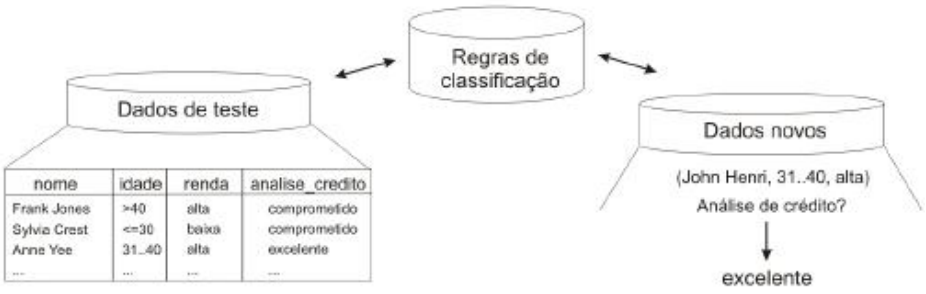

FIGURA 12 - PROCESSO DE CLASSIFICAÇÃO DE DADOS- 2ª ETAPA.

Fonte: Adaptado de Han e Kamber (2006)

Nessa tarefa, cada tupla pertence a uma classe entre um conjunto pré-definido de classes. A classe de uma tupla, ou registro, é indicado um valor pelo usuário em um atributo meta (FREITAS, 1998). As tuplas consistem de atributos preditivos e um objetivo, esse último indicando a que classe essa tupla pertence. O atributo objetivo é do tipo categórico, ou discreto, determinando classes ou categorias. Esse atributo pode ter valores como SIM ou NÃO, um código pertencente a um intervalo de números inteiros, tais como {1...10}, etc.

O princípio da tarefa de classificação é descobrir algum tipo de relacionamento entre os atributos preditivos e o atributo objetivo, de modo a descobrir um conhecimento que possa ser utilizado para prever a classe de uma tupla desconhecida, ou seja, que ainda não possui uma classe definida.

Em um processo de mineração de dados, a classificação está especificamente voltada à atribuição de uma das classes pré-definidas pelo analista a novos fatos ou objetos submetidos à classificação. Essa técnica pode ser utilizada tanto para entender dados existentes quanto para prever como novos dados irão se comportar (FREITAS, 1998).

A tarefa de classificar geralmente exige a comparação de um objeto ou dado com outros dados ou objetos que supostamente pertençam a classes anteriormente definidas. Para comparar os dados ou objetos utiliza-se uma métrica ou forma de medida de diferenças entre eles.

3.3.1 ALGORITMOS PARA CLASSIFICAÇÃO DE DADOS

Existe uma grande variedade de algoritmos para classificação de dados. Neste trabalho, será citado apenas três: ID3, J48 e C4.5. A citação dos três algoritmos deve-se ao fato dos mesmos serem populares e muito utilizados no meio acadêmico e por fazer parte do pacote de ferramentas Weka. O algoritmo J48 foi utilizado em um estudo de caso como processo de aquisição do conhecimento da tarefa de classificação.

3.3.1.1 ALGORITMO ID3

Segundo Mota (2004), o algoritmo ID3 utiliza a lógica e a matemática para processar, organizar e simplificar um grande conjunto de dados. Além disso, possui habilidade para operar dados não numéricos, sendo essa uma diferença entre ele e os métodos estatísticos, pois enquanto o ID3 assume atributos nominais, os métodos estatísticos utilizam atributos numéricos.

De acordo com Motta (2004), o algoritmo ID3 emprega a medida do ganho de informação para reduzir a incerteza sobre o valor do objeto de saída. O ganho de informação consiste em uma medida estatística utilizada para a construção das árvores de decisão a fim de escolher o atributo de teste entre todos os envolvidos com o nó em questão. O atributo que possui o maior ganho de informação é aquele que melhor classifica o conjunto de amostras de treinamento.

De acordo com Mendonça Neto (2001), a indução de árvores de decisão por meio do algoritmo ID3 é constituída dos seguintes passos:

1) selecionar um atributo para ser raiz da árvore e criar tantos ramos quantos valores tiver esse atributo;

2) utilizar a árvore gerada para classificar o conjunto de treinamento. Se todos os exemplos em uma folha tiverem o mesmo valor para o objeto de saída, retorne ao nó folha este valor;

3) senão, crie um nó com um atributo que ainda não foi utilizado em seus nós ancestrais, e crie todos os ramos possíveis para ele, a seguir retorne ao segundo passo.

A forma como o ID3 constrói árvores de decisão é *top-down*, ou seja, de cima para baixo. Este algoritmo consiste num processo recursivo, pois ao escolher um atributo para um nó, partindo da raiz, aplica o mesmo algoritmo aos descendentes desse nó, até que certos critérios de parada sejam atingidos.

Segundo Mitchell (1997), o algoritmo básico, ID3, constrói árvores de decisão a partir da raiz e começa com a pergunta “que atributo deveria ser testado na raiz da árvore?”. Para responder à pergunta, cada atributo da instância é avaliado usando um teste estatístico para determinar como este classifica os exemplos de treinamento. O melhor atributo é selecionado e é usado como o teste no nodo raiz da árvore. Um descendente do nodo raiz é criado então para cada possível valor deste atributo, e os exemplos de treinamento são particionados e associados a cada nodo descendente para selecionar o melhor atributo para testar naquele ponto na árvore. Isso forma uma procura por uma árvore de decisão aceitável na qual o algoritmo nunca retrocede para reconsiderar escolhas feitas anteriormente.

3.3.1.2 ALGORITMO J48

De acordo com Martins e Costa (2009), o algoritmo J48 permite a criação de modelos de árvore de decisão. O modelo de árvore de decisão é construído pela análise dos dados de treino e o modelo utilizado para classificar dados ainda não classificados. O J48 gera árvores de decisão em que cada nó da árvore avalia a existência ou significância de cada atributo individual. As árvores de decisão são construídas do topo para a base, através da escolha do atributo mais apropriado para cada situação. Uma vez escolhido o atributo os dados de treino são divididos em subgrupos, correspondendo aos diferentes valores dos atributos e o processo é repetido para cada subgrupo até que uma grande parte dos atributos em cada subgrupo pertença a uma única classe.

Segundo os mesmos autores, a indução por árvore de decisão é um algoritmo que habitualmente aprende um conjunto de regras com elevada acuidade. Esse algoritmo é escolhido para comparar a percentagem de acerto com outros algoritmos.

3.3.1.3 ALGORITMO C4.5

O algoritmo C4.5 é uma extensão do ID3, com alguns aperfeiçoamentos, tais como: redução em erros de poda na árvore; implementação de regras de verificação após poda; capacidade de trabalhar com atributos contínuos; capacidade de trabalhar com atributos de valores ausentes e aperfeiçoamento da eficiência computacional.

Esse algoritmo constrói um modelo de árvore de decisão baseado num conjunto de dados de treinamento, sendo que esse modelo é utilizado para classificar as instâncias do conjunto de teste (QUINLAN, 1993).

De acordo com Quinlan (1993) *apud* Collin *et al.* (2009), C4.5 é um algoritmo de indução de árvore de decisão baseado num conjunto de dados de treinamento, sendo que esse modelo é utilizado para classificar as instâncias do conjunto de teste.

Este programa gera um classificador na forma de uma árvore de decisão, com uma estrutura composta por (QUINLAN, 1993):

- uma folha, indicando uma classe ou um nó de decisão que especifica um teste a ser realizado no valor de um único atributo, com um galho, para cada resposta possível do teste, que levará para um sub-árvore ou uma folha.

Em uma árvore de decisão, a classificação de um caso se inicia pela raiz da árvore, e essa árvore é percorrida até que se chegue a uma folha. Em cada nó de decisão será feito um teste que irá direcionar o caso para uma sub-árvore. Esse processo irá guiar-se para uma folha. A classe do caso se pressupõe que seja a mesma que está armazenada na folha e de acordo

com Quinlan (1993), uma das variantes mais conhecidas e usadas de árvores de decisão é a do algoritmo C4.5.

4 MODELAGEM E TESTE DA BASE DE CONHECIMENTO

Para haver decisão sobre um determinado assunto, um especialista o faz a partir de fatos que encontra e de hipóteses enunciadas de uma maneira precisa, investigando em sua memória um conhecimento previamente armazenado durante anos, à época de sua formação e no tempo decorrido de sua vida profissional, sobre esses fatos e hipóteses. É feito de acordo com conhecimentos resultantes de vivências subjetivas, isto é, com o seu conhecimento previamente reunido sobre o assunto e, com esses fatos e hipóteses, enuncia a decisão.

Durante o decurso do raciocínio, observa-se qual o significado e a importância dos fatos que encontra, comparando-os com as informações já inseridas no seu discernimento acumulado desses fatos e hipóteses. Assim, nesse processo, formula novas hipóteses e observando novos fatos; e esses vão exercer influências no processo de raciocínio. Esse raciocínio é sempre baseado no conhecimento previamente acumulado. Um especialista, com esse processo de raciocínio poderá não atingir uma decisão se os fatos de que dispõe para aplicar o seu conhecimento prévio não forem aptos. Pode, inclusive, por esse motivo, chegar a um resultado errôneo, sendo esse justificado em função dos fatos encontrados e do seu conhecimento anteriormente acumulado.

Um sistema especialista deve, além de inferir resultados concretos, ter a capacidade de adquirir novos conhecimentos e, assim, tornar melhor o seu desempenho de raciocínio e também a qualidade de suas deliberações.

Apresentamos neste capítulo os trabalhos realizados: rede semântica, base de conhecimento (composta por perguntas de direcionamento da tarefa de classificação), regras de produção e testes da base de conhecimento.

4.1 REDE SEMÂNTICA

De acordo com Passos (1989), a representação do conhecimento através de redes semântica, é uma tentativa de simular o modelo psicológico de memória associativa humana, modelando o conhecimento como um conjunto de pontos chamados nós ou nodos, conectados por ligações chamados arcos que descrevem as relações entre os nós.

A figura 13 representa o processo de mineração de dados utilizando a tarefa de classificação, processo este representado por meio de uma rede semântica.

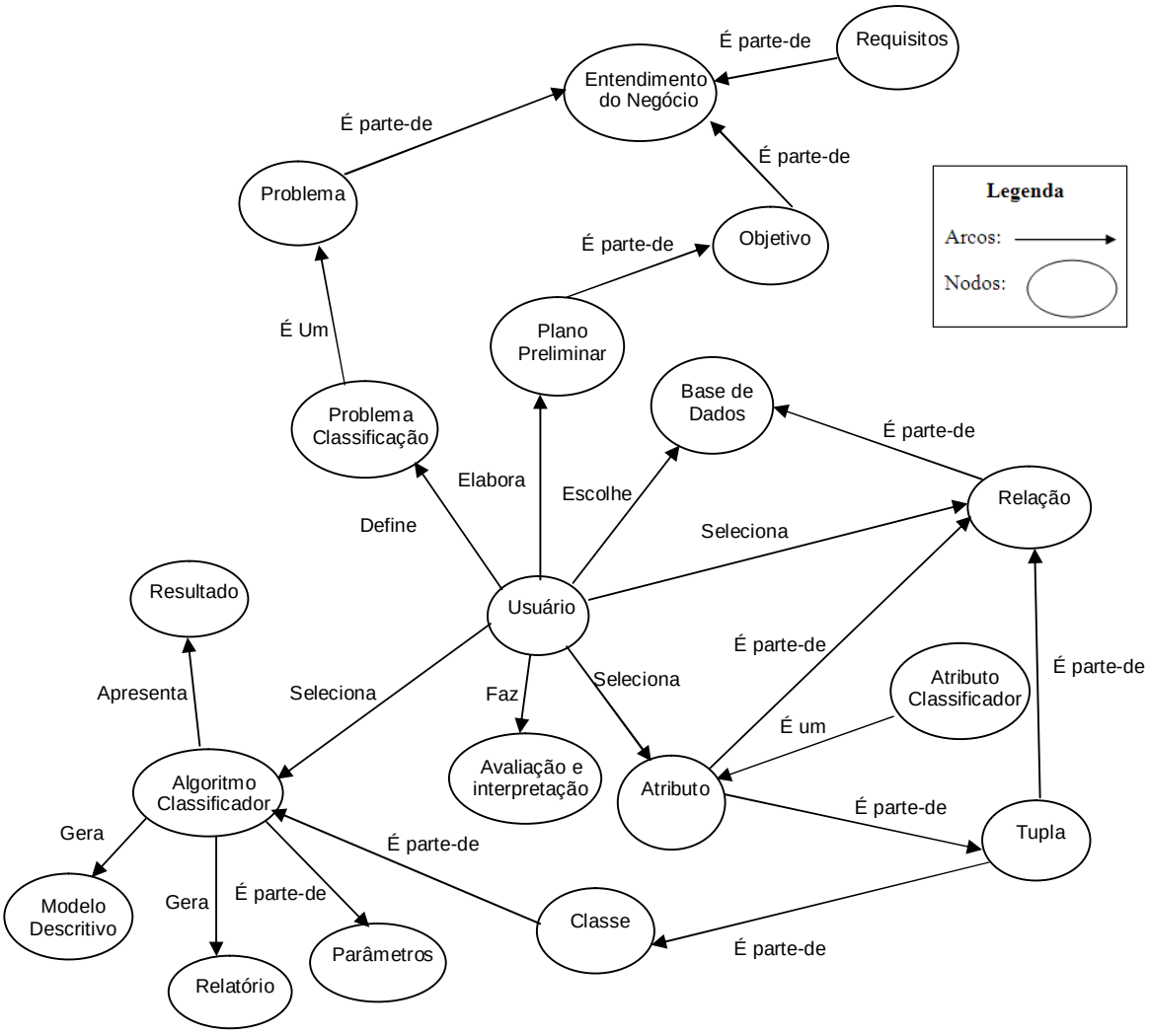

FIGURA 13: REDE SEMÂNTICA - TAREFA DE CLASSIFICAÇÃO

Nesta rede, as elipses representam conceitos presentes no domínio de mineração de dados com ênfase na tarefa de classificação. Esses conceitos foram semanticamente relacionados por meio de arestas. As setas indicam o sentido da leitura do relacionamento.

Muitos dos conceitos utilizados na criação da rede semântica foram retirados de estudos realizados na metodologia CRISP-DM, tais como: entendimento de negócios, plano preliminar, requisitos, avaliação e interpretação, etc.

4.2 Base de Conhecimento

A fim de construir uma base de conhecimento para instruir a escolha da tarefa de classificação em mineração de dados, o problema do domínio foi estudado de acordo com as características que possam conduzir à escolha ou a rejeição da tarefa de classificação.

A base de conhecimento é considerada a parte principal de um sistema especialista; contem conhecimento sob a forma de regras de produção, quadros, redes semânticas, ou seja, de várias formas. Uma das mais comuns é por sentenças do tipo: “SE – ENTÃO”.

Neste trabalho, para a construção da base de conhecimento foram utilizadas várias técnicas de aquisição de conhecimento, dentre elas destacam-se: entrevista com especialista, análise de discurso e a realização de estudo de caso utilizando a tarefa de classificação em banco de dados real.

A partir da aplicação dessas técnicas, várias perguntas foram elaboradas com o objetivo de direcionar a escolha da tarefa de classificação.

Considerando as respostas à essas perguntas, a base de conhecimento é executada. As perguntas foram fundamentadas consoantes com vários autores que tratam do tema mineração de dados. A seguir as perguntas são apresentadas e a explicação teórica de cada uma delas é mostrada de acordo com a literatura pesquisada.

➔ **Você deseja identificar muitos atributos para fazerem parte do processo de mineração?**

Segundo Bing Liu, Wynne Hsu, Yiming Ma (1998), conjuntos de dados de classificação contêm frequentemente muitos atributos. As bases de dados podem conter inúmeros atributos, que podem ser categóricos ou numéricos. Um atributo é um valor de dado assumido pelos objetos de uma classe. Nome, idade e peso são exemplos de atributos de objetos Pessoa. Cor, peso e modelo são possíveis atributos de objetos Carro. Cada atributo tem um valor para cada instância de objeto (HAN e KAMBER, 2006).

Dados categóricos são dados discretos. Atributos categóricos têm um finito - mas possivelmente grande número de valores distintos, sem nenhuma ordem entre os valores. Exemplos incluem localização geográfica, da categoria profissional e tipo de item. Existem vários métodos para a geração de hierarquias de conceitos para dados categóricos (HAN e KAMBER). Atributos categóricos são também chamados de atributos nominais, porque seus valores são "nomes das coisas" (HAN e KAMBER).

Um exemplo de um atributo categórico é cor, cujo domínio inclui valores como castanho, preto, branco, etc. (GUHA; RASTOGI; SHIM, 2000).

O fato de que o banco de dados do usuário apresenta vários atributos, especialmente do tipo categórico, é, então, um indicador de que os algoritmos de classificação de mineração de dados podem ser adequados para serem aplicados.

➔ **Os atributos escolhidos são todos numéricos?**

Outra característica a ser considerada na escolha da tarefa de classificação é o tipo de atributos apresentado em bases de dados do usuário: se os atributos são todos numéricos.

Um atributo é dito numérico quando ele pode ser dimensionado e representado por um número. Por exemplo, se tivéssemos registros de

pessoas, exemplos de atributos numéricos poderiam ser: idade, altura, peso, entre outros.

Os atributos numéricos eventualmente podem ser transformados em categóricos, por exemplo, a saber, o atributo altura poderia ter seus valores numéricos representados pelos valores categóricos: “alto” ou “baixo” (HAN E KAMBER, 2006).

Técnicas de discretização de dados podem ser usadas para reduzir o número de valores para um determinado atributo contínuo, dividindo o intervalo do atributo em intervalos. Intervalo de etiquetas pode ser usado para substituir os valores reais de dados. Substituindo os inúmeros valores contínuos de um atributo por um pequeno número de etiquetas de intervalo, assim, reduz e simplifica os dados originais. Isso leva a uma representação do conhecimento conciso e fácil de usar - em nível de resultados de mineração.

Técnicas de discretização podem ser classificadas com base em como a discretização é realizada, como ela usa informações de classe ou de que direção ela procede, ou seja, de cima para baixo ou de baixo para cima (HAN E KAMBER, 2006).

Atributos quantitativos são numéricos e têm implícita uma ordenação entre os valores (por exemplo, idade, renda, preços). Técnicas de mineração de regras de associação multidimensionais podem ser categorizadas em duas abordagens básicas sobre o tratamento de atributos quantitativos.

Na primeira abordagem, os atributos quantitativos são discretizados usando hierarquias de conceitos pré-definidos. Essa discretização ocorre antes da mineração, por exemplo, uma hierarquia de conceitos para a renda pode ser usada para substituir os valores numéricos originais desse atributo por rótulos de intervalo, como: "0. . . 20K "," 21K. . . 30K "," 31K. . . 40K ", e assim por diante. Na segunda abordagem, os atributos quantitativos são discretos ou agrupados em "caixas" com base na distribuição dos dados. Essas caixas podem ainda ser combinadas durante o processo de

mineração. O processo de discretização é dinâmico e estabelecido de modo a satisfazer alguns critérios de mineração, tais como maximizar a confiança das regras mineradas, pois essa estratégia trata os valores do atributo numérico como quantidades e não como intervalos predefinidos ou categorias (HAN E KAMBER, 2006).

Podemos converter os atributos numéricos para nominais usando um método de discretização simples. Em primeiro lugar, classificar os exemplos de treinamento de acordo com os valores do atributo numérico. Isso produz uma sequência de valores de classe (WITTEN, 2005).

A possibilidade de categorizar atributos numéricos torna-se possível a aplicação de algoritmos de classificação, no entanto, se existem vários atributos numéricos na base de dados e não é possível transformá-los em categóricos, então não é aconselhável a utilização de algoritmos de classificação.

➔ O volume de dados é grande?

De acordo com Berry e Linoff (2004, p. 12), a mineração de dados faz mais sentido, quando há grandes volumes de dados. Na verdade, a maioria dos algoritmos de mineração de dados requer grandes quantidades de dados a fim de construir e treinar os modelos que serão usados para executar a tarefa de classificação em mineração de dados. Grandes quantidades de dados é uma característica desejável do banco de dados do usuário para que se possam aplicar algoritmos de classificação.

➔ O atributo classificador é uma categoria?

Outra característica na tarefa de classificação, é que o usuário pode ser capaz de identificar um atributo categórico. Pois, na tarefa de classificação, um modelo ou classificador é construído para predizer rótulos categóricos, tais como "seguro" ou "de risco" para os dados do pedido de empréstimo; "sim" ou "não" para os dados de comercialização, ou "tratamento A", "tratamento B", ou "tratamento C" para os dados médicos.

Essas categorias podem ser representadas por valores discretos, por exemplo, os valores 1, 2 e 3 podem ser usadas para representar os tratamentos A, B, e C, em que não há ordenação implícita entre este grupo de regimes de tratamento (HAN E KAMBER, 2006, P. 286). Os atributos categóricos podem assumir apenas uma quantidade finita de valores. Dizemos que cada um desses valores pertence a uma classe. Por exemplo, se tivéssemos registros de pessoas, um atributo categórico poderia ser o atributo sexo, que pode assumir um dos valores categóricos: “masculino” ou “feminino”.

➔ **Os atributos são parte de um banco de dados Transacional?**

Para aplicar algoritmos de classificação é importante saber se o banco de dados do usuário é um banco de dados transacional.

Segundo Han e Kamber (2001, p. 12), em geral, um banco de dados transacional consiste de um arquivo, onde cada registro representa uma transação. Uma transação normalmente inclui um número de identificação único de transação (trans ID), e uma lista dos itens que compõem a operação (tais como itens comprados em uma loja).

TABELA 1: TABELA TRANSAÇÃO

Conta / Empréstimo						
ID_Empréstimo	ID_Conta	Valor	Duracão	Pagamento	Frequência	Data
1	124	1000	12	120	Mensal	03/12/2009
2	108	10000	24	500	Semanal	02/12/2009

(Han e Kamber, 2006, p. 575)

Portanto, se o banco de dados do usuário estiver sob a forma de um banco de dados transacional é mais adequado utilizar algoritmos de associação ao invés de algoritmos de classificação.

➔ **Para o seu problema de mineração você precisa pré-determinar um único alvo para a mineração?**

Um único alvo corresponde à escolha de uma classe. Classe é o atributo objetivo do problema, a característica que se quer classificar, ou, de outra forma, é a variável dependente dos outros atributos.

Segundo Bing Liu, Wynne Hsu, Yiming Ma (1998), na mineração de dados com regras de associação, o objetivo não é predeterminado, enquanto que na mineração com regras de classificação é um e apenas um alvo predeterminado, ou seja, a classe.

Já para Elmasri & Navathe (2006), classificação é o processo de encontrar um modelo que descreva classes diferentes de dados. Essas classes são predeterminadas.

Considerando-se o problema que o usuário precisa aplicar algoritmos de mineração de dados, é importante conhecer os seus objetivos em termos de dados. Se o usuário quer pré-determinar uma meta para realizar a mineração, então é altamente aconselhável a utilização de algoritmos de classificação.

➔ **Você consegue identificar esse alvo entre os seus atributos?**

Além de ter como objetivo de mineração de dados um processo de classificação, o usuário tem que ser capaz de identificar em seu banco de dados entre os atributos, um que poderá ser usado como rótulo de classe ou atributo alvo. Sem esse atributo a tarefa de classificação não é possível ser aplicada.

O atributo alvo é do tipo categórico, ou discreto, determinando classes ou categorias. Esse atributo pode ter valores como SIM ou NÃO, um código pertencente a um intervalo de números inteiros, tais como {1...10}, etc (FREITAS, 1998). Atributos que assumem uma grande variedade de valores

podem ser agrupados em algumas categorias, por exemplo, a renda mensal do cliente pode ser agrupada em 4 categorias: baixa, média, média-alta e alta, etc.

Classe é o atributo objetivo do problema. É a característica que se quer classificar, ou, de outra forma, é a variável dependente dos outros atributos, enquanto que categoria significa o valor que a classe recebe, como por exemplo, SIM, NÃO, VERDADEIRO, FALSO, etc.

Em mineração de dados usando a tarefa de classificação, o atributo alvo deverá ser identificado ou criado.

➔ **Você possui um conjunto de dados que pode ser utilizado como treinamento?**

Segundo Han e Kamber, (2006) os registros ou amostras individuais que formam o conjunto de dados de treinamento são denominados amostras de treino e são selecionadas aleatoriamente da população total (recomenda-se a utilização de técnicas de redução de dados e amostragem para a seleção desses objetos). Grande parte dos problemas de classificação de padrões de interesse real tem duas fases: Treinamento (aprendizado) executado a partir do banco de dados existente e a fase de Generalização, em que são apresentados dados que não foram utilizados no treinamento. Já para Amo, (2003), a classificação é um processo de 3 etapas. A primeira é a da criação do modelo de classificação, ou seja, a etapa de treinamento. Esse modelo é constituído de regras que permitem classificar as tuplas do banco de dados dentro de um número de classes pré-determinado.

A segunda etapa é a da verificação do modelo ou Etapa de Classificação: as regras são testadas sobre outro banco de dados, completamente independente do banco de dados de treinamento, chamado de banco de dados de testes.

A terceira etapa é a de utilização do modelo em novos dados: após o modelo ter sido aprovado nos testes das etapas anteriores, ele é aplicado em novos conjuntos de dados.

Cada instância do conjunto de dados de treinamento deve pertencer a uma classe previamente definida por um atributo chamado de atributo de classe, quanto maior a quantidade de dados mais informações poderão ser utilizadas em treinamento.

➔ **O conjunto de execução é diferente do conjunto de treinamento?**

Segundo Amo (2003), na classificação, as regras são testadas sobre outro banco de dados, completamente independente do banco de dados de treinamento, chamado de banco de dados de testes ou conjunto de execução. A qualidade do modelo é medida em termos da porcentagem de tuplas do banco de dados de testes que as regras do modelo conseguem classificar de forma satisfatória. É claro que, se as regras forem testadas no próprio banco de dados de treinamento, elas terão alta probabilidade de estarem corretas, uma vez que este banco foi usado para extraí-las, por isso a necessidade de um banco de dados completamente novo.

Segundo Han e Kamber (2006), com a classificação, o conjunto de treinamento usado para construir um indicador não deve ser usado para avaliar a sua precisão. Um conjunto de teste independente deve ser utilizado. Portanto, na tarefa de classificação é necessário obter dois conjuntos de dados diferentes cada um com vista a atingir objetivos diferentes.

➔ **Você deseja utilizar o resultado da mineração para poder aplicá-lo a outros dados?**

A tarefa de classificação é uma forma de análise dos dados que pode ser utilizada na extração de modelos que descrevem importantes classes de dados ou de predição de tendências futuras. (HAN e KAMBER, 2006). A tarefa de classificação ou predição tem por objetivo construir um modelo que será utilizado para classificar dados, visando a categorizá-los em classes. Empresas de qualquer segmento de mercado, que necessitem de conhecimento estratégico, vêem na mineração de dados forte aliado para

melhorar a lucratividade da empresa e tomar decisões mais acertadas com relação à previsão futura.

De acordo com Han e Kamber (2006), as regras descobertas no processo de mineração de dados utilizando a tarefa de classificação podem ser usadas para categorizar amostras de dados futuros, bem como proporcionar uma melhor compreensão do conteúdo do banco de dados.

Desta forma, se as empresas desejam ter um modelo para classificar as futuras entradas de dados, então a tarefa de classificação pode ser uma possibilidade.

➔ **Você deseja utilizar o resultado da mineração para poder tomar alguma decisão imediata na organização?**

Outro ponto que suporta a escolha da tarefa de classificação é a utilização dos dados resultantes do processo de mineração. Ao contrário dos resultados de regras de associação, por exemplo, os resultados da tarefa de classificação não têm que ser imediatamente utilizados. As classes podem ser estudadas por um período de tempo e as decisões em médio prazo poderão ser tomadas durante este período.

Segundo Harrison (1998), a tarefa de classificação consiste na construção de um modelo de algum tipo que possa ser aplicado a dados não classificados visando a categorizá-los em classes. Um objeto é examinado de acordo com uma classe definida. Pode-se citar como exemplo a classificação de pedidos de crédito como sendo de baixo risco, médio risco e alto risco, baseando-se em classes previamente definidas (Han e Kamber, 2006).

Na tarefa de classificação, os dados da mineração serão utilizados depois de alguma análise relativa às classes formadas.

Considerando as variáveis associadas às perguntas mencionadas, a base de conhecimento foi construída. Essa base é representada por meio de regras SE-ENTÃO. O raciocínio utilizado para a construção da base foi o encadeamento para frente.

No encadeamento para frente, também chamado encadeamento dirigido por dados, a parte esquerda da regra é comparada com a descrição da situação atual, contida na memória de trabalho. As regras que satisfazem a essa descrição têm sua parte direita executada, o que, em geral, significa a introdução de novos fatos na memória de trabalho.

Segundo Rezende *et. al.* (2003), no processo de encadeamento para frente, chega-se a uma solução para o problema a partir das informações fornecidas pelo usuário. Esse processo consiste em analisar esses dados, baseando-se no conhecimento, até chegar a uma conclusão.

O mecanismo de inferência que utiliza encadeamento para frente é baseado na busca do sucesso de uma regra através da checagem do Antecedente de uma Regra e depois de seu Consequente. A solução é encontrada partindo-se do antecedente e tentando-se provar o consequente.

De acordo com as respostas às perguntas, as regras de produção são formadas por um conjunto de predicados demonstrados na tabela 02.

Conforme Russel e Norvig (2004), predicado é uma proposição da lógica, e na lógica proposicional, símbolos representam proposições inteiras (fatos). A lógica consiste no seguinte:

a) Um sistema formal para descrever estados de coisas, que consiste em:

- a sintaxe da linguagem, que descreve como fazer frases, e;

- a semântica da linguagem, que dispõe sobre as restrições sistemáticas como as sentenças se relacionam com estados de coisas.

b) a teoria da prova - um conjunto de regras para deduzir as vinculações de um conjunto de frases dispostas.

Para cada pergunta, relativa ao processo de mineração, um predicado é criado. Dependendo da resposta à pergunta, o predicado tem

determinado valor associado à ele. Esses valores são utilizados para executar a base de conhecimento de acordo com as regras. Os predicados formados estão apresentados na tabela 01.

TABELA 2- PREDICADOS

Perguntas relacionadas ao predicado	Predicados	Explicação
Você deseja identificar muitos atributos para fazer parte do processo de mineração?	Vários_atributos	Dado o conjunto de classes pré-definidas, uma série de atributos, e um conjunto de treinamento, os métodos de classificação podem automaticamente prever a classe de outros dados não confidenciais de um conjunto de aprendizagem. Conjuntos de dados de classificação contêm frequentemente muitos atributos.
Os atributos escolhidos são todos numéricos?	Todos_numéricos	Na tarefa de classificação, um modelo ou classificador é construído para predizer rótulos categóricos.
O volume de dados é grande?	Volume_dados _grande	A maioria dos algoritmos de mineração de dados requer grande quantidade de dados a fim de construir e treinar os modelos que irão ser usado para executar a classificação.
O atributo classificador é uma categoria?	Categoria	Na tarefa de classificação, um modelo ou classificador é construído para predizer rótulos categóricos.
Os atributos são parte de um banco de dados transacional?	Transação	Um banco de dados transacional consiste de um arquivo, onde cada registro representa uma transação. Na tarefa de classificação, utiliza-se dados do passado para prever dados futuros.
Para o seu problema de mineração você precisa pré-determinar um único alvo para a mineração?	Alvo_único	Na mineração de dados com regras de classificação, um e apenas um alvo é pré-determinado, ou seja, a classe.
Você consegue identificar esse alvo entre os seus atributos?	Identificar	O atributo alvo é do tipo categórico, ou discreto, determinando classes ou categorias. Para que haja classificação, o atributo alvo deve ser identificado ou criado.
Você possui um conjunto de dados que pode ser utilizado como treinamento?	Dados_treinamento	O algoritmo de classificação constrói um classificador, analisando um conjunto de dados de treinamento. Cada instância do conjunto de dados de treinamento deve pertencer a uma classe previamente definida por um atributo chamado atributo classe. quanto maior a quantidade de dados mais informação poderá ser usada no treinamento.
O conjunto de execução é diferente do conjunto de treinamento?	Conjunto_execução	Na classificação as regras são testadas sobre um outro banco de dados, completamente independente do banco de dados de treinamento, chamado de banco de dados de testes ou conjunto de execução.
Você deseja utilizar o resultado da mineração para poder aplicá-lo a outros dados?	Aplicar_resultado	Na classificação, a precisão preditiva do classificador é avaliada. A precisão de um classificador é avaliada e se esse teste de precisão produzir resultados aceitáveis, o modelo poderá ser utilizado na classificação

		de novas instâncias. Essa técnica pode ser utilizada tanto para entender dados existentes quanto para prever como novos dados irão se comportar.
Você deseja utilizar o resultado da mineração para poder tomar alguma decisão imediata na organização?	Resultado_decisão	Na tarefa de classificação, os dados da mineração serão utilizados depois de alguma análise relativa às classes formadas.

Observa-se na Tabela 2, que os predicados classificação e tarefa_nao_identificada não aparecem nas perguntas. Esses predicados são usados para formular o resultado final das regras. Os possíveis valores assumidos pelas variáveis são descritos por dois conjuntos nomeados predicado_valores_binário e Predicado_valores_linguísticos. Apenas um predicado pode assumir valor linguístico: classificação (VIEIRA *et al*, 2009). Ao assumir valor linguístico esse predicado é capaz de apresentar níveis diferentes de diagnóstico para a tarefa de classificação, tais como: baixo, médio, alto e muito_alto, conforme Figura 14 a seguir.

```
Valores_Predicado_Binário = {sim, não}
Predicados_Valores_Linguísticos = { baixo, medio, alto, muito_alto}
```

FIGURA 14 - PREDICADOS BINÁRIO E LINGÜÍSTICOS

Na figura 15, são apresentadas as regras de produção que representam a base de conhecimento da tarefa de classificação.

R01. **Se** vários atributos = sim
Então Classificação = médio
R02. **Se** vários atributos = não
Então Classificação = baixo
R03. **Se** todos_numericos = sim e Classificação = médio
Então Classificação = baixo
R04. **Se** todos_numericos = não e varios_atributos = sim
Então Classificação = médio
R05. **Se** Volume_dados_grande = sim e classificação = médio
Então Classificação = alto
R06. **Se** volume_dados_grande = não
Então Classificação = baixo
R07. **Se** categoria = sim e classificação = médio
Então classificação = alto
R08. **Se** categoria = não e (classificação = médio ou classificação = alto)

Então tarefa_não_identificada = sim
R09. **Se** transação = sim
Então classificação = baixo
R10. **Se** transação = não e classificação = médio
Então classificação = alto
R11. **Se** alvo_unico = sim e classificação = médio
Então classificação = alto
R12. **Se** alvo_unico = não e classificação = baixo
Então classificação = baixo
R13. **Se** identificar = sim e classificação = médio e alvo_unico = sim
Então classificação = alto
R14. **Se** identificar = não e alvo_unico = não
Então tarefa_não_identificada = sim
R15. **Se** dados_treinamento = sim e classificação = médio
Então classificação = alto
R16. **Se** dados_treinamento = não
Então classificação = baixo
R17. **Se** dados_treinamento = sim e alvo_unico = sim e identificar = sim e classificação = alto
Então classificação = muito_alto
R18. **Se** conjunto_execução = sim e classificação = medio
Então classificação = alto
R19. **Se** conjuntos_execução = não
Então classificação = baixo
R20. **Se** conjunto_execução = não e dados_treinamento = não e classificação = alto
Então tarefa_não_identificada = sim
R21. **Se** aplicar_resultado = sim e classificação = médio
Então classificação = alto
R22. **Se** aplicar_resultado = não
Então classificação = baixo
R23. **Se** resultado_decisão = sim e aplicar_resultado = sim e Classificação = alto
Então Classificação = muito_alto
R24. **Se** resultado_decisão = não e classificação = baixo
Então classificação = baixo
R25. Se tarefa_não_identificada = sim
Então classificação = não_e_possivel

FIGURA 15 - REGRAS DE PRODUÇÃO DA BASE DE CONHECIMENTO

A figura 16 mostra um exemplo de execução da base conforme respostas dadas pelo usuário:

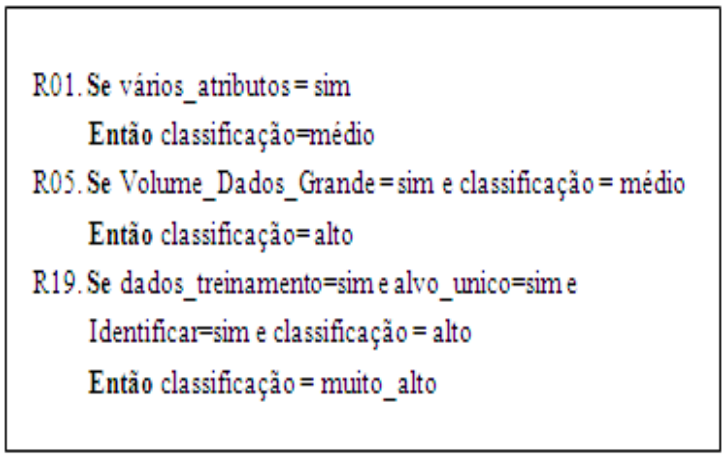

FIGURA 16 - EXEMPLO DE EXECUÇÃO DA BASE DE CONHECIMENTO

4.2.1 TESTES DA BASE DE CONHECIMENTO

Os testes da base de conhecimento foram realizados no programa Chimera. Este programa é uma Shell e tem como objetivo a implementação de sistemas especialistas. Esse é um trabalho de iniciação científica realizado pelo aluno André Bindilatti da Universidade Metodista de Piracicaba (BINDILATTI, 2009). Essa ferramenta permite que o programador do sistema preocupe-se somente com a representação do conhecimento do especialista do domínio, ficando para a Shell a ocupação de interpretar o conhecimento representado e executá-lo, além de permitir depurações e explicações de como o computador chegou àquela conclusão. O principal papel de uma Shell é simplificar ao máximo o trabalho de implementação de um sistema especialista e permitir seu uso por qualquer pessoa sem conhecimento de informática.

Para a realização dos testes na base de conhecimento, foi necessário implementar a base na ferramenta Chimera.

A primeira etapa realizada foi a implementação dos fatos, ou seja, um conjunto de predicados, necessários na criação das regras de produção, onde cada pergunta, relativa ao processo de mineração de dados, recebe um predicado. Dependendo da resposta à pergunta, o predicado tem determinado valor associado à ele. A implementação dos fatos pode ser observado na figura 17.

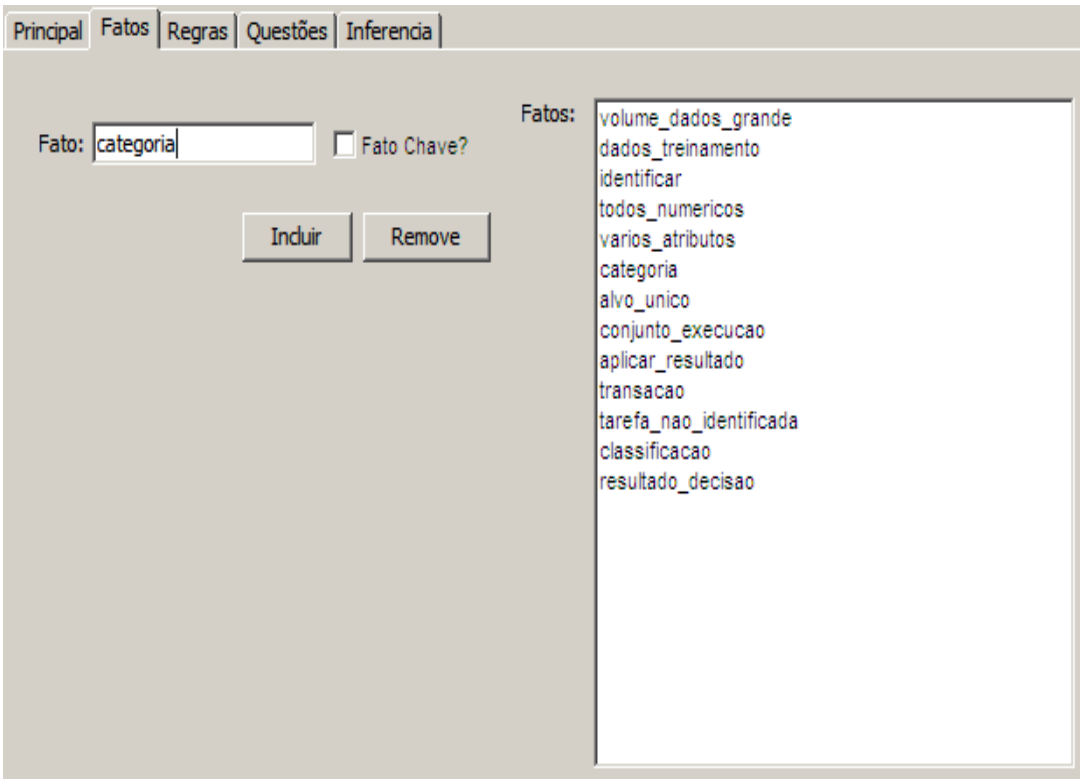

FIGURA 17 - IMPLEMENTAÇÃO DOS PREDICADOS.

Na etapa a seguir foram implementadas as regras de produção referente à base de conhecimentos de instrução para a tarefa de classificação na forma de Se <condição> Então <ação>. Foram criadas 25 regras específicas para a tarefa de classificação de acordo com a figura 18.

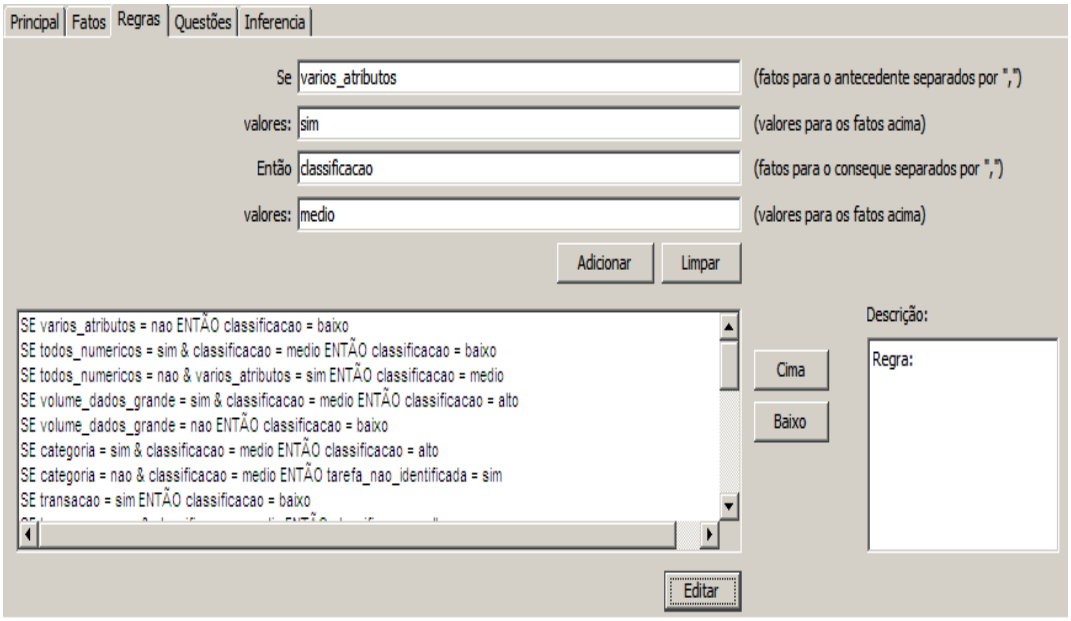

FIGURA 18 - IMPLEMENTAÇÃO DAS REGRAS

A figura 19 mostra a implementação das perguntas de direcionamento da tarefa de classificação. As perguntas possuem enfoque instrucional proporcionando uma representação do conhecimento necessário para o entendimento e escolha da aplicação da tarefa de classificação em problemas de mineração de dados.

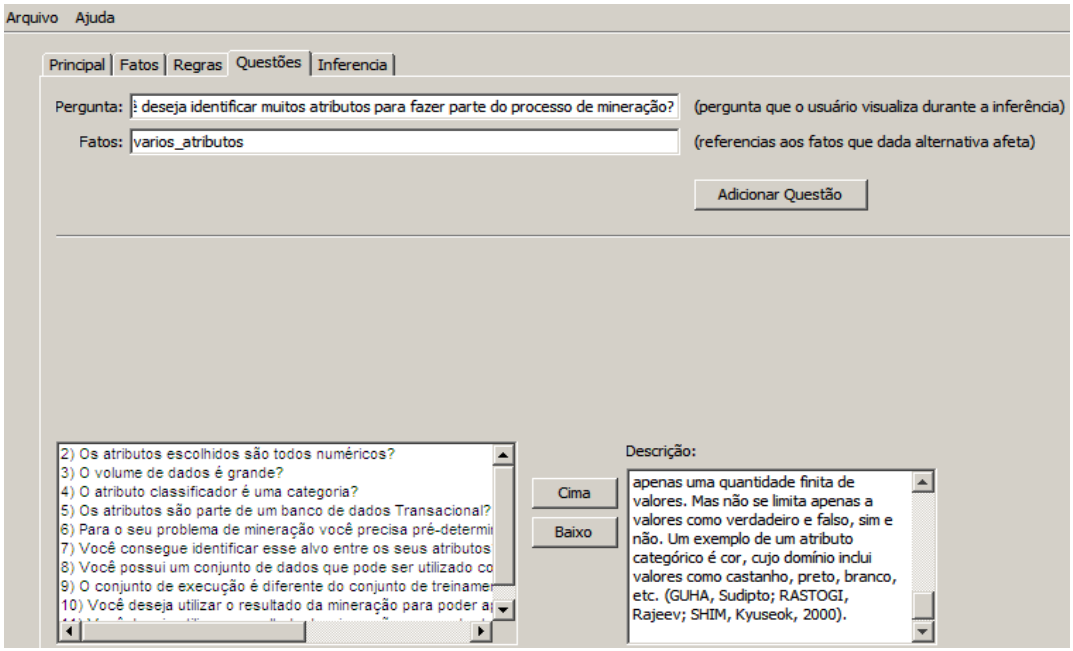

FIGURA 19 – IMPLEMENTAÇÃO DAS PERGUNTAS DE DIRECIONAMENTO

Ao executar as perguntas e o usuário tiver alguma dúvida, basta clicar no botão ajuda e ler as instruções de acordo com a figura 20.

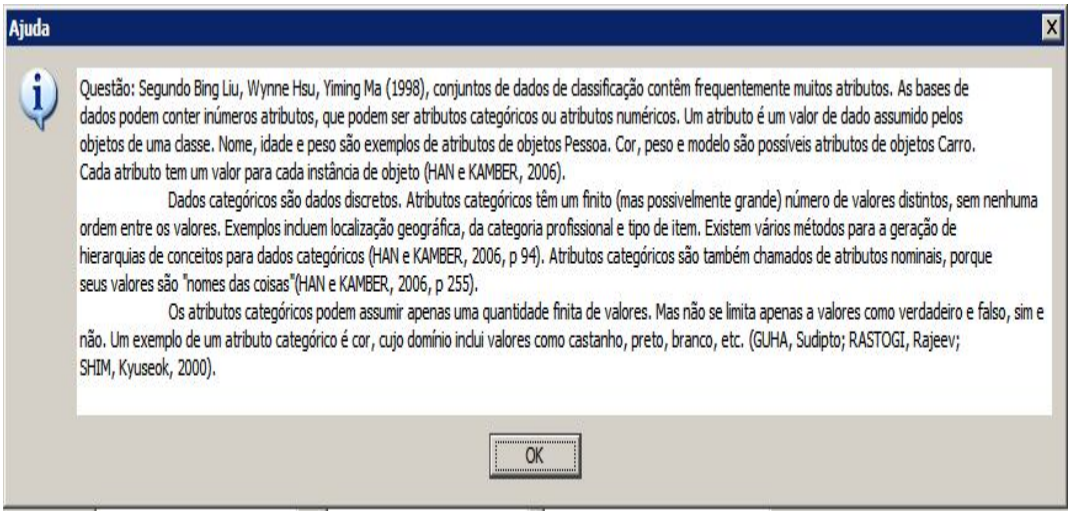

FIGURA 20 – TELA DE INSTRUÇÃO

Estas instruções são dicas que poderão orientar os usuários no momento em que estiverem respondendo as perguntas oferecidas pelo sistema com informações fundamentadas em pesquisas bibliográficas.

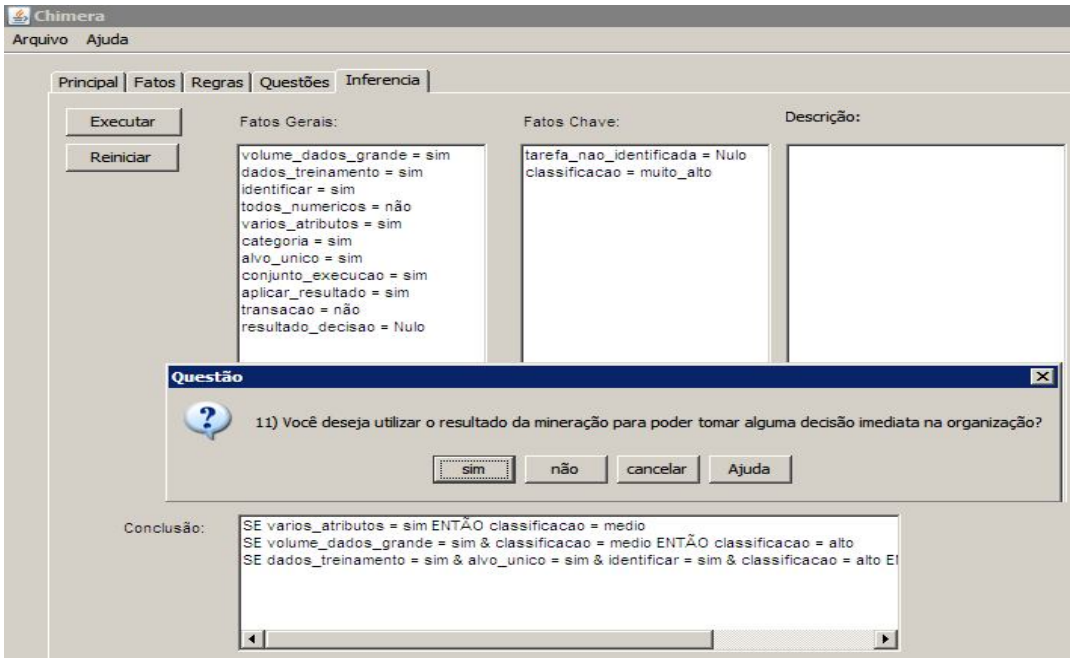

FIGURA 21 - TELA DE EXECUÇÃO E APRESENTAÇÃO DOS RESULTADOS

Após a execução do conjunto de perguntas existentes, o sistema faz o processamento das informações e em seguida exibe o resultado para a consulta realizada, conforme pode ser observado na figura 21.

E por fim, após a implementação da base de conhecimento, os testes foram realizados na ferramenta Chimera. Paralelamente aos testes elaborados com o uso da ferramenta, também foram realizados testes manualmente, conforme tabela 3 a seguir.

TABELA 3: TESTE NA BASE DE CONHECIMENTO

COD	VARIOS ATRIBUTOS	TODOS NÚMERICOS	VOLUME DADOS GRANDE	CATEGORIA	TRANSAÇÃO	ALVO UNICO	IDENTIFICAR	DADOS TREINAMENTO	CONJUNTO EXECUÇÃO	APLICAR RESULTADO	RESULTADO DECISÃO	RESPOSTA OBTIDA	RESPOSTA ESPERADA
1	SIM	NÃO	SIM	SIM	NÃO	SIM	SIM	SIM	SIM	SIM	SIM	MUITO ALTO	MUITO ALTO
2	SIM	NÃO	NÃO	SIM	NÃO	SIM	SIM	NÃO	NÃO	SIM	NÃO	ALTO	ALTO
3	SIM	SIM	SIM	SIM	SIM	SIM	SIM	SIM	SIM	SIM	SIM	MUITO ALTO	MUITO ALTO
4	SIM	SIM	SIM	NÃO	SIM	NÃO	SIM	NÃO	NÃO	NÃO	SIM	BAIXO	BAIXO
5	SIM	SIM	SIM	SIM	SIM	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO_É_ POSSIVEL	NÃO_É_ POSSIVEL
6	NÃO	NÃO	NÃO	NÃO	NÃO	SIM	SIM	SIM	SIM	SIM	SIM	BAIXO	BAIXO
7	SIM	SIM	SIM	NÃO	NÃO	SIM	SIM	SIM	NÃO	NÃO	NÃO	BAIXO	BAIXO
8	SIM	NÃO	SIM	NÃO	SIM	NÃO	NÃO	SIM	SIM	NÃO	NÃO	NÃO_É_ POSSIVEL	NÃO_É_ POSSIVEL
9	NÃO	SIM	NÃO	SIM	SIM	NÃO	SIM	NÃO	NÃO	SIM	SIM	BAIXO	BAIXO
10	NÃO	NÃO	SIM	NÃO	NÃO	SIM	NÃO	SIM	NÃO	NÃO	NÃO	NÃO_É_ POSSIVEL	NÃO_É_ POSSIVEL
11	SIM	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	MÉDIO	MÉDIO
12	NÃO	SIM	SIM	SIM	SIM	SIM	SIM	SIM	SIM	SIM	SIM	BAIXO	BAIXO
13	SIM	NÃO	SIM	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO_É_ POSSIVEL	NÃO_É_ POSSIVEL
14	NÃO	SIM	NÃO	SIM	SIM	SIM	SIM	SIM	SIM	SIM	NÃO	MÉDIO	BAIXO
15	NÃO	SIM	SIM	NÃO	NÃO	SIM	NÃO	NÃO	SIM	SIM	NÃO	NÃO_É_ POSSIVEL	NÃO_É_ POSSIVEL
16	SIM	NÃO	SIM	NÃO	NÃO	SIM	SIM	SIM	NÃO	NÃO	NÃO	NÃO_É_ POSSIVEL	NÃO_É_ POSSIVEL
17	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	SIM	SIM	SIM	SIM	NÃO_É_ POSSIVEL	NÃO_É_ POSSIVEL
18	SIM	SIM	SIM	SIM	SIM	SIM	SIM	NÃO	NÃO	NÃO	NÃO	MÉDIO	MÉDIO
19	SIM	NÃO	SIM	SIM	SIM	SIM	SIM	SIM	SIM	SIM	SIM	MUITO ALTO	MUITO ALTO
20	NÃO	SIM	SIM	SIM	SIM	SIM	SIM	SIM	SIM	SIM	NÃO	MÉDIO	MÉDIO
21	SIM	NÃO	SIM	SIM	SIM	NÃO	SIM	SIM	SIM	SIM	NÃO	BAIXO	BAIXO
22	NÃO	NÃO	NÃO	NÃO	NÃO	SIM	SIM	NÃO	NÃO	NÃO	NÃO	BAIXO	BAIXO
23	SIM	SIM	SIM	SIM	SIM	NÃO	NÃO	SIM	SIM	SIM	SIM	NÃO_É_ POSSIVEL	NÃO_É_ POSSIVEL

Conforme demonstrado na tabela 2, apenas o predicado classificação pode assumir um valor linguístico e ao fazê-lo apresenta níveis de diagnósticos diferentes para a tarefa de classificação.

A base de conhecimento de instrução foi criada apenas para a tarefa de classificação com os seguintes níveis de adequação: baixo, médio, alto, muito alto e não é possível. Estes níveis foram escolhidos para demonstrar as diferentes possibilidades de aplicação para a tarefa de classificação uma vez que existem vários parâmetros a serem analisados e uma resposta binária, como adequado (não adequado) restringiria o raciocínio aplicado à base de conhecimento.

A tabela 3 mostra alguns testes realizados. A escolha dos testes foi realizada observando as perguntas e as combinações de respostas mais óbvias, ou seja, testes que poderiam resultar em respostas como nível de adequação para tarefa de classificação muito alto, alto, médio e baixo. Os resultados esperados foram calculados anteriormente - de acordo com a literatura apresentada - e, em seguida, comparados aos resultados obtidos. Os testes foram realizados na ferramenta chimera (BINDILATTI, 2009) e para fins de amostras são apresentados a seguir quatro exemplos de respostas obtidas:

Ao responder as perguntas conforme a primeira linha da tabela 3, o resultado esperado é um nível de classificação muito_alto. Os atributos escolhidos com a resposta sim foram:

```
Varios_atributos=sim,
Volume_dados_grande=sim,
Categoria=sim,
Alvo_unico=sim,
Identificar=sim,
Dados_treinamento=sim,
Conjunto_execucao=sim,
Aplicar_resultado=sim,
Resultado_decisao=sim.
```

De acordo com as respostas das perguntas, foi aplicada na base de conhecimento a sequência do encadeamento para a frente, em que o

algoritmo é aplicado da seguinte forma: em primeiro lugar ele obtém os valores da parte Antecedente de uma Regra e se os valores são verdadeiros então os valores da parte Consequente são executados. Esse consequente torna-se um predicado verdadeiro e a parte Antecedente das regras devem ser testadas novamente, considerando esse novo predicado como verdadeiro. Esses passos são repetidos até que todas as regras tenham sido testadas.

Das 26 regras testadas, três regras foram verificadas com sucesso na base de conhecimento, são elas:

R01. **Se** vários_atributos = sim
Então classificação=médio
R05. **Se** Volume_Dados_Grande = sim e classificação = médio
Então classificação= alto
R19. **Se** dados_treinamento=sim e alvo_unico=sim e
Identificar=sim e classificação = alto
Então classificação = muito_alto

Dessa maneira, para esse exemplo, a tarefa de classificação poderá ser aplicada pelo usuário de acordo com o sistema especialista com nível de adequação muito_alto. O resultado faz sentido, já que as características principais para a aplicação da tarefa de classificação são apresentadas no problema do domínio do usuário.

Outro exemplo pode ser aplicado utilizando as respostas na quarta linha da tabela 3. O resultado esperado é de um nível de classificação baixo. Os valores dos predicados para esse exemplo são dados pelos valores a seguir.

Vários_atributos=sim,
Todos_numericos=sim,
Volume_dados_grande=sim,
Categoria=não,
Transação=sim,
Alvo_unico=não,
Identificar=sim,
Dados_treinamento=não,
Conjunto_execução=não,
Aplicar_resultado=não,
Resultado_decisão=sim.

Com a aplicação da inferência para à frente, apenas uma regra foi verificada com sucesso, conferindo uma classificação com um nível baixo de adequação a esta tarefa.

R01. **Se** vários atributos=sim
Então Classificação=médio
R03. **Se** todos_numericos= sim e Classificação= médio
Então Classificação =baixo
R09. **Se** transação=sim
Então classificação=baixo

Na linha de número 20 da tabela 02, resultado esperado é um nível de adequação da tarefa de classificação médio, conforme respostas a seguir:

Vários_atributos=sim,
Volume_dados_grande=não,
Todos_numericos=não,
Categoria=não,
Transação=não,
Volume_transação_alto=não,
Alvo_unico=não,
Identificar=não,
Dados_treinamento=não,
Conjunto_execução=não,
Aplicar_resultado=não,
Resultado_decisão=sim.

Após a aplicação da inferência para frente, as regras verificadas com sucessos conferindo um nível de adequação médio para a tarefa de classificação foram:

R01. Se vários_atributos = sim
Então classificação=médio

Na linha 02 da tabela 3, esperava-se obter um nível de classificação alto com as seguintes respostas:

Vários_atributos=sim,
Todos_numericos=não,
Volume_dados_grande=sim,
Categoria=sim,
Transação=não,
Alvo_unico=não,
Identificar=sim,
Dados_treinamento=não,
Conjunto_execução=não,
Aplicar_resultado=não,
Resultado_decisão=não.

Quando aplicada à inferência para frente obteve-se um nível de adequação para a tarefa de classificação como alto, em que as seguintes regras foram verificadas com sucesso:

R01. **Se** vários_atributos = sim
Então classificação=médio
R07. **Se** categoria=sim e classificação = médio
Então classificação=alto

Na linha 14 da Tabela 3, O resultado esperado seria um nível baixo de adequação, o nível de adequação esperada foi diferente do obtido, porque os predicados que receberam um valor não (vários_atributos, volume_dados_grande e aplicar_resultado) não proibe a aplicação da classificação porque todos os outros predicados receberam um valor sim. O nível de adequação para a tarefa de classificação obtido foi médio. Isto significa que a base de conhecimentos está sendo severa com alguns dos predicados.

Os testes permitiram descobrir vários níveis de adequação da tarefa de classificação de acordo com diferentes respostas dadas às perguntas. Para a realização de todos os testes, fazendo a combinação das 11 variáveis da base de conhecimento com a possibilidade de resposta de cada variável, (sim e não) há 2.048 possibilidades de respostas. Esse resultado equivale a duas possibilidades de resposta das variáveis - Sim e Não- elevados a doze – número de variáveis da base de conhecimento (2^{11}).

Dessa forma, para um primeiro nível de teste, apenas os valores mais óbvios de respostas às perguntas foram escolhidos. Para esse conjunto de respostas, os testes mostraram que o resultado obtido estava de acordo com o resultado esperado.

5 CONCLUSÃO

Este trabalho nos permitiu realizar a modelagem e teste de uma base de conhecimento de instrução para a tarefa de classificação com a finalidade de instruir o usuário através de perguntas que o direcionem a descobrir se a tarefa de classificação é adequada para ser utilizada em seu domínio de problema. Um dos objetivos também foi o de colaborar com a ferramenta Kira, no sentido de proporcionar uma representação do conhecimento necessário para o entendimento e escolha da aplicação da tarefa de classificação em problemas de mineração de dados.

A base de conhecimento foi criada tendo 11 (onze) perguntas sobre a tarefa de classificação, com comentários/dicas que poderão nortear os usuários na ocasião em que estiverem respondendo às perguntas apresentadas pelo sistema com informações baseadas em pesquisas bibliográficas. Assim, o usuário do sistema pode ver que as dicas são baseadas na literatura sobre mineração de dados, especificamente, sobre classificação.

Para validação das regras de produção criadas foram realizados testes utilizando a ferramenta Chimera onde, os testes permitiram verificar a funcionalidade dessa ferramenta e testar a base de conhecimento.

5.1 CONTRIBUIÇÃO

As contribuições deste trabalho são:

- uma rede semântica para representação do conhecimento da tarefa de classificação;

- onze perguntas de direcionamento da tarefa de classificação;

- 25 Regras de Produção para representação e execução do conhecimento relativo à escolha da tarefa de classificação como tarefa de mineração de dados para determinado domínio;

5.2 TRABALHOS FUTUROS

Como trabalho futuro fica a sugestão de continuidade deste trabalho. Aqui foram realizados a modelagem e teste de uma base de conhecimento de instrução para mineração de dados relacionais utilizando a tarefa de classificação. A continuidade seria:

- ➔ realizar mais testes com a base para validar o conhecimento representado pelas regras;
- ➔ utilizar alguns banco de dados de aplicações (reais) para testar a base de conhecimento.
- ➔ fazer uma base de conhecimento que contemple as tarefas de associação e agrupamento;
- ➔ implementar um sistema especialista de instrução com todas as tarefas citadas.

REFERÊNCIAS BIBLIOGRÁFICAS

ADRIAANS, P., ZANTINGE, D. **Data Mining**. Addison-Wesley, 1997.

AGRAWAL, R.; SRIKANT, R. **Fast Algorithms for Mining Association Rules.** In: Proc. of the Int'l Conf. on Very Large Databases, Santiago de Chile, Chile, 1994.

_____. **Mining Sequencial Patterns.** In: Eleventh International Conference on Data Engineering, 1995, Taipei, Taiwan. Anais. Taipei, Taiwan, 1995, p. 3 -14.

ALECRIM, E. **Redes Neurais Artificiais.** Publicado em maio 2004. Disponível em: <http://www.infowester.com/redesneurais.php>. Acesso em: 20 Out. 2009.

ALVAREZ, R. **Discourse Analysis of Requirements and Knowledge Elicitation Interviews.** University of Massachusetts. In: IEEE. Proceedings of the 35th Hawaii International Conference on System Sciences - 2002

AMO, S. **Curso de Data Mining**. Programa de Mestrado em Ciência da Computação, Universidade Federal de Uberlândia, 2003. Disponível <http://www.deamo.prof.ufu.br/CursoDM.html>. Acesso em: 01 Out. 2009.

BARR, A; FEIGENBAUM, E. **The Handbook of Artificial Intelligence**. Los Altos, California: William Kaufmann Inc., 1981. v.I –II.

BARRETO, J. M. **Inteligência artificial no Limiar do Século XXI.** 3 ed. Florianópolis: Duplic, 2001.

BERRY, MICHAELl J. A.; LINOFF, G. **Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management.** 2 ed. Wiley Publishing. USA, 2004.

BEYON-DAVIS, P. **Expert database systems – agente introduction.** London: Mc GrawHill, 1991.

BINDILATTI, A. A. **Chimera**. Unimep: Universidade Metodista de Piracicaba. Piracicaba, São Paulo, 2009. Relatório Técnico.

BITTENCOURT, G. **Inteligência artificial: ferramentas e Teorias.** Florianópolis: Editora da UFSC, 1998.

BUCHANAN, B., D. BARSTOW & R. BECHTEL. Building Expert Systems, Capítulo **constructing an Expert System**, pp. 127-169. Reading, MA: Addison-Wesley, 1983.

CARVALHO, L. A. V. **Data Mining – A Mineração de Dados no Marketing, Medicina, Economia, Engenharia e Administração.** 2005.

CARVALHO, D. R. **Data Mining através de indução de Regras e Algoritmos Genéticos.** Dissertação de Mestrado em Informática Aplicada, PUCPR, PR, 1999.

CHAPMAN, P.; *et al.* **CRISP-DM 1.0. Step-by-step data mining guide**.[S.l] 2000. Disponível em: http://www.spss.com/media/collateral/CRISP-DM_1.0__Step-by-Step_Data_Mining_Guide.pdf. Acesso em: 05 Mar. 2009.

CHECLAND, P. B. **Systems Thinking, Systems Practice**. Chichester, England: John Wiley, 1981.

CRISP-DM Consortium. **Process Model.** Disponível em: <http://www.crisp-dm.org/Process/index.htm>. Acesso em: 05 Mar 2009.

CRISP-DM Consortium. **CRISP-DM – Cross Industry Standard Process for Data Mining.** Disponível no site CRISP-DM (2000). URL: http://www.crispdm.org/. Acesso em: 05 Mar. 2009.

DAVIS, R., H. SHROBE, & P. SZOLOVITS. **What is a knowledge representation?** AI Magazine, 1993.

DAYHOFF, J. **Neural Network Architectures: A Introductin**. Van Nostrand Reinhold, New York, NY: 1990.

ELMASRI, R. & NAVATHE, S.B. **Sistemas de Banco de Dados**. 4 ed. Pearson Brasil, 2005.

FAYYAD,U.M.,G.Piatetsky–Shapiro, P.Smyth. **Knowledge Discovery and Data Mining:Towards a Unifying Framework.** Proceeding of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, august, 1996.

FEIGENBAUM, E.; BUCHANAN, B.; LEDERBERG, J. On generality and problem solving: a case study using the dendral program. In: MACHINE INTELLIGENCE, 1971, Edinburgh, GB. **Anais**. . . Edinburgh University Press, 1971. v.6, p.165.190.

FEIGENBAUM, E. A.; McCORDUCK, P. **The Fifth Generation: Artificial Intelligence and Japan´s Computer Challenge to the World**. Michel Joseph ed., 1983.

FERREIRA, J. B. **Mineração de dados na Retenção de clientes em Telefonia celular**. 2005. P. 93. Dissertação (Mestrado em Engenharia Elétrica) – Faculdade de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2005. Disponível em: <http://www2.dbd.puc-rio.br/pergamum/tesesabertas/0310411_05_pretextual.pdf>. Acesso em: 09 Dez. 2009.

FERREIRA, M. C. L. **Análise do Discurso no Brasil: notas à sua história. (**20/06/2006) Disponível em: <http://www.discurso.ufrgs.br/>. Acesso em: 12 Mai. 2009.

FLORES, C. D. Fundamentos dos Sistemas especialistas. In: BARONE, D.A. C. (Ed.). **Sociedades Artificiais: a nova fronteira da inteligência nas máquinas.** Porto Alegre: Bookman, 2003. p.332.

FREITAS, A. A., Lavington, S. H. **Mining Very Large Databases With Parallel Proccessing.** Kluwer Academic Publishers. 1998.

GIARRATANO, J.; RILEY, G. **Expert systems: principles and programming.** Third edition. PWS Publishing Company. 1998.

_____. **Expert systems: principles and programming**. 2 ed. Boston: PWS Publishing, 1994.

GUHA, Sudipto; RASTOGI, Rajeev; SHIM, Kyuseok. **ROCK: A robust clustering algorithm for categorical attributes.** Information Systems. Stanford University. Stanford, CA 94305, 25(5), 2000.

HARMON, Paul. l; KING, David. **Sistemas especialistas: A inteligência artificial chega ao mercado.** Campus, Rio de Janeiro, 1988.

HAYKIN, S. **Neural Networks: A comprehensive Foundation**. Macmillam College Publishing Company, New York, NY, 1994.

HAN, J.; KAMBER, M. **Data Mining - Concepts and Techniques.** 2 ed. Nova York: Morgan Kaufmann, 2006.

_____. **Data Mining - Concepts and Techniques.** San Francisco, EUA: Morgan Kaufmann, 2001. 550 p.

HARRISON, Thomas H. **Intranet Data Warehouse.** São Paulo, Berkeley Brasil, 1998.

HEISSERMAN, J., S. CALLAHAN, & R. MATTIKALI. **A design representation to support automated design generation.** Em Proceedings of the Sixth International Conference on Artificial Intelligence in Design, pp. 545-566.

KELLY, S. **Data Warehouse applications in the telecommunications industry.** Proc. Conf. Commercial Parallel Processing. London, IBC, 1995.

KLÕSGEN, W. (Org). **Handbook of Data Mining and Knowledge discovery.** New York, EUA: Oxford University Press, 2002.

KUIPERS, B.; KASSIRER, J. P. **Knowledge Acquisition by Analysis of Verbatim Protocols.** This article appeared in A. Kidd (Ed.), Knowledge Acquisition for Expert Systems. New York: Plenum, 1987. Cognitive Science 8: 363-385, 1984.

LAROSE, D. T. **Discovering Knowledge: introduction to data mining.** Printed in the United States Of America. 2005.

LIEBOWITZ, J. **The Handbook of Applied Expert Systems**. CRC Press, 1999.

LYNCH Collin, ASHLEY Kevin D., PINKWART Niels, ALEVEN Vincent. **Argument Graph Classification via Genetic Programming and C4.5.** Intelligent Systems Program, University of Pittsburgh. This research is supported by NSF Award IIS-0412830. Disponível em: <http://www.pdf-search-engine.com/c4-5-ross-quinlan-pdf.html>. Acesso em: 10 Out. 2009.

LUGER, G. F. **Artificial Intelligence Structures and Strategies for Complex Problem Solving. Fifth Edition.** England. Addison-Wesley. 2005

MARTINS, A. C.; COSTA, P. D. C. **Estudo Comparativo de três Algoritmos de Machine Learning na Classificação de Dados Electrocardiográficos.**

Dissertação de Mestrado. Faculdade de Medicina da Universidade do Porto. Mestrado em Informática Médica. Porto, 2009.

MENDES, E. F. **Automatização da Técnica de Mineração de Dados Auxiliada por Guias.** Dissertação de Mestrado. Universidade Metodista de Piracicaba. Piracicaba. São Paulo, 2009.

MENDONÇA NETO, M. G. **Mineração de Dados.** In: 6º Escola Regional de Informática de São Paulo, São Carlos, 2001.

MINSKY, M. A Framework for Representing Knowledge In: **Readings In Knowledge Representation.** USA: Morgan Kaufmann Publishers, Inc.1985.

MITTCHEL, Tom M. **Machine Learning**. McGraw-Hill Science/Engineering/Math; March 1, 1997.

MOTTA, C. G. L. **Sistema Inteligente para Avaliação de Riscos em Vias de Transporte Terrestre.** Tese de Mestrado. Universidade Federal do Rio de Janeiro, Rio de Janeiro (2004). Disponível em: <http://www.coc.ufrj.br/teses/mestrado/inter/ 2004/Teses/MOTTA_CGL_04_t_M_int.pdf>. Acesso em: 20 Out. 2009.

NIKOLOPOULOS, C. **Expert systems: introduction to first and second generation and hybrid knowledge based systems.** New York: Marcel Dekker, Marcel Dekker Inc, 1997. 331 p.

NILSSON, J. **Artificial intelligence: a new synthesis.** San Francisco, CA: Morgan Kaufmann, 1998. 513 p.

NILSON, Neils S. **Principles of Artificial Intelligence**. Springer Verlag, Berlin, 1982.

OLSON, D.L., DELLEN, D. **Advanced Data Mining Techniques.** Springer. Verlag – Berlin – Heidelberg. 2008.

PASSOS, E. L. **Inteligência artificial e sistemas especialistas ao Alcance de Todos**. Rio de Janeiro, Soc. Ben. Guilherme Guinle, 1989.

PIATETSKY-SHAPIRO, G. **Knowlwdge Discovery in real database: A reporto the IJCAI-89 Workshop.** AI Magazine, Vol. 11, nº 5, Jan. 1991, Special Issue, 68-70.

PYLE, D. **Data Preparation for Data Mining.** Morgan Kaufmann, Califórnia, USA. 1999.

REZENDE, S. O.; et al. Mineração de Dados. In Rezende, S. O. et al: **Sistemas Inteligentes – Fundamentos e Aplicações.** Barueri, SP: Manole, 2003.

RICH, E.; KNIGHT, K.. **Artificial Intelligence**. 2. ed. New York McGraw-Hill, New York. 1993.

RODEN, J.; BURL, M. C.; FOWLKES, C;. *et al*. **Mining for image Content.** In Systemics, 1999.

ROIGER R. J., GEATZ M.W., **Data Mining: A Tutorial-Based Primer.** New York, Addison Wesley, 2003.

ROVER, Aires J. **Representação do conhecimento através de sistemas especialistas.** Disponível em: <http://www.infojur.ufsc.br/aires/arquivos/representacao%20do%20conhecimento%20sistemas%20especialistas.pdf>. Acesso em: 02 Jan 2010. Material Didático.

RUSSELL, S. J.; NORVIG, P. **Artificial Intelligence, A Modern Approach.** 2nd ed. Upper Saddle River, New Jersey, USA: Pearson Education, 2003.

QUINLAN J. Ross. **C4.5: Programs for Machine Learning.** Morgan Kaufmann Publishers; San Mateo, 1993.

SAVARIS, Silvana V. A. Michelotto. **Sistema especialista para primeiros socorros para cães**. 2002. 156 f. Tese (Pós-Graduação em Ciência da Computação) – Universidade Federal de Santa Catarina, Florianópolis.

SILVA, A. E. A. Inteligência artificial II – **Regime Especial: Introdução a Sistemas Especialistas.** Disponível em:< www.unimep.br/~aeasilva/regesp.pdf>. Acesso em: 31 Mar. 2009. Material Didático.

SOWA, J. F. **Semantic networks.** 2002. Disponível em: <http://www.jfsowa.com/pubs/semnet.htm>. Acesso em: 05 Mar. 2009.

TWO CROWS CORPORATION. **Introduction to Data Mining and Knowledge Discovery.** Third Edition. USA. 1999.

VIEIRA, M. T. P.; SILVA, A. E. A.; PEIXOTO, C. S. A..; GOMIDE, R. S.; MENDES, E. F. **KIRA - a tool based on guides and domain knowledge to instruct data mining.** Proceedings Of The IADIS International Conference Applied Computing. pp 12-16. 2009 ISBN: 978-972-8924-97-3. Vol II. Rome, Italy, 2009.

WATERMAN, D. A**. A Guide to Expert Systems**. Canada: Addison-Wesley, 1986.

WEKA. Disponível em: http://www.cs.waikato.ac.nz/ml/weka/. Acesso em: 20 Dez. 2009.

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

APÊNDICE A: ANÁLISE DE DISCURSO

Guia 01: Auxilia o analista de dados a definir o problema de mineração de dados.

Nº	Conhecimento
01	É necessário que o analista de dados conheça o ambiente de negócio em que será aplicado à mineração de dados.
02	É necessário que o analista de dados conheça a etapa do processo de KDD responsável pela definição do problema.
03	É necessário que o analista de dados saiba definir um problema de mineração de dados.

TABELA 5.1 – CONHECIMENTOS DA ETAPA DE DEFINIÇÃO DO PROBLEMA

Guia 02: Auxilia o analista de dados a definir o objetivo que deseja cumprir para resolver o problema de mineração de dados.

Nº	Conhecimento
01	É necessário que o analista de dados saiba em qual etapa do processo de KDD é necessário definir o objetivo de mineração de dados.
02	É necessário que o analista de dados tenha conhecimento sobre o que é um objetivo de mineração de dados.
03	É necessário que o analista de dados saiba definir corretamente um objetivo de mineração de dados.

TABELA 5.2 – CONHECIMENTOS DA ETAPA DE DEFINIÇÃO DO OBJETIVO

Guia 03: Auxilia o analista de dados na identificação da tarefa de mineração que será usada para resolver o problema e objetivo de mineração.

Nº	Conhecimento
01	É necessário que o analista de dados saiba em qual etapa do processo de KDD é necessário selecionar a tarefa de mineração de dados.
02	É necessário que o analista de dados tenha conhecimento sobre as tarefas de mineração de dados.
03	É necessário que o analista de dados saiba associar qual tarefa de mineração é mais apropriada ao problema e objetivo de mineração de dados.

TABELA 5.3 – CONHECIMENTOS DA ETAPA DE DEFINIÇÃO DO OBJETIVO

Guia 04: Auxilia o analista de dados a identificar e selecionar os dados relevantes ao problema.

Nº	Conhecimento
01	É necessário que o analista de dados conheça a etapa do processo de KDD responsável pela seleção dos dados.
02	É necessário que o analista de dados tenha entendimento sobre os dados armazenados, ou seja, tenha conhecimento sobre os dados relevantes ao problema de mineração.
03	É necessário que o analista de dados saiba quais colunas são mais apropriadas para determinadas tarefas de mineração
04	É necessário que o analista de dados domine as regras do modelo relacional, ou seja, tenha conhecimento de como as tabelas estão relacionadas umas as outras
05	É necessário que o analista de dados saiba executar instruções SQL, ou seja, após a seleção dos dados, é necessária a criação de uma tabela com os dados selecionados para utilização nas etapas seguintes.

TABELA 5.4 – CONHECIMENTOS DA ETAPA DE SELEÇÃO DOS DADOS

Guia 05: Auxilia o analista de dados a avaliar individualmente cada coluna da tabela e sugere o que deve ser feito para deixar a coluna apta para a mineração.

Nº	Conhecimento
01	É necessário que o analista de dados conheça a etapa do processo de KDD responsável pela transformação dos dados.
02	É necessário ter conhecimento sobre o formato exigido dos dados para a tarefa de mineração.
03	É necessário que o analista de dados tenha conhecimento sobre as técnicas utilizadas para preparação dos atributos.

TABELA 5.5 – CONHECIMENTOS DA ETAPA DE TRANSFORMAÇÃO

Guia 06: Auxilia o analista de dados na parametrização do algoritmo que será usado para a execução da mineração de dados.

Nº	Conhecimento
01	É necessário que o analista de dados conheça a etapa do processo de KDD responsável pela mineração de dados.
02	É necessário que o analista de dados tenha conhecimento sobre os algoritmos utilizados dentro de cada tarefa de mineração de dados.
03	É necessário que o analista de dados tenha conhecimento sobre como escolher o melhor algoritmo baseado no problema de mineração.
04	É necessário que o analista de dados tenha conhecimento sobre como parametrizar corretamente o algoritmo de mineração.

TABELA 5.6 – CONHECIMENTOS DA ETAPA DE MINERAÇÃO DE DADOS

Guia 07: Auxilia o analista de dados a ler, interpretar e avaliar a qualidade das regras geradas.

Nº	Conhecimento
01	É necessário que o analista de dados conheça a etapa do processo de KDD responsável pela avaliação das regras.
02	É necessário que o analista de dados tenha conhecimento sobre o problema e objetivo que a mineração tem que resolver.
03	É necessário que o analista de dados tenha conhecimento de como ler e interpretar corretamente os valores das regras geradas.
04	É necessário que o analista de dados tenha conhecimento sobre como avaliar a qualidade das regras geradas.

TABELA 5.7 – CONHECIMENTOS DA ETAPA DE AVALIAÇÃO

APÊNDICE B – ENTREVISTA AO ESPECIALISTA

1 - Quais as dificuldades encontradas na utilização dos softwares para mineração de dados?
R: Softwares sem instrução de uso ou com poucas informações.

2 – O que é entendimento de negócios?
R: É a etapa que incide sobre a compreensão dos objetivos do projeto e os requisitos a partir de uma perspectiva empresarial.

3 - O que é um problema de mineração de dados?
R: É o que se deseja resolver com a mineração de dados

4 - O que é preciso saber ou ter conhecimento para definir o problema?
R: É preciso ter conhecimento do ambiente de negócios e o que se deseja resolver com a mineração de dados.

5 – Quais os conhecimentos exigidos do analista de dados para a definição do problema de mineração de dados?
R: É necessário que o analista de dados conheça a etapa do processo de KDD responsável pela definição do problema e saiba definir um problema de mineração de dados.

6 - Na resolução do problema que objetivos são necessários cumprir?
R: O objetivo da aplicação da mineração de dados compreende as características esperadas do modelo de conhecimento a ser produzido no final do processo. Tais objetivos retratam, portanto, restrições e expectativas dos analistas de dados no domínio da aplicação acerca do modelo de conhecimento a ser gerado. Definir o objetivo que deseja cumprir para a resolução do problema de mineração de dados é fundamental para o sucesso da mineração.

7- Como selecionar corretamente a tarefa de mineração de dados?
R: A escolha da tarefa depende do problema de mineração. Se o analista quer encontrar itens com uma grande probabilidade de acontecerem juntos em uma mesma transação, então a tarefa será associação. Se ele deseja classificar um determinado objeto em uma entre várias possíveis classes, então deverá escolher classificação. Mas se ele deseja a organização de uma coleção de objetos em grupos de objetos similares, então a tarefa será agrupamento.

8 - Quais são as principais tarefas de mineração de dados?
R: Associação, Classificação, agrupamento.

9 - Que conhecimentos são exigidos do analista de dados para a identificação da tarefa de mineração?
R: Que o analista de dados saiba em qual etapa do processo de KDD é necessário selecionar a tarefa de mineração de dados, tenha conhecimento sobre as tarefas de mineração de dados e saiba associar qual tarefa de mineração é mais apropriada ao problema e objetivo de mineração de dados.

10 – Porque a escolha da tarefa é fato primordial na mineração de dados?
R: por que a escolha irá afetar diretamente as etapas seguintes e, conseqüentemente, o resultado das regras geradas.

11 – Quais os conhecimentos exigidos do analista de dados para a execução da etapa de seleção dos dados?
R: Que o analista de dados conheça a etapa do processo de KDD responsável pela seleção dos dados, tenha entendimento sobre os dados armazenados, ou seja, tenha conhecimento sobre os dados relevantes ao problema de mineração, saiba quais colunas são mais apropriadas para determinadas tarefas de mineração, domine as regras do modelo relacional, ou seja, tenha conhecimento de como as tabelas estão relacionadas umas as outras e saiba executar instruções SQL, ou seja, após a seleção dos dados, é necessária a criação de uma tabela com os dados selecionados para utilização nas etapas seguintes.

12 - Porque selecionar os dados e gerar uma única tabela?
R: porque a maioria dos algoritmos de mineração de dados trabalha com dados armazenados em uma única tabela.

13 - Que conhecimentos são exigidos para executar a etapa de transformação?
R: É necessário que o analista de dados conheça a etapa do processo de KDD responsável pela transformação dos dados, ter conhecimento sobre o formato exigido dos dados para a tarefa de mineração e tenha conhecimento sobre as técnicas utilizadas para preparação dos atributos.

14 - É preciso conhecer o formato dos dados para minerar?
R: Sim, pois os dados devem ser colocados em um formato que possam servir de entrada para um algoritmo de mineração.

15 – É necessário conhecer os algoritmos para mineração de dados?
R: Sim, ter conhecimento dos algoritmos disponíveis de cada tarefa de mineração, suas características e como parametrizar corretamente é fundamental para o sucesso da mineração de dados. Este problema pode facilmente aumentar o número de iterações do processo, na medida em que diversos algoritmos, com diferentes parametrizações sejam experimentados na busca por resultados promissores.

16 - Na mineração dos dados porque é necessário parametrizar os algoritmos?
R: Além de conhecer os algoritmos e associá-los ao problema, sua parametrização é um ponto crítico. Por exemplo, se um determinado algoritmo de regras de associação for parametrizado de forma errada, nenhuma regra poderá ser gerada.

17 - Como parametrizar um algoritmo corretamente?
R: Para parametrizar o algoritmo de forma correta, é necessário que o analista de dados tenha conhecimento sobre todos os parâmetros disponíveis para cada algoritmo, quais as características provocadas por cada valor de parâmetro para a execução do algoritmo e quais os valores que podem ser utilizados para cada parâmetro

18 - Quais conhecimentos são exigidos para a execução da etapa de avaliação?
R: É necessário que o analista de dados conheça a etapa do processo de KDD responsável pela avaliação das regras, tenha conhecimento sobre o problema e objetivo que a mineração tem que resolver, tenha conhecimento de como ler e interpretar corretamente os valores das regras geradas e tenha conhecimento sobre como avaliar a qualidade das regras geradas.

APÊNDICE C: DICIONÁRIO DE CONHECIMENTO

Os conceitos apresentados a seguir foram retirados durante a análise de discurso realizada nos guias da ferramenta Kira e nas entrevistas com especialistas nas áreas de mineração de dados e SE.

Tabela	é um conjunto de dados dispostos em número finito de colunas e número ilimitado de linhas (ou tuplas). As colunas são tipicamente consideradas os *campos* da tabela, e caracterizam os tipos de dados que deverão constar na tabela (numéricos, alfa-numéricos, datas, coordenadas, etc). O número de linhas pode ser interpretado como o número de combinações de valores dos campos da tabela, e pode conter linhas idênticas, dependendo do objetivo.
Base de Dados	são conjuntos de registros dispostos em estrutura regular que possibilita a reorganização dos mesmos e produção de informação. Um banco de dados normalmente agrupa registros utilizáveis para um mesmo fim.
Atributo	As colunas de uma tabela são também chamadas de Atributos. Ao conjunto de valores que um atributo pode assumir chama-se domínio. Por exemplo: em um campo do tipo numérico, serão somente armazenados números.
Tarefa de Mineração	Consiste na especificação do que estamos querendo buscar nos dados, que tipo de regularidades ou categoria de padrões temos interesse em encontrar, ou que tipo de padrões poderiam nos surpreender (por exemplo, um gasto exagerado de um cliente de cartão de crédito, fora dos padrões usuais de seus gastos.
Técnica de mineração	Consiste na especificação de métodos que nos garantam como descobrir os padrões que nos interessam. Dentre as principais técnicas utilizadas em mineração de dados, temos técnicas estatísticas, técnicas de aprendizado de máquina e técnicas baseadas em crescimento-poda-validação.
Classificação	Processo de encontrar um conjunto de modelos que descrevem e distinguem classes de dados ou conceitos com o propósito de utilizar o modelo para predizer a classe de objetos que ainda não foram classificados. Esses modelos são usados para predição de objetos cujas classes são desconhecidas, baseada na analise de um conjunto de dados de treinamento.
Tupla ou Registro	Cada linha formada por uma lista ordenada de colunas representa um registro, ou tupla. Os registros não precisam conter informações em todas as colunas, podendo assumir valores nulos quando assim se fizer

	necessário.Resumidamente, um registro é uma instância de uma tabela, ou entidade.
Usuário	Pessoa que executa todo o processo da mineração.
Entendimento do Negócio	Conhecimentos sobre os objetivos do negócio e seus requisitos. (Define o problema e o objetivo a serem alcançados).
Problema de mineração de dados	O que se deseja resolver com a mineração de dados.
Objetivo da tarefa de classificação	O objetivo dessa tarefa é descobrir um relacionamento entre os atributos previsores e o atributo meta, usando registros cuja classe é conhecida, para que posteriormente esses atributos previsores possam ser utilizados para prever a classe de um registro cuja classe é desconhecida.
Etapas da tarefa de classificação	A tarefa de classificação consiste em 2 etapas: 1 – Construção de um modelo descrevendo um conjunto de dados predeterminado, chamado de conjunto de dados de treinamento; 2 – Utilização do modelo criado para classificar novos dados.
Passos da tarefa de classificação	Entendimento do negócio, Identificação do problema e objetivo, Entendimento dos dados, avaliação e obtenção dos dados, preparação dos dados, transformação dos dados, preparação do conjunto de treinamento, escolher classificador, aplicar classificador e avaliar o resultado.
Conjunto de treinamento	Conjunto de dados predeterminado utilizado para o treinamento do algoritmo indutor.
Conjunto de execução	Dados que não pertencem ao conjunto de treinamento – geralmente coletados cronologicamente após os dados de treinamento.
Conjunto de testes	O conjunto de exemplos utilizado para testar a qualidade do algoritmo indutor.
Classificador	Algoritmo indutor de classificação baseado em um modelo de classificação a partir do conjunto de treinamento.
Objetivo principal do classificador	Gerar conhecimento a partir de um conjunto de dados passado como entrada. O conhecimento pode ser então usado para classificar novos dados.
Utilidade de um classificador	É útil para os seguintes propósitos: - análise descritiva: o classificador pode servir para explicar as características dos objetos de diferentes classes. Por exemplo, pode ser usado para informar as características dos clientes que são bons pagadores. Essas informações são obtidas apenas com a interpretação do modelo de classificação gerado através de exemplos conhecidos. - análise preditiva: o classificador é utilizado para

	classificar objetos que não foram utilizados na construção do modelo. Por exemplo, pode ser usado para predizer se um novo cliente vai ser um bom ou mau pagador
Regras de classificação	Uma regra de classificação é um par <condição➔classe>, onde condição é o conjunto de condições para que o problema seja classificado como um dos valores do conjunto de classes.
Atributo Classe	É o atributo objetivo do problema. E a característica que se quer classificar, ou, de outra forma, e a variável dependente dos outros atributos.
Árvore de decisão	É uma estrutura de árvore onde: (1) cada nó interno é um atributo do banco de dados de amostras, diferente do atributo-classe, (2) as folhas são valores do atributo-classe, (3) cada ramo ligando um nó-_lho a um nó-pai é etiquetado com um valor do atributo contido no nó-pai. Existem tantos ramos quantos valores possíveis para este atributo. (4) um atributo que aparece num nó não pode aparecer em seus nós descendentes.
Entropia	Grau de Pureza de um atributo num nó. Este grau de pureza representa a quantidade de informação esperada que seria necessária para especificar se uma nova instância seria classificada em 'Sim' ou 'Não', uma vez chegado a este nó.
Definição do problema	Essa etapa visa à assimilação dos objetivos do projeto a análise dos requisitos de negócio. Com base no conhecimento adquirido, o problema de mineração é definido e um plano preliminar deverá ser projeto para atingir os objetivos.
Limpeza dos dados	etapa onde são eliminados ruídos e dados inconsistentes.
Integração dos dados	etapa onde diferentes fontes de dados podem ser combinadas produzindo um único repositório de dados.
Seleção	Etapa onde são selecionados os atributos que interessam ao usuário. Por exemplo, o usuário pode decidir que informações como endereço e telefone não são de relevantes para decidir se um cliente é um bom comprador
Preparação dos dados	Operações básicas tais como remoção de ruídos quando necessário, coleta da informação necessária para modelar ou estimar ruído, escolha de estratégias para manipular campos de dados ausentes, formatação de dados de forma a adequá-los à ferramenta de mineração
Transformação	Os dados são modificados ou transformados em formatos apropriados à mineração.
Seleção	São selecionados os dados que serão utilizados no processo de mineração de dados
Mineração	É o processo de explorar grandes quantidades de dados à procura de padrões consistentes, como regras de

	associação, para detectar relacionamentos sistemáticos entre variáveis, detectando assim novos subconjuntos de dados.
Avaliação	Avaliação dos resultados obtidos pela etapa de mineração, após a aplicação do algoritmo.
Apresentação do conhecimento	Apresenta-se o conhecimento através de técnicas de visualização e representação do conhecimento.
Interpretabilidade	Refere-se ao nível de entendimento que o modelo fornece, isto é, o quanto as regras fornecidas são entendidas pelos usuários do classificador.
Escalabilidade	Refere-se à eficiência do processo de aprendizado (construção do modelo) em presença de grandes volumes de dados de treinamento.
Robustez	Refere-se a habilidade do modelo em fazer uma classificação correta mesmo em presença de ruídos ou valores desconhecidos em alguns campos dos registros.
Rapidez	Refere-se ao tempo gasto no processo de classificação.
O grau de acertos (accuracy)	Este critério refere-se a capacidade do modelo em classificar corretamente as tuplas do banco de dados de testes. É medida pela porcentagem de instâncias do conjunto de teste que foram corretamente classificadas pelo classificador, de acordo com a classificação inferida. (grau de acertos)
Sistema de Instrução	Tem um mecanismo para verificar e corrigir o comportamento do aprendizado dos estudantes. Normalmente, incorporam como subsistemas um sistema de diagnóstico e de reparo, e tomam por base uma descrição hipotética do conhecimento do aluno. Seu funcionamento consiste em ir interagindo com o treinando, em alguns casos apresentando uma pequena explicação e, a partir daí, ir sugerindo situações para serem analisadas pelo treinando. Dependendo do comportamento deste, se vai aumentando a complexidade das situações e encaminhando o assunto, de maneira didática, até o nível intelectual do treinamento.
Sistema Especialista	Programa inteligente de computador que usa conhecimentos e procedimentos inferenciais, para resolver problemas que são bastante difíceis, de forma a requererem para sua solução, muita perícia humana. Podem ser classificados quanto às características de seu funcionamento, e estão distribuídos em 10 categorias: interpretação, diagnósticos, monitoramento, predicação, planejamento, projeto, depuração, reparo, instrução e controle.
Fases de	As fases de desenvolvimento de um SE são 7 fases

desenvolvimento de um SE	distintas: identificação, conceituação, formalização, implementação, teste e avaliação, e a fase de revisão.
Base de conhecimento	Responsável por estruturar todo o conhecimento sobre o domínio da aplicação. A base de conhecimento é um elemento permanente, mas específico de um sistema especialista. Contém conhecimento, sob a forma de regras de produção, quadros, redes semânticas, ou seja, de várias formas
Motor de inferência	Implementa os algoritmos que decidirão quais as regras que serão satisfeitas pelos fatos ou objetos. Posteriormente, prioriza essas regras e executa aquela de maior prioridade