UNIVERSIDADE REGIONAL DE BLUMENAU

CENTRO DE CIÊNCIAS EXATAS E NATURAIS

CURSO DE CIÊNCIAS DA COMPUTAÇÃO

(BACHARELADO)

PROTÓTIPO DE UM SISTEMA ESPECIALISTA PARA ANÁLISE DE CRÉDITO DE PESSOAS FÍSICAS

TRABALHO DE CONCLUSÃO DE CURSO SUBMETIDO À UNIVERSIDADE REGIONAL DE BLUMENAU PARA OBTENÇÃO DOS CRÉDITOS DE DISCIPLINA COM NOME EQUIVALENTE NO CURSO DE CIÊNCIAS DA COMPUTAÇÃO - BACHARELADO

WANTOIR FEITEN

BLUMENAU (SC), NOVEMBRO/1999.

1999/2-40

UNIVERSIDADE REGIONAL DE BLUMENAU

CENTRO DE CIÊNCIAS EXATAS E NATURAIS

CURSO DE CIÊNCIAS DA COMPUTAÇÃO

(Bacharelado)

PROTÓTIPO DE UM SISTEMA ESPECIALISTA PARA ANÁLISE DE CRÉDITO DE PESSOAS FÍSICAS

TRABALHO DE CONCLUSÃO DE CURSO SUBMETIDO À UNIVERSIDADE REGIONAL DE BLUMENAU PARA A OBTENÇÃO DOS CRÉDITOS NA DISCIPLINA COM NOME EQUIVALENTE NO CURSO DE CIÊNCIAS DA COMPUTAÇÃO — BACHARELADO

WANTOIR FEITEN

BLUMENAU, NOVEMBRO/1999

PROTÓTIPO DE UM SISTEMA ESPECIALISTA PARA ANÁLISE DE CRÉDITO DE PESSOAS FÍSICAS

WANTOIR FEITEN

ESTE TRABALHO DE CONCLUSÃO DE CURSO, FOI JULGADO ADEQUADO PARA OBTENÇÃO DOS CRÉDITOS NA DISCIPLINA DE TRABALHO DE CONCLUSÃO DE CURSO OBRIGATÓRIA PARA OBTENÇÃO DO TÍTULO DE:

BACHAREL EM CIÊNCIAS DA COMPUTAÇÃO

_____

Prof. Roberto Heinzle — Orientador na FURB

_____

Prof. José Roque Voltolini da Silva — Coordenador do TCC

BANCA EXAMINADORA

_____

Prof. Roberto Heinzle

_____

Prof. Marcel Hugo

_____

Prof. Oscar Dalfovo

Este trabalho de conclusão de curso é dedicado à minha noiva, pelo apoio e incentivos recebidos ao longo destes anos de graduação.

AGRADECIMENTOS

Ao Professor e Orientador Roberto Heinzle pelo acompanhamento e incentivo na realização do trabalho.

A todos os amigos, professores e colegas do curso de Ciências da Computação pelo incentivo, ajuda, apoio e compreensão recebidos durante os anos de graduação.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

O presente estudo consiste em desenvolver um protótipo que auxilie na determinação do limite de crédito que poderá ser concedido à uma pessoa física, minimizando o risco de inadimplência. O protótipo a ser implementado é um sistema especialista que possibilita a tomada de decisões utilizando uma base de dados existente e de alguns questionamentos específicos para a ocasião. Para a implementação do protótipo será utilizada uma filosofia de *data mining*, para a obtenção dos dados da pessoa para serem usados no processo de decisão de limite de crédito a ser sugerido. Esta sugestão de limite de crédito será obtida através de iterações, realizadas com um conjunto de regras baseadas em Lógica Difusa, sobre informações extraídas de um banco de dados.

ABSTRACT

This work consist of developing a prototype auxiliary in the determination of credit limit of a person, minimizing the risc of inadimplence. The prototype to be implemented is an expert system which alow decision making by using an existent data base and a few specific question to the occasion. For the implementation of the prototype is used data mining tecnic to obtain person datas to be used in the process of the credit limit decision. This limit is calculated by iterations, realized with one rules set based in Fuzzy Logic, and informations extract of one data base.

1 INTRODUÇÃO

Acompanhando a economia nacional nos últimos anos, verifica-se que as diversas mudanças promovidas pelo governo impactaram diretamente a concessão de crédito pessoal, tanto no setor bancário/financeiro como no comércio. O processo de mudança de atitude, no que tange o crédito para pessoas físicas, dos tempos da inflação elevada para o momento de estabilidade, gerou uma desorientação para as pessoas e para as instituições financeiras, acarretando um aumento considerável na inadimplência [ALM92].

Neste instante, aprofundaram-se os estudos na análise de características das pessoas, para desenvolver Sistemas de Apoio à Decisão capazes de auxiliar na definição de quanto poderia ser concedido de crédito. Muitos dos modelos de análise de crédito para pessoas físicas desenvolvidos até o presente baseiam-se em cálculos matemáticos, suscetíveis a falhas por não considerar fatores de natureza humana. Desenvolveram-se também estudos utilizando Inteligência Artificial, especificamente com Sistemas Especialistas Probabilísticos, que utilizam Bases de Conhecimento e regras de produção.

Este trabalho consiste numa pesquisa sobre o processo de análise e concessão de crédito para pessoas físicas, e desenvolvimento de um protótipo para este fim, utilizando Inteligência Artificial, através de tecnologia de Sistemas Especialistas associado a métodos de *data mining* e lógica difusa. As informações sobre os clientes serão coletadas através de módulo de cadastro, para posteriormente serem utilizadas na definição do perfil do cliente ideal para concessão de crédito, minimizando o risco de inadimplência.

Para a determinação do perfil do cliente ideal será utilizada a filosofia de *data mining*, para vasculhar os dados cadastrais dos clientes e descobrir características associadas aos clientes que não apresentem históricos de inadimplência. Este perfil será utilizado então, pelo analista de crédito, para calcular o valor considerado tecnicamente possível de ser emprestado, para determinado cliente, com menor risco de inadimplência. Este cálculo será baseado em informações obtidas pelo estudo dos conceitos e métodos para concessão de crédito utilizados atualmente, mas aplicados de forma distinta da usual. A definição deste valor será feita com uso de lógica difusa, através de inferência com regras de produção, que farão ponderações entre as características do perfil do cliente ideal e as características do cliente para o qual se deseja conceder crédito.

1.1 OBJETIVOS DO TRABALHO

Com este trabalho objetiva-se desenvolver um protótipo para auxílio na determinação de um limite de crédito para empréstimo à pessoas físicas, de forma a minimizar o risco de inadimplência. Para tanto utilizar-se-á um sistema especialista que associa o uso de Lógica Difusa e *data mining* aplicados sobre um conjunto de informações extraídas de um Banco de Dados.

Será estudado *Data Mining* e Sistemas Especialistas, com utilização conjunta de Lógica Difusa, utilizando a ferramenta de desenvolvimento de aplicativos Delphi. Descrição da especificação e implementação de um protótipo de Sistema Especialista para Análise de Crédito, para apoio à decisão de crédito para instituições financeiras ou empresas comerciais.

1.2 ORGANIZAÇÃO DO TRABALHO

O trabalho está organizado em oito capítulos, descrevendo:

- Capítulo 1 - Introdução ao trabalho, com breve descrição do contexto deste, seus objetivos e sua organização.

- Capítulo 2 - Conceituação de Inteligência Artificial, descrição de seus objetivos e aplicações. Fundamentação teórica relativa a Sistemas Especialistas, abrangendo desde conceituação até suas limitações.

- Capítulo 3 - Fundamentação teórica da Teoria dos Conjuntos Difusos.

- Capítulo 4 - Fundamentação teórica de Data Mining, compreendendo conceituação, descrição de prospecção de conhecimento, utilidades e técnicas.

- Capítulo 5 - Consiste na fundamentação teórica da Análise de Crédito;

- Capítulo 6 - Apresentação do protótipo desenvolvido;

- Capítulo 7 - Conclusões e sugestões para futuros trabalhos.

2 INTELIGÊNCIA ARTIFICIAL

2.1 CONCEITUAÇÃO

Há muito tempo os pesquisadores e cientistas estudam a inteligência humana com o intuito de entender seu funcionamento. Diversas pesquisas foram desenvolvidas tentando reproduzir a forma humana de pensar.

No âmbito da computação, utiliza-se a expressão Inteligência Artificial para designar o estudo do comportamento inteligente. Mas o que é a Inteligência Artificial? Segundo [RAB95], "é o resultado da aplicação de técnicas e recursos, especialmente de natureza não numérica, viabilizando a solução de problemas que exigiriam do humano certo grau de raciocínio e de perícia". Para [HAR88], "a Inteligência Artificial é um campo de estudos que busca o desenvolvimento de sistemas inteligentes. Um sistema inteligente é aquele capaz de resolver problemas, que, quando resolvidos por humanos, exigem um comportamento dito inteligente".

2.2 OBJETIVOS DA INTELIGÊNCIA ARTIFICIAL

A Inteligência Artificial tem como objetivo “compreender os princípios que permitem simular a inteligência humana por meio da criação de modelos computacionais de processos cognitivos”, bem como “desenvolver sistemas (“hardware e software”) mais úteis e com capacidade de dedução e percepção” [RAB95].

2.3 APLICAÇÕES DA INTELIGÊNCIA ARTIFICIAL

A Inteligência Artificial (IA) pode ser aplicada onde existe inferência humana e esta necessita de alguma forma de auxílio ou automatização. Dentre as diversas áreas da IA, cita-se como exemplo a robótica, o processamento de linguagem natural, a computação algébrica, os sistemas especialistas, o reconhecimento de padrões, as bases de dados inteligentes, a prova de teoremas e os jogos. Podem existir outras aplicações, porém estas são as que mais tem se destacado nos últimos anos.

2.4 SISTEMAS ESPECIALISTAS

2.4.1 CONCEITUAÇÃO

[LEV88] afirma que "Sistemas Especialistas são programas de computador que usam conhecimento especializado e procedimentos de inferência para resolver problemas que normalmente são solucionados por especialistas humanos altamente experientes". Alguns problemas somente conseguem ser resolvidos por pessoas com conhecimento especializado sobre o assunto, treinamento e experiência. Estas pessoas são denominadas especialistas. Para solucionar estes mesmos problemas utilizando os recursos computacionais, utiliza-se os Sistemas Especialistas, que visam reproduzir o processo de resolução de um especialista humano.

Já [RIB87] escreve que "um sistema especialista é aquele que é projetado e desenvolvido para atender a uma aplicação determinada e limitada do conhecimento humano. É capaz de emitir uma decisão, com o apoio em conhecimento justificado, a partir de uma base de informações, tal qual um especialista de determinada área do conhecimento humano". Os Sistemas Especialistas na maioria das vezes, quando não solucionam o problema, diminuem o universo onde localiza-se a solução, facilitando a busca do resultado desejado.

2.4.2 EVOLUÇÃO DOS SISTEMAS ESPECIALISTAS

No final da década de 60, os primeiros pesquisadores de Inteligência Artificial concluíram que apenas produzindo um conjunto resumido de regras, e associá-las à um computador potente, não seria possível alcançar o desempenho da mente humana. Não conseguiriam obter sequer o desempenho da mente humana para solucionar um problema específico. Notaram que o objetivo procurado era demasiadamente grande.

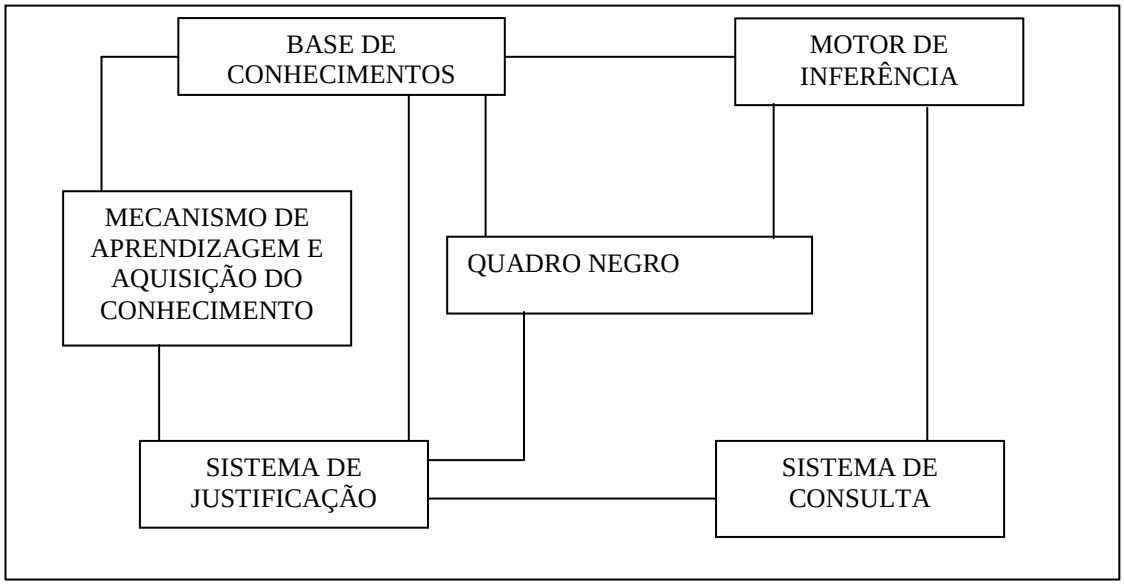
Trataram então de desenvolver pesquisas no sentido de conseguir atender à solução de problemas específicos, sem alternativas de aprendizado ou com aprendizado reduzido. Apesar dos resultados alcançados serem pequenos, satisfizeram os pesquisadores, que durante a década de 70 desenvolveram diversos Sistemas Especialistas. Podem ser destacados o Prospector e o Mycin, desenvolvidos para Geólogos e Médicos, respectivamente. Desde

então, as pesquisas tem evoluído de forma rápida, principalmente em torno da aquisição e representação do conhecimento.

2.4.3 ARQUITETURA DE UM SISTEMA ESPECIALISTA

Para exemplificar a estrutura de um Sistema Especialista genérico, a figura 1 mostra o modelo proposto por [HEI95].

Figura 1: Componentes de um Sistema Especialista.



Fonte: [HEI95]

A representação da figura 1 não é unânime entre os autores, porém é aceita pela maioria deles. Também cabe ressaltar que esta estrutura pode sofrer alterações, dependendo da implementação adotada e da forma de representação do conhecimento. Dentre os módulos constantes na figura 1, serão descritos a seguir os que tem maior relevância para o presente trabalho.

2.4.3.1 BASE DE CONHECIMENTOS

É o módulo principal de todo Sistema Especialista, pois contém o conhecimento necessário para que se consiga alcançar o objetivo pretendido. Pode-se dizer que a base de conhecimentos contém um somatório de crenças, fatos e heurísticas.

Este conhecimento é passado ao sistema pelo especialista e armazenado de uma forma própria que permite ao sistema fazer posteriormente o processamento ou inferência. A forma como o conhecimento é representado pode variar, sendo as mais comuns, regras de produção, *frames* e redes semânticas.

A fase de construção da base de conhecimentos é uma das mais complexas na implementação de um sistema especialista pois o conhecimento de um especialista não se encontra formalizado, precisando portanto de um trabalho prévio para tal. A base de conhecimentos está interligada com quase todos os demais elementos do sistema [HEI95].

2.4.3.2 MOTOR DE INFERÊNCIA

As informações armazenadas numa base de conhecimentos são, evidentemente, estáticas até que uma força externa analise e processe este conhecimento para ele tirar proveito. Este mecanismo, também conhecido como máquina de inferência, é responsável por buscar na base o conhecimento necessário a ser avaliado em cada situação, direcionar o processo de raciocínio, gerenciar situações de incerteza e levar ao resultado final.

Entretanto, de forma geral, pode-se afirmar que o processo envolve um encadeamento lógico que permita tirar conclusões a partir do conhecimento existente. O motor de inferência é, portanto, o responsável pela ação repetitiva de buscar, analisar e gerar novos conhecimentos [HEI95]. A forma de análise e interpretação envolve diversos tipos de soluções e às vezes até manipulação de incertezas, variando conforme o problema que se deseja resolver.

2.4.4 REPRESENTAÇÃO DO CONHECIMENTO

Segundo [LAP93], a representação do conhecimento constitui-se no conjunto de mecanismos usados para armazenar e manipular o conhecimento. Para [PER95], a representação do conhecimento caracteriza-se por métodos usados para modelar os conhecimentos de especialistas em algum campo, de forma eficiente, e colocá-los prontos para serem acessados pelo usuário de um sistema inteligente.

Existem várias maneiras de representar o conhecimento, sendo as principais as descritas a seguir:

a) regras de produção: é uma maneira bastante utilizada nos diversos sistemas especialistas existentes no mercado mundial [PER95]. Sua estrutura constitui-se basicamente de uma premissa, ou conjunto de premissas, e uma conclusão, ou conjunto de conclusões;

b) redes semânticas: são estruturas formadas por nós, conectados entre si através de arcos rotulados. Os nós representam objetos, conceitos, situações ou ações, e os arcos representam relações entre os nós [LAP93];

c) *frames*: também são chamados de quadros e compõem-se de estruturas de preenchimento que descrevem uma entidade real ou imaginária. Um *frame* é constituído por um nome, uma coleção de atributos, chamados de escaninhos ou *slots*, e valores associados a eles.

2.4.5 FERRAMENTAS DE CONSTRUÇÃO DE SISTEMAS ESPECIALISTAS

Conforme [HEI95] uma das maiores dificuldades na implementação de Sistemas Especialistas é quanto ao ambiente de programação. Desde 1958, quando foi criado o LISP, diversas ferramentas foram criadas para os mais diversos tipos de aplicações da inteligência artificial. As linguagens de programação que tem maior destaque são o LISP, o FORTRAN e o PROLOG, sendo que cada ambiente possui suas características específicas, definindo a área em que são utilizados.

Visando facilitar o desenvolvimento de aplicações, pois considerou-se que vários sistemas utilizariam uma mesma máquina de inferência, foram criadas ferramentas, denominadas Shell, para transcrever para o computador os Sistemas Especialistas.

2.4.6 LIMITAÇÕES DE UM SISTEMA ESPECIALISTA

Os pontos negativos, comumente ressaltados, segundo [RAB95], são que sistemas especialistas não são bons em representar o conhecimento temporal e espacial, em executar raciocínio de senso comum, em manipular conhecimento inconsciente e em reconhecer os seus próprios limites. Adicionalmente há muitas falhas nas ferramentas de IA disponíveis, especialmente no que concerne à manipulação concomitante de várias formas de representação do conhecimento e de sua aquisição.

Para [PAC91], uma análise do processo de resolução de problemas por parte do ser humano evidencia que este freqüentemente considera situações com informações de natureza qualitativa, incompleta ou incerta. O ser humano, quando busca por determinada solução, pressupõe que o tratamento de informações de tal natureza não se constitui em um obstáculo intransponível. Para os sistemas especialistas, no entanto, o tratamento deste tipo de informações é problemático e tem sido alvo de amplos estudos.

Contudo, para manipular informações de natureza qualitativa, incompleta ou incerta, pode-se associar aos sistemas especialistas a Teoria dos Conjuntos Difusos. Esta teoria será abordada no capítulo 3.

3 TEORIA DOS CONJUNTOS DIFUSOS

Segundo [RAB95], "existe em nossa comunicação quotidiana muitas palavras e sentenças com significado não preciso ou vago. Isto acontece porque, tanto quem fala como quem houve, não necessita de informações mais precisas e está acostumado a lidar com tais tipos de imprecisão. Por exemplo, alguém que no restaurante solicita uma sopa bem quente de barbatana de tubarão, não está preocupado com a real temperatura da sopa. O que ele deseja é que a temperatura da sopa esteja bastante acima do que ele considera como morna. Certamente ninguém é capaz de determinar o ponto preciso em que a sopa passa de morna para quente".

Interessado em representar tais imprecisões, o professor Lofti A Zadeh, da Universidade da Califórnia, Berkeley, desenvolveu a teoria dos conjuntos difusos, publicando um primeiro artigo sobre o assunto em 1962. Zadeh tratou o assunto pela denominação *fuzzy sets*, que é traduzida para o português como conjuntos difusos ou conjuntos nebulosos.

Esta teoria define que um conjunto não apresenta limites bem definidos, podendo um elemento pertencer parcialmente a ele, ou pertencer a dois conjuntos ao mesmo tempo. Os conjuntos difusos são classes que possuem elementos que estão associados a estas por graus de pertinência, que é uma medida que quantifica o grau ou a força com que estes elementos pertencem a um determinado conjunto. O mundo real indica estas classes através da incerteza, imprecisão ou do duvidoso.

3.1 DEFINIÇÃO

Na teoria clássica dos conjuntos, um elemento pertence ou não a um determinado conjunto, restringindo as fronteiras dos conjuntos e dando o mesmo peso a diferentes objetos. Para conjuntos ordinários, podemos associar o valor 1 aos elementos pertencentes a um conjunto e 0 para os elementos que não pertencem a ele. A função que associa estes valores é denominada de função característica do conjunto.

A teoria dos conjuntos difusos também permite que se tenha uma função característica, a qual é chamada de função de pertinência. Esta função de pertinência, em geral, assume valores no intervalo [0,1] e faz com que um objeto passe a não mais ser classificado como

estritamente pertencente ou não a um conjunto, mas sim, lhe designa graus de pertinência em relação a diferentes conjuntos [RAU96]. Tomando por exemplo um conjunto X:

a) para os elementos que com certeza pertencem ao conjunto X, é atribuído um grau de pertinência igual a 1;

b) para os elementos que com certeza não pertencem ao conjunto X, é atribuído um grau de pertinência igual a 0;

c) para os elementos que não se pode afirmar com certeza se pertencem ao conjunto X, é atribuído um valor intermediário, que tende para 1, quanto maiores forem as razões para crer que ele pertença ao conjunto X.

3.2 FUNÇÃO DE PERTINÊNCIA

O componente crucial de um conjunto difuso é sua função de pertinência, a qual quantifica o quanto cada objeto pertence ao conjunto. Assim, as operações sobre conjuntos difusos são definidas pela função de pertinência.

Segundo [WEL94], funções de pertinência são mecanismos através dos quais conjuntos difusos interagem com o mundo real. O domínio de uma função de pertinência é o conjunto de valores possíveis para uma dada variável.

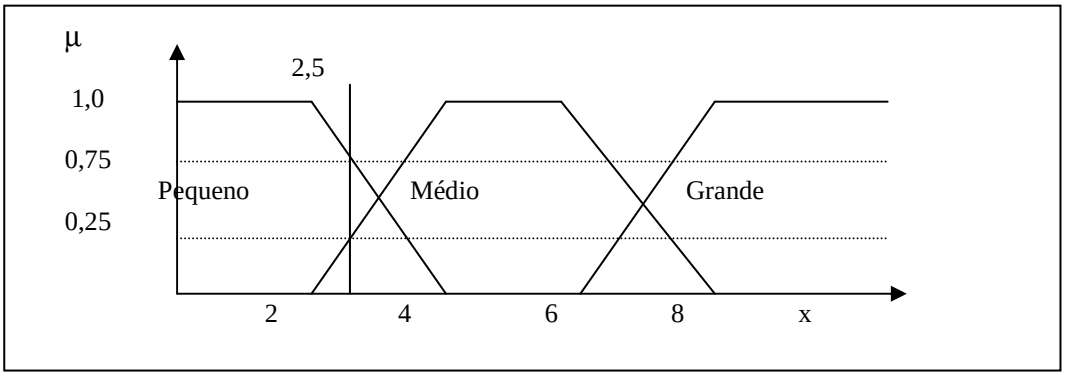
A figura 2 demonstra um exemplo empregando a função trapezoidal, onde x representa o lucro anual (em milhões de US$) de uma empresa. O lucro pode ser caracterizado como:

a) $\mu_{pequeno}(x) = (4 - x)/2$ se $2 < x < 4$, 0 se $x \geq 4$ e 1 se $x \leq 2$

b) $\mu_{médio}(x) = (x - 2)/2$ se $2 < x < 4$, $(8 - x)/2$ se $6 < x < 8$, 0 se $2 \geq x \geq 8$ e 1 se $4 \leq x \leq 6$

c) $\mu_{grande}(x) = (x - 6)/2$ se $6 < x < 8$, 0 se $x \leq 6$ e 1 se $x \geq 8$

Figura 2: Exemplo de função de pertinência.



Fonte: [WEL94]

Utilizando este exemplo, assumindo que o valor de x seja 2,5, tem-se um valor que está contido no conjunto difuso pequeno com um grau de pertinência igual a 0,75 e pertencendo a médio com um grau de pertinência igual a 0,25.

3.3 SISTEMAS DIFUSOS

Um sistema difuso consiste na combinação de conjuntos difusos definidos por variáveis de entrada e saída, junto com um conjunto de regras difusas, ligando um ou mais conjuntos difusos de entrada a um conjunto difuso de saída.

Os sistemas difusos são utilizados geralmente em aplicações que envolvem controle e redes neurais como reconhecimento de padrões. Sabe-se que os japoneses utilizam sistemas difusos em seus carros para controle de frenagem, suspensão ativa, controles de ignição e transmissão automática. Sistemas difusos também podem ser aplicados em modelos não lineares. Nestes casos, para executar um processo de tomada de decisão, por exemplo, estes sistemas baseiam-se em regras que utilizam variáveis lingüísticas difusas.

As regras que compõem estes sistemas são do tipo "SE ENTÃO", onde variáveis utilizadas nos antecedentes e nos conseqüentes são variáveis lingüísticas. Os antecedentes situam-se entre o SE e o ENTÃO, e os conseqüentes são posicionados após o ENTÃO.

3.4 RACIOCÍNIO DIFUSO

O raciocínio difuso é o processo pelo qual obtêm-se uma conclusão, geralmente imprecisa, deduzida através de um conjunto de premissas, também imprecisas.

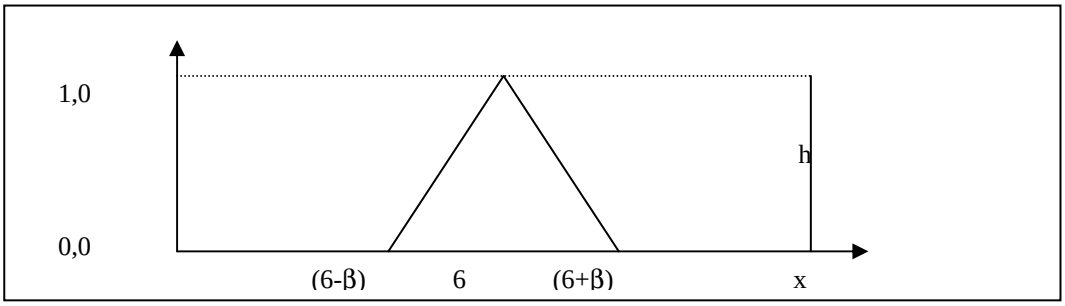
Na lógica clássica, os valores verdade são zero ou um, e o vocabulário é definido através desses valores verdade sob a forma de tabelas verdade. A lógica difusa baseia-se nas lógicas multivaloradas, em que os valores verdade variam no intervalo [0,1], podendo assumir os "meios termos". Zadeh diz que a lógica difusa é uma extensão da lógica multivalorada, em que os valores verdade são variáveis lingüísticas.

Para [RAB95], os melhores argumentos a favor da lógica difusa estão localizados não em seus fundamentos conceituais, mas em suas potenciais aplicações. Verifica-se que a lógica difusa tem larga aplicabilidade em áreas de controle e processos de tomada de decisão, onde a modelagem matemática precisa se torna inviável ou até impossível, dada a imprecisão dos elementos envolvidos, ou da existência de informações imprecisas e incompletas.

3.5 NÚMEROS DIFUSOS

[PER95] afirma que "um número difuso é um conjunto difuso que simultaneamente é convexo e normalizado, ou seja, é um subconjunto difuso de números reais". Um conjunto difuso é dito normalizado quando sua altura for 1. A altura (h) do conjunto é o limite superior do próprio conjunto [PAC91]. Na figura 3 vê-se uma exemplificação de um conjunto difuso normalizado.

Figura 3: Conjunto difuso de números reais próximos de 6.



Fonte: [RAU96]

Para [ROS95], um conjunto difuso convexo é descrito por uma função de pertinência cujos valores de pertinência são crescentes, ou decrescentes, ou ainda crescentes e

decrescentes, a medida que se incrementa os valores para os elementos no universo. Em outras palavras, para quaisquer elementos x, y e z pertencentes ao conjunto difuso A, a relação $x < y < z$ implica que $\mu_A(y) \geq \min [\mu_A(x),\mu_A(z)]$. Na figura 4 é mostrado um conjunto difuso convexo e na figura 5 um conjunto difuso não convexo.

Figura 4: Conjunto difuso convexo.

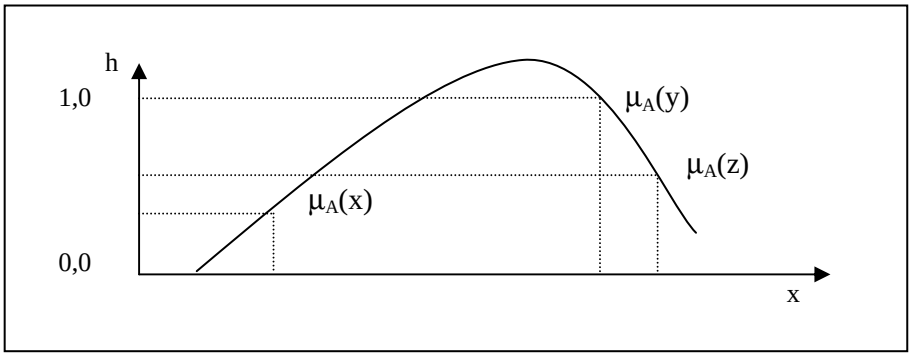
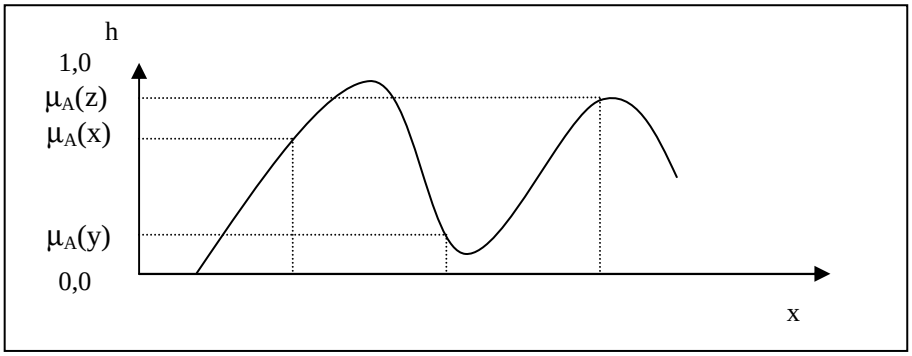


Figura 5: Conjunto difuso não convexo.



Segundo [PER95], um exemplo de números difusos é o conjunto difuso formado pelas expressões "pequeno", "aproximadamente 8" e "mais ou menos grande".

3.6 VARIÁVEIS LINGÜÍSTICAS

Sendo a teoria dos conjuntos difusos capaz de tratar a modelagem de situações complexas e imprecisas, esta também permite trabalhar com variáveis menos numéricas e menos precisas, chamadas de variáveis lingüísticas. [ROS95] descreve que uma variável lingüística difere de uma variável numérica já que seus valores não são números, mas palavras ou sentenças em uma linguagem natural ou artificial. Já que palavras, em geral, são menos precisas que números, o conceito de variável lingüística serve ao propósito de prover um meio

de aproximar caracterizações de fenômenos, os quais são muito complexos ou mal definidos quando descritos em termos quantitativos convencionais.

[PER95] afirma que "o uso deste tipo de variável permite que se faça estimativas numéricas de termos da linguagem natural. Para ele, uma área de aplicação particularmente importante das variáveis lingüísticas é a do raciocínio aproximado.

3.7 DESFUSIFICAÇÃO

Existem situações em que a saída de um processo difuso necessita ser um valor quantitativo. Este valor pode ser apurado a partir de um número difuso, através do processo denominado desfusificação. Segundo [ROS95], existem, pelo menos, sete métodos pesquisados e popularizados, de desfusificação, destacando-se o da pertinência máxima, o método da centróide e a média ponderada da pertinência máxima.

4 DATA MINING

4.1 DEFINIÇÃO

A evolução tecnológica dos últimos anos tornou relativamente fácil o acúmulo de dados. Como conseqüência surgiram grandes repositórios de dados, agregados de forma organizada e eficiente. Ao mesmo tempo, informação passa a ser valorizada como nunca antes na história, e os dados armazenados, vasculhados por especialistas, a procura de tendências e padrões.

Entretanto, a análise desses dados ainda é demorada, dispendiosa, pouco automatizada, e sujeita a erros, mal-entendidos e falta de acurácia. A automatização dos processos de análise de dados, com a utilização de softwares ligados diretamente à massa de informações, se tornou uma necessidade, já que o aproveitamento das informações já existentes, transformando-as em conhecimento, permite avanços sem paralelo na história do desenvolvimento dos bancos de dados [FIG98].

Neste capítulo é apresentado o *Data Mining*, que é a exploração e análise, por meios automáticos ou semi-automáticos, de uma grande quantidade de dados para descobrir padrões e regras significativos [BER97]. Serão descritas as etapas do Processo de KDD (*Knowledge Discovery in Databases* - KDD) e as tarefas que o *Data Mining* pode desempenhar.

4.2 PROSPECÇÃO DE CONHECIMENTO

Atribuindo algum significado especial a um dado, este se transforma em uma informação. Se especialistas elaboram uma regra, a interpretação do confronto entre a informação e a regra constitui um conhecimento [FIG98].

Prospecção de conhecimento em bases de dados (*Knowledge Discovery in Databases - KDD*) é um processo que envolve a automação da identificação e do reconhecimento de padrões em um banco de dados. Trata-se de uma pesquisa de fronteira, que começou a se expandir mais rapidamente nos últimos cinco anos. Sua principal característica é a extração não-trivial de informações a partir de uma base de dados de grande porte. Essas informações são necessariamente implícitas, previamente desconhecidas, e potencialmente úteis [FIG98].

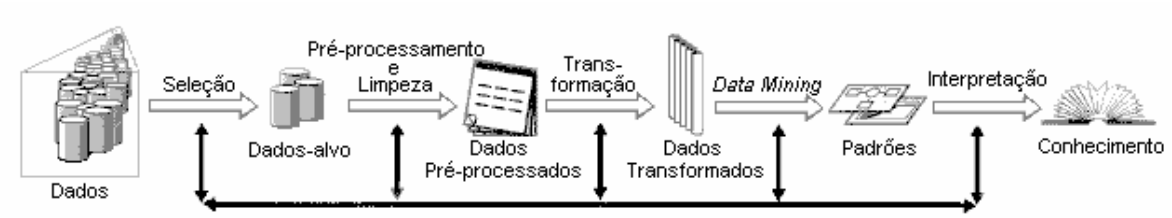
Devido a essas características incomuns, todo o processo de KDD depende de uma nova geração de ferramentas e técnicas de análise de dados, e envolve diversas etapas. A principal, que forma o núcleo do processo, e que muitas vezes se confunde com ele, chama-se *Data Mining*, ou Mineração de Dados, também conhecido como processamento de padrões de dados, arqueologia de dados, ou colheita de informação (*information harvesting*).

O KDD compreende todo o processo de descoberta de dados, enquanto o *Data Mining* refere-se a aplicação de algoritmos para extração de padrões de dados, sem os passos adicionais do KDD e da análise dos resultados [AVI98].

4.3 AS ETAPAS DO PROCESSO DE KDD

O processo de KDD (figura 6) começa com o entendimento do domínio da aplicação e dos objetivos finais a serem atingidos. Em seguida, é feito um agrupamento organizado de uma massa de dados, alvo da prospecção. A etapa da limpeza dos dados (*data cleaning*) vem a seguir, através de um pré-processamento dos dados, visando adequá-los aos algoritmos. Isso se faz através da integração de dados heterogêneos, eliminação de incompletude dos dados, repetição de registros, problemas de tipagem, etc. Essa etapa pode tomar até 80% do tempo necessário para todo o processo, devido às bem conhecidas dificuldades de integração de bases de dados heterogêneas [FAY96].

Figura 6: Os passos do processo de KDD.



Fonte: [FIG98]

Os dados pré-processados devem ainda passar por uma transformação que os armazena adequadamente, visando facilitar o uso das técnicas de *Data Mining*. Em algumas aplicações de *Data Mining* mais específicas, ferramentas avançadas de representação de conhecimento podem descrever o conteúdo de um banco de dados por si só, usando esse mapeamento como uma meta-camada para os dados.

Prosseguindo no processo, chega-se à fase de *Data Mining* especificamente, que começa com a escolha dos algoritmos a serem aplicados. Essa escolha depende fundamentalmente do objetivo do processo de KDD: classificação, segmentação, agrupamento por afinidades, estimativas, etc. De modo geral, na fase de *Data Mining*, ferramentas especializadas procuram padrões nos dados. Essa busca pode ser efetuada automaticamente pelo sistema ou interativamente com um analista, responsável pela geração de hipóteses.

Diversas ferramentas distintas, como redes neurais, indução de árvores de decisão, sistemas baseados em regras e programas estatísticos, tanto isoladamente quanto em combinação, podem ser então aplicadas ao problema. Em geral, o processo de busca é interativo, de forma que os analistas revêem o resultado, formam um novo conjunto de questões para refinar a busca em um dado aspecto das descobertas, e realimentam o sistema com novos parâmetros.

Ao final do processo, o sistema de *Data Mining* gera um relatório das descobertas, que passa então a ser interpretado pelos analistas de mineração. Somente após a interpretação das informações obtidas encontra-se o conhecimento.

Uma diferença significante entre *Data Mining* e outras ferramentas de análise está na maneira como exploram as interrelações entre os dados. As diversas ferramentas de análise disponíveis dispõem de um método baseado na verificação, isto é, o usuário constrói hipóteses sobre interrelações específicas e então verifica ou refuta, através do sistema. Esse modelo torna-se dependente da intuição e habilidade do analista em propor hipóteses interessantes, em manipular a complexidade do espaço de atributos, e em refinar a análise baseado nos resultados de consultas ao banco de dados potencialmente complexas. Já o processo de *Data Mining* fica responsável pela geração de hipóteses, garantindo mais rapidez, acurácia e completude aos resultados.

Estas etapas são interdependentes, pois os resultados de cada uma são a entrada da próxima etapa. Toda a abordagem é dirigida por resultados e cada estágio depende dos resultados do estágio anterior [FIG98]. Mas não existe uma ordem ou seqüência totalmente única para o andamento deste processo, porque isso depende das técnicas empregadas e dos dados sobre os quais o KDD está sendo aplicado [AVI98]. A qualquer momento, por

exemplo, pode-se voltar o processo de KDD para uma etapa anterior, desde que a técnica e os dados empregados permitam.

4.4 UTILIDADES DO DATA MINING

O *Data Mining* pode desempenhar uma série limitada de tarefas dependendo das circunstâncias. Cada classe de aplicação em *Data Mining* tem como base um conjunto de algoritmos que serão usados na extração de relações relevantes dentro de uma massa de dados [FIG98]:

a) classificação;
b) estimativa;
c) agrupamento por afinidade;
d) previsão;
e) segmentação.

Cada uma destas propostas difere quanto à classe de problemas que o algoritmo será capaz de resolver.

4.4.1 CLASSIFICAÇÃO

Classificação é uma técnica que consiste na aplicação de um conjunto de exemplos pré-classificados para desenvolver um modelo capaz de classificar uma população maior de registros. Em geral, algoritmos de classificação incluem árvores de decisão ou redes neurais, e começam com um treinamento a partir de transações-exemplo. O algoritmo classificador usa estes exemplos para determinar um conjunto de parâmetros, codificados em um modelo, que será mais tarde utilizado para a discriminação do restante dos dados.

Uma vez que o algoritmo classificador foi desenvolvido de forma eficiente, ele será usado de forma preditiva para classificar novos registros naquelas mesmas classes pré-definidas.

Alguns exemplos de Classificação são:

a) classificar pedidos de créditos como de baixo, médio e alto risco;
b) esclarecer pedidos de seguro fraudulentos;
c) atribuir palavras-chave a artigos jornalísticos.

Um modelo de classificação apanha um novo registro e atribui ao mesmo uma classificação existente. Um modelo de previsão é semelhante a um modelo de classificação, exceto por não ser limitado a um conjunto de número de classes. Um modelo de agrupamento toma vários registros e retorna um número menor de grupos. Esses grupos podem então ser aplicados a novos registros, criando um modelo de classificação. Um modelo de séries temporais é como um modelo de classificação ou de previsão, exceto por incluir dados tomados com o decorrer do tempo [BER97].

4.4.2 ESTIMATIVA

Uma variação do problema de classificação envolve a geração de valores ao longo das dimensões dos dados: são os chamados algoritmos de estimativa. A estimativa lida com resultados contínuos, ao contrário da classificação que lida com resultados discretos. Fornecidos alguns dados, usa-se a estimativa para estipular um valor para alguma variável contínua desconhecida como receita, altura ou saldo de cartão de crédito.

Ao invés de um classificador binário determinar um risco “positivo” ou “negativo”, a técnica gera valores de “escore”, dentro de uma determinada margem. A abordagem de estimativa tem a grande vantagem de que os registros individuais podem ser agora ordenados por classificação, e as redes neurais são adequadas a esta tarefa.

Exemplos de estimativa incluem:

a) estimar o número de filhos numa família;
b) estimar a renda total de uma família;
c) estimar o valor em tempo de vida de um cliente.

4.4.3 AGRUPAMENTO POR AFINIDADE

Este algoritmo identifica afinidades entre itens de um subconjunto de dados. Essas afinidades são expressas na forma de regras: “72% de todos os registros que contém os itens A, B, e C também contém D e E”. A porcentagem de ocorrência (72 no caso) representa o fator de confiança da regra, e costuma ser usado para eliminar tendências fracas, mantendo apenas as regras mais fortes. Dependências funcionais podem ser vistas como regras de associação com fator de confiança igual a 100%.

Trata-se de um algoritmo tipicamente endereçado à análise de mercado, onde o objetivo é encontrar tendências dentro de um grande número de registros de compras, por exemplo, expressas como transações. Essas tendências podem ajudar a entender e explorar padrões de compra naturais, e podem ser usadas para ajustar mostruários, modificar prateleiras ou propagandas, e introduzir atividades promocionais específicas. Um exemplo mais distinto, onde essa mesma técnica pode ser utilizada, é o caso de um banco de dados escolar, relacionando alunos e disciplinas. Uma regra do tipo “84% dos alunos inscritos em ‘Introdução ao Unix’ também estão inscritos em ‘Programação em C’ ” pode ser usada pela direção ou secretaria para planejar o currículo anual, ou alocar recursos como salas de aula e professores [FIG98].

4.4.4 PREVISÃO

A previsão é o mesmo que classificação ou estimativa, exceto pelo fato de que os registros são classificados de acordo com alguma atitude futura prevista. Em um trabalho de previsão, o único modo de confirmar a precisão da classificação é esperar para ver.

Essa tarefa é uma variante do problema de agrupamento por afinidades, onde as regras encontradas entre as relações podem ser usadas para identificar seqüências interessantes, que serão utilizadas para predizer acontecimentos subsequentes. Nesse caso, não apenas a coexistência de itens dentro de cada transação é importante, mas também a ordem em que aparecem, e o intervalo entre elas. Seqüências podem ser úteis para identificar padrões temporais, por exemplo entre compras em uma loja, ou utilização de cartões de crédito, ou ainda tratamentos médicos.

Exemplos de tarefas de previsão:

a) previsão de quais clientes sairão nos próximos seis meses;
b) previsão da quantia de dinheiro que um cliente utilizará caso seja oferecido a ele um certo limite de cartão de crédito.

4.4.5 SEGMENTAÇÃO

A segmentação é um processo de agrupamento de uma população heterogênea em vários subgrupos ou *clusters* mais homogêneos. O que a distingue da classificação é que a segmentação não depende de classes pré-determinadas.

Essa segmentação é realizada automaticamente por algoritmos que identificam características em comum e particionam o espaço n-dimensional definido pelos atributos. Os registros são agrupados de acordo com a semelhança e depende do usuário determinar qual o significado de cada segmento, caso exista algum. Muitas vezes a segmentação é uma das primeiras etapas dentro de um processo de *Data Mining*, já que identifica grupos de registros correlatos, que serão usados como ponto de partida para futuras explorações.

O exemplo clássico é o de segmentação demográfica, que serve de início para uma determinação das características de um grupo social, visando desde hábitos de compras até utilização de meios de transporte.

4.5 TÉCNICAS DE DATA MINING

Muitas das técnicas usadas em ferramentas atuais de *Data Mining* se originaram na pesquisa em inteligência artificial da década de 80 e princípio da década de 90. Entretanto, agora essas técnicas passaram a ser utilizadas em sistemas de banco de dados de grande escala, devido a confluência de diversos fatores que aumentaram o valor líquido da informação, dentre os quais se destacam [FIG98]:

a) a expansão e difusão de sistemas transacionais volumosos: nos últimos 15 ou 20 anos, computadores estão sendo usados para capturar e armazenar informações detalhadas de processos transacionais intensivos, como vendas, telecomunicações, bancos e operações com cartões de crédito. Os SGBDs saltaram de algumas centenas de transações por minuto para mais de 10.000/min, com exceções que chegam a 30.000. Esse crescimento da capacidade de processamento é acompanhado de uma redução equivalente do custo por processamento, que ajuda a disseminar a tecnologia e integrá-la ao mercado, gerando uma proliferação ainda maior de sistemas de transações geradores de informação.
b) informação como vantagem competitiva: a necessidade da informação resulta na proliferação de grandes repositórios de dados (*Data Warehouses*) que integram múltiplos sistemas operacionais para suporte a decisão, muitas vezes incluindo dados de fontes externas, como registros demográficos.

c) a difusão de tecnologia de informação escalável: a busca da interoperabilidade levou à recente adoção da sistemas de informação escaláveis, incluindo SGBDs, ferramentas analíticas e troca de informações via serviços de Internet/Intranet.

Por outro lado, a quantidade de dados brutos armazenados está crescendo rapidamente, tornando o "espaço de decisão" muito extenso e complexo para os atuais sistemas de suporte a decisão.

[FIG98] explica que, por causa desta grande quantidade de dados brutos, todo o processo de KDD atual ainda requer pré/pós-processamentos dos dados, necessários para assegurar o melhor aproveitamento da aplicação e a consistência dos resultados. Atividades de pré-processamento incluem a seleção apropriada de subconjuntos de dados, por razões de desempenho, assim como complexas transformações de dados que servem de ponte para o chamado "gap representacional", separação entre os dados e seu significado real. Pós-processamento envolve a subseleção de resultados volumosos e a aplicação de técnicas de visualização para auxiliar o entendimento. Essas atividades são críticas para contornar alguns problemas de implementação, tais como:

a) alta suscetibilidade a dados "sujos": as ferramentas de *Data Mining* via de regra não possuem uma estrutura dotada de semântica, orientada a aplicação, e como tal, tomam todos os dados factualmente. Torna-se necessário tomar precauções para assegurar que os dados analisados são "limpos", o que pode significar uma exaustiva análise dos atributos que alimentam os algoritmos. Entretanto, um bom processo de "limpeza de dados" (*data cleaning*), certamente beneficia o processo de *Data Mining*.
b) inabilidade para "explicar" resultados em termos humanos: mesmo em aplicações utilizando árvores de decisão e regras de indução, que são capazes de gerar informação sobre os atributos utilizados, o volume e formato da informação encontrada pode ser inútil sem um processamento adicional.
c) "gap" representacional: a maior parte das fontes de dados das aplicações de *Data Mining* atuais está armazenada em grandes sistemas relacionais, e seus dados estão em geral normalizados, com os atributos espalhados em múltiplas tabelas. Além disso, a maioria das ferramentas é restrita em termos dos tipos de dados com as quais podem operar, tornando-se necessário categorizar variáveis ou remapeá-las.

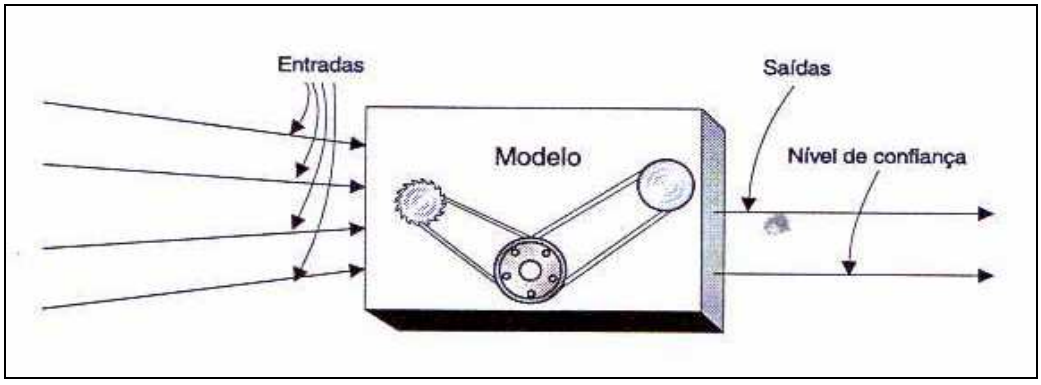
Conforme salienta [FIG98], um modelo produz um ou mais valores a partir de um dado conjunto de entradas. A análise dos dados é, com freqüência, o processo de construção de um modelo apropriado para os dados (figura 7). Um exemplo disso é uma regressão linear, onde é construída sobre um modelo em linha com a seguinte forma:

$$aX + bY + c = 0$$

Onde a, b, c são os parâmetros e X e Y são as variáveis. Para um dado valor de X, estima-se o valor de Y. Este tipo de modelo é um dos mais simples existentes.

O fato de um modelo existir não significa que proporcionará resultados precisos. Existem bons e maus modelos e, medir seus resultados é um passo crítico em seu uso e desenvolvimento [FIG98].

Figura 7: Modelo recebe entradas e produz informações.



[BER97]

Na criação dos modelos, a entrada é geralmente especificada claramente. Geralmente, preparar os dados de sistemas para preencher o domínio de um modelo – chamado de depuração de dados ou *data scrubbing* – é mais desafiador do que a própria criação do modelo. Os dados que alimentarão o modelo podem afetar a escolha da técnica. Para problemas físicos, com muitas variáveis contínuas de entrada, as técnicas de regressão estatísticas normalmente funcionam muito bem. Quando as entradas têm muitas variáveis de categorias, as árvores de decisão funcionam melhor. Quando a relação entre as entradas e a saída de dados é difícil de ser estabelecida, as redes neurais são as melhores opções.

Freqüentemente a saída de dados de um modelo é especificada em primeiro lugar e geralmente é uma categoria ou uma variável contínua.

Segundo [BER97], para criar um modelo para *Data Mining*, deve-se ter em mente o seguinte:

a) um dos perigos no uso de modelos é o excesso ou a carência de dados;
b) tanto o *Data Mining* direto quanto o indireto usam modelos, mas de maneira diversa;
c) alguns modelos expõem sua finalidade melhor que outros;
d) alguns modelos são mais fáceis de aplicar que outros.

Cada técnica de *data mining* possui tarefas onde elas são melhores aplicáveis. Cada classe de aplicação em *data mining* tem como base um conjunto de algoritmos que serão usados na extração de relações relevantes dentro de uma massa de dados: análise de seqüências, clusterização, classificação, estimativas e regras de associação. Outras técnicas mais recentes incluem lógica difusa (*fuzzy logic*) e algoritmos genéticos. Cada uma destas propostas difere quanto à classe de problemas que o algoritmo será capaz de resolver.

4.5.1 ANÁLISE DE SELEÇÃO ESTATÍSTICA

A análise de seleção estatística é uma forma de agrupamento usada para encontrar grupos de itens que tendem a ocorrer em conjunto em uma seleção estatística. Como técnica de agrupamento, ela é útil quando se deseja saber quais itens ocorrem ao mesmo tempo ou em uma seqüência particular [FIG98].

4.5.2 MBR

O MBR (*Memory-Based Reasoning* – raciocínio baseado em memória) é uma técnica de *data mining* dirigida que usa exemplos conhecidos como modelo para fazer previsões sobre exemplos desconhecidos. O MBR procura os vizinhos mais próximos nos exemplos conhecidos e combina seus valores para atribuir valores de classificação ou de previsão [BER97].

Os elementos-chave no MBR são a função de distância usada para encontrar os vizinhos mais próximos e a função de combinação, que combina valores dos vizinhos mais próximos para fazer uma previsão. Uma vantagem do MBR é sua habilidade de aprender sobre novas classificações simplesmente introduzindo novos exemplos no banco de dados. Uma vez encontrada a função de distância e a função de combinação corretas tendem a

permanecer muito estáveis, mesmo com a incorporação de novos exemplos para novas categorias nos dados conhecidos. Aliás, esta é uma característica que diferencia o MBR da maior parte das outras técnicas de *data mining*.

4.5.3 ALGORITMOS GENÉTICOS

Os algoritmos genéticos aplicam a mecânica da genética e seleção natural à pesquisa usada para encontrar os melhores conjuntos de parâmetros que descrevem uma função de previsão. Eles são utilizados no *data mining* dirigido e são semelhantes à estatística, em que a forma do modelo precisa ser conhecida em profundidade. Os algoritmos genéticos usam os operadores seleção, cruzamento e mutação para desenvolver sucessivas gerações de soluções. Com a evolução do algoritmo, somente os mais previsíveis sobrevivem, até as funções convergirem em uma solução ideal [BER97].

Esta técnica é apropriada para resolver os mesmos tipos de problemas que as outras técnicas de *data mining*, mas ela também pode ser usada para aprimorar MBRs e redes neurais.

4.5.4 DETECÇÃO DE AGRUPAMENTOS

Esta técnica constitui-se na construção de modelos para encontrar dados semelhantes, e estas reuniões por semelhança são chamadas de grupos (*clusters*). É uma forma de *data mining* não-direcionado, onde a meta é encontrar similaridades não conhecidas anteriormente. Existem muitas técnicas para encontrar grupos, incluindo métodos geométricos, estatísticos e redes neurais [HAR98].

4.5.5 ANÁLISE DE VÍNCULOS

A análise de vínculos segue as relações entre registros para desenvolver modelos baseados em padrões nas relações. Esse é um aplicativo de construção de teoria gráfica de *data mining*. Esta técnica não é muito compatível com a tecnologia de banco de dados relacionais e sua maior área de aplicação é a área policial, onde pistas são ligadas entre si para solucionar os crimes. As poucas ferramentas que existem, enfocam mais a visualização de vínculos que a análise de padrões [HAR98].

4.5.6 ÁRVORES DE DECISÃO E INDUÇÃO DE REGRAS

As árvores de decisão são usadas para o *data mining* dirigido, mais especificamente a classificação. Esta técnica divide os registros do conjunto de dados de treinamento em subconjuntos separados, cada um descrito por uma regra simples em um ou mais campos [HAR98].

Uma grande vantagem nesta técnica é que o modelo é bem explicável, já que tem a forma de regras explícitas. Isto permite às pessoas avaliarem os resultados, identificando os atributos-chave do processo.

4.5.7 REDES NEURAIS ARTIFICIAIS

As redes neurais são modelos simples de interconexões neurais no cérebro, adaptados para o uso em computadores e são, provavelmente, a técnica de *data mining* mais utilizada. Elas aprendem com um conjunto de dados de treinamento, generalizando modelos para classificação e previsão. Esta técnica pode também ser aplicada ao *data mining* não-dirigido (na forma de redes Kohonen e estruturas relacionadas) e às previsões em séries temporais [HAR98].

Uma das principais vantagens na utilização desta técnica é a sua variedade de aplicação. Elas são interessantes porque detectam padrões nos dados de forma análoga ao pensamento humano. Mas existem duas desvantagens em seu uso:

a) a dificuldade de interpretar os modelos produzidos por elas;
b) a sensibilidade ao formato dos dados que as alimentam, pois representações de dados diferentes podem produzir resultados diversos.

5 ANÁLISE DE CRÉDITO

5.1 INTRODUÇÃO

A decisão em finanças é sempre objeto de muitos cuidados por parte dos executivos financeiros. Em condições limites, poderá significar o fracasso ou o sucesso de toda uma administração. O ato de decidir, segundo [SEC96], é a mais importante função do administrador e a que envolve a maior relação custo-benefício, quando se trata do administrador financeiro.

Pode-se compreender então, que uma decisão tomada hoje deverá causar uma série de efeitos no futuro, embora existam grandes dificuldades em estabelecer claramente estes efeitos, ou mesmo instrumentos para detectá-los e quantificá-los. O executivo financeiro está constantemente tomando decisões dentro de um ambiente de mudanças, onde o risco e a incerteza preponderam em termos de conjuntura econômica, política e social [SEC96]. Apresenta-se nos tópicos a seguir as questões consideradas mais importantes para um processo de análise de crédito.

5.2 PROCESSO DE TOMADA DE DECISÃO

Em praticamente todas as atividades econômicas, "os homens de finanças estão constantemente sujeitos às tomadas de decisão que, muitas vezes, podem representar o fracasso ou o sucesso de determinado projeto, principalmente em economias tão atribuladas como a brasileira" [SEC96]. Toda vez que se toma uma decisão, utiliza-se dados conhecidos sobre o passado e faz-se previsões sobre o futuro. Segundo [LEM76], previsão é o processo pelo qual a partir de informações existentes, admitidas certas hipóteses e através de algum método de geração, chega-se a informações sobre o futuro, com uma determinada finalidade.

Embora haja dificuldades no estabelecimento de um processo para tomada de decisão, [SEC96] identifica três elementos que podem influenciá-lo:

a) experiência;

b) julgamento;

c) ambiente.

5.2.1 EXPERIÊNCIA

A experiência provém de um conjunto de situações vividas por uma pessoa ou empresa, e é tanto maior quanto maior for o número de exposições ao processo decisório. Quando se pensa em experiência é importante considerar o número de exposições a diferentes processos de decisão, o nível de responsabilidade do envolvido na tomada da decisão, com quem a compartilhou ou se era única e, finalmente, os resultados obtidos [SEC96].

5.2.2 JULGAMENTO

Normalmente uma decisão é tomada com base em experiências do passado, porém, em algumas situações, se é obrigado a contrariar a experiência. "É nestas ocasiões que aparece certa habilidade inata aos tomadores de decisão" [SEC96].

A ocorrência destes fatos amplia a experiência e conduz a um ciclo que envolve experiências e julgamentos. O julgamento tem muito a ver com a questão política dentro da organização; isso deve levar a uma postura objetiva na realização de um trabalho, de forma a tomar-se a decisão de qualidade boa ao invés de ficar procurando a ótima [SEC96]. O julgamento é, ainda, o responsável pelo exame da possibilidade de a decisão ser ou não efetivada.

5.2.3 AMBIENTE

O ambiente, dentro de um processo de tomada de decisão, deve ser analisado sob dois aspectos, antes e depois da decisão. É fundamental que o ambiente, do ponto de vista anterior ao instante da decisão, seja cultivado para que se facilite o processo decisório. É importante a diversificação do fluxo de informações e a consciência do grupo que decide em relação a sua cultura [SEC96].

No aspecto do ambiente pós decisão, [SEC96] declara que deve ser levado em conta que as decisões podem afetar pessoas, suas crenças, opiniões e conceitos pré-estabelecidos. Dentro deste quadro torna-se importante ao administrador não só a tomada de decisão, mas também a explicação da mesma.

5.3 RISCO

Na área financeira, o risco e a incerteza estão presentes em um grande número de decisões que, em conjunto, podem conduzir ao fracasso ou ao sucesso. A definição mais simples e prática parece ser dada por [SOL81]: "risco é o grau de incerteza a respeito de um evento". Um "evento certo" é tratado no estudo das probabilidades como correspondendo à probabilidade de 100% de que ocorra. Desta forma, "sempre que estivermos diante de eventos que apresentam certo grau de incerteza, podemos estabelecer uma correspondente probabilidade de ocorrência do evento" [SEC96].

O grau de incerteza, ou seja, o risco, estará intimamente ligado com a probabilidade de ocorrência dos eventos em estudo. A condição limite será a condição de incerteza plena em que não se quer ou não se tem condições de usar o conceito de probabilidades para a avaliação do evento. É por meio da probabilidade que se capta a influência da experiência, do julgamento e do ambiente, em diferentes condições de projeções de resultados, relativos a uma questão que será objeto de decisão. Nestas condições, a forma mais comum de tratamento da questão é a obtenção de uma distribuição de probabilidades, sua média e desvio.

5.4 ANÁLISE DISCRIMINANTE

Quando se tem uma situação em que se necessita classificar uma pessoa, por exemplo, como participante de um grupo, dentro de um universo de grupos, utiliza-se a análise discriminante.

Segundo [PET73], a análise discriminante destina-se a estabelecer um método para atribuir itens a populações predeterminadas. Também pode ser definida como um instrumento usado para encontrar semelhanças e diferenças entre dois ou mais conjuntos. No caso da análise discriminante aplicada ao processo de concessão de crédito, interessa saber quais características definem o "bom" cliente. Para isso, define-se como são o bom e o mau cliente para determinada instituição. Esta definição normalmente é baseada em dados como rentabilidade, fidelidade, número de produtos adquiridos, etc. A definição de bom e mau cliente varia de instituição para instituição. Como um segundo passo, seleciona-se dois grupos

de clientes daquela instituição, um de bons e outro de maus, para que sejam determinadas quais as características comuns.

Por característica, neste caso, devem ser consideradas variáveis como idade, tempo de emprego, valor do patrimônio, etc. Os valores que as características assumem são denominados de atributos. Abaixo, um exemplo extraído de [PET73] ilustra a análise discriminante.

Se aceitar todos os clientes que procurarem a instituição, encontrar-se-á dezesseis clientes bons para cada cliente mau. Esta relação pode ser expressa dizendo-se que a "probabilidade da população" [PET73] é de dezesseis para um.

Supondo-se também que cada cliente mau gere um prejuízo médio de R$ 400,00 e que cada cliente bom gere um lucro médio de R$ 20,00. O ponto de equilíbrio será de 400/20, ou seja, vinte para um. Isto equivale a dizer que será preciso vinte clientes bons para pagar um mau.

Imaginando-se que na amostra, as variáveis (características) idade e tipo de residência tenham apresentado o seguinte comportamento:

Tabela 1: Tabela de escore de duas características - idade

Idade	% de bons	% de maus	Probabilidade
Até 30	10	40	1/4
31 a 40	20	30	2/3
41 a 50	30	20	3/2
Acima de 50	40	10	4/1
	100 %	100 %	

Na tabela 1 verifica-se que as probabilidades para um cliente de 25 anos de idade serão de 16/1x(1/4)= 4/1. Partindo do pressuposto que o ponto de equilíbrio é de vinte para um, não se pode aceitar este cliente. Clientes com este perfil representam um risco além do estatisticamente aceitável.

No entanto, para um cliente de 50 anos de idade, as probabilidades serão de 16/1x(4/1)=64/1. Neste caso, o risco apresentado é inferior ao representado pelo ponto de equilíbrio (20/1), mostrando ser um cliente aceitável, pois neste grupo da população são encontrados sessenta e quatro clientes bons para cada mau. Esta tabela, porém, não é de muita utilidade, pois permitiria trabalhar apenas com clientes com idade acima de 50 anos.

Acrescentando-se então outra característica à pesquisa, registra-se também se a pessoa mora em casa própria ou alugada, supondo que o estudo desta caraterística tenha a seguinte distribuição:

Tabela 2: Tabela de escore de duas características - casa própria/alugada

Imóvel	% de bons	% de maus	Probabilidade
Próprio	60	30	2/1
Alugado	30	60	1/2
Outros	10	10	1/1
	100 %	100 %	

Tem-se então as seguintes relações de probabilidade:

a) 25 anos de idade e casa alugada = 16/1 x (1/4 x 1/2) = 2/1;

b) 50 anos de idade e casa própria = 16/1 x (4/1 x 2/1) = 128/1;

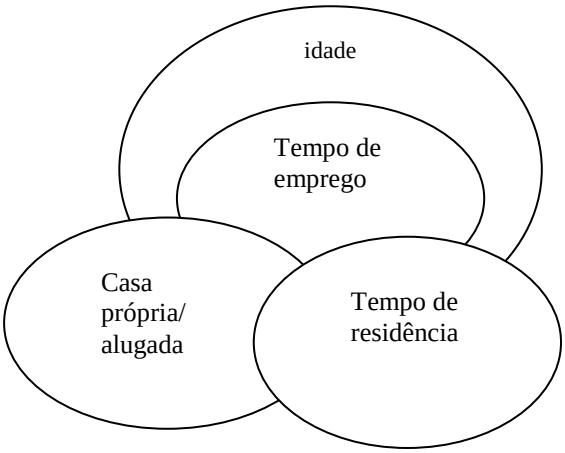
c) 35 anos de idade e casa própria = 16/1 x (2/3 x 4/1) = 43/1.

Percebe-se que o acréscimo de mais uma característica ampliou a área de atuação, levando a concluir que quanto mais características se estudar, mais refinado será o nosso modelo. Na prática são utilizadas entre oito e doze características [PET73].

5.5 MODELO DE ESCORAGEM

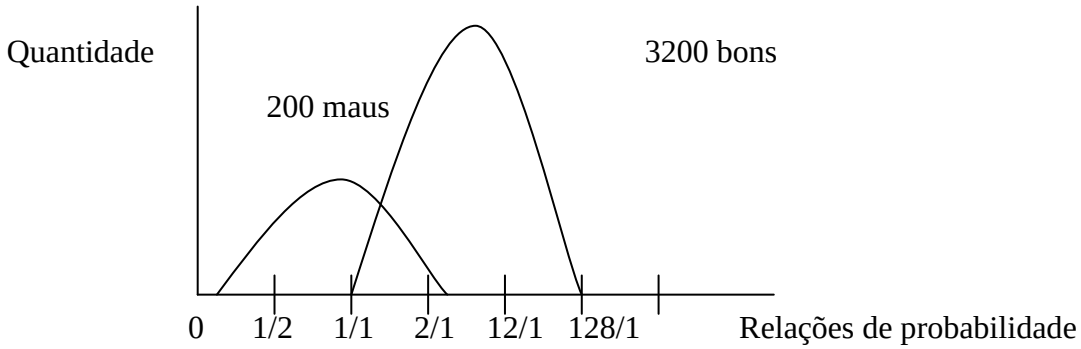
Ao estudar várias características, deve-se analisar o grau de correlação entre elas, como observado na figura 8.

Figura 8: Correlação entre características.



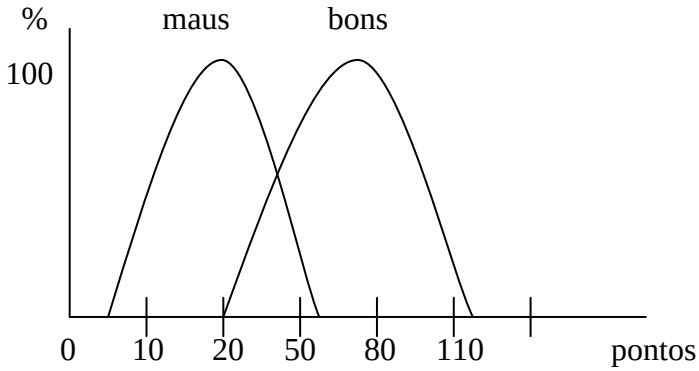
Transformando-se as probabilidades em pontos e montando um gráfico de sua distribuição, obtem-se algo próximo da distribuição normal, as conhecidas curvas em forma de sino (figura 9).

Figura 9: Distribuição de escore - 1.



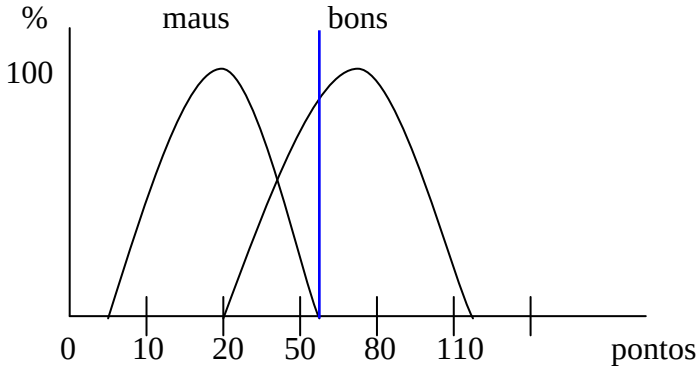
Este gráfico, em números absolutos de bons e maus clientes não é muito útil para leitura e análise. O mesmo gráfico, registrando porcentagens destes mesmos clientes facilita a visualização, como observado na figura 10.

Figura 10: Distribuição de escore - 2.



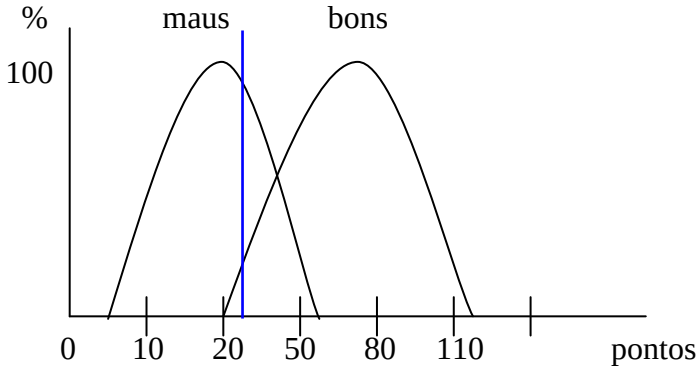
Supondo que se quisesse colocar um ponto de corte a 65 pontos, para eliminar todos os maus clientes, perder-se-ía cerca de 40% dos clientes potencialmente bons.

Figura 11: Ponto de corte 1.



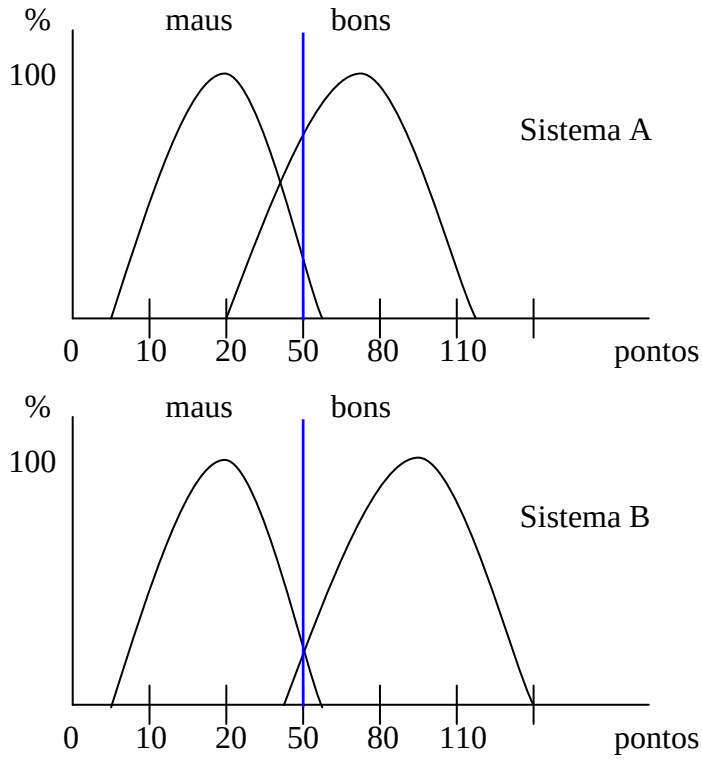
Entretanto, colocando-se o ponto de corte a 25, com o intuito de incluir quase todos os bons clientes, aceitar-se-á cerca de 50% dos clientes potencialmente maus.

Figura 12: Ponto de corte 2.



Assim, quanto mais afastadas estiverem as curvas, mais eficiente será o sistema. A distância entre os picos das curvas é chamada de "divergência" e mede o "poder discriminante" do sistema.

Figura 13: Divergência.



6 O PROTÓTIPO

6.1 INTRODUÇÃO

A proposta de construir um protótipo de um sistema especialista utilizando-se a teoria dos conjuntos difusos e a técnica de *data mining* para análise de seleção estatística, aplicados na área de análise de crédito, objetiva apoiar o especialista na realização de suas tarefas. O desenvolvimento de um protótipo deve necessariamente possuir uma especificação, onde define-se os requisitos da aplicação.

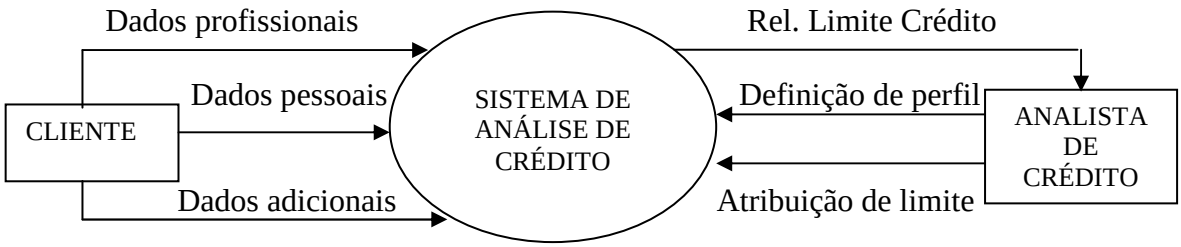
O passo inicial da fase de especificação do sistema é um levantamento de dados e informações para nortear o desenvolvimento do protótipo. Este levantamento é feito contatando-se o usuário, e elaborando-se uma descrição textual, por exemplo. A partir desta descrição, parte-se para a etapa de modelagem do protótipo. No capítulo 5 foi feita a descrição do processo de análise de crédito, e com as informações coletadas pode-se realizar diversas representações que auxiliam a implementação do protótipo, de acordo com a modelagem essencial.

6.2 MODELAGEM ESSENCIAL

Esta modelagem mostra, como o próprio nome diz, a essência do sistema a ser desenvolvido. É composta pelos modelos ambiental e comportamental. O modelo ambiental visa mostrar como o sistema interage com o ambiente externo e o modelo comportamental indica o que o sistema deve fazer para interagir com o ambiente externo.

O modelo ambienteal é composto por um diagrama de contexto, que representa o fluxo de dados, e por uma lista de eventos, que representa as tarefas que devem ser executadas no sistema. A figura 14 mostra o diagrama de contexto.

Figura 14: Modelo Ambiental - diagrama de contexto.

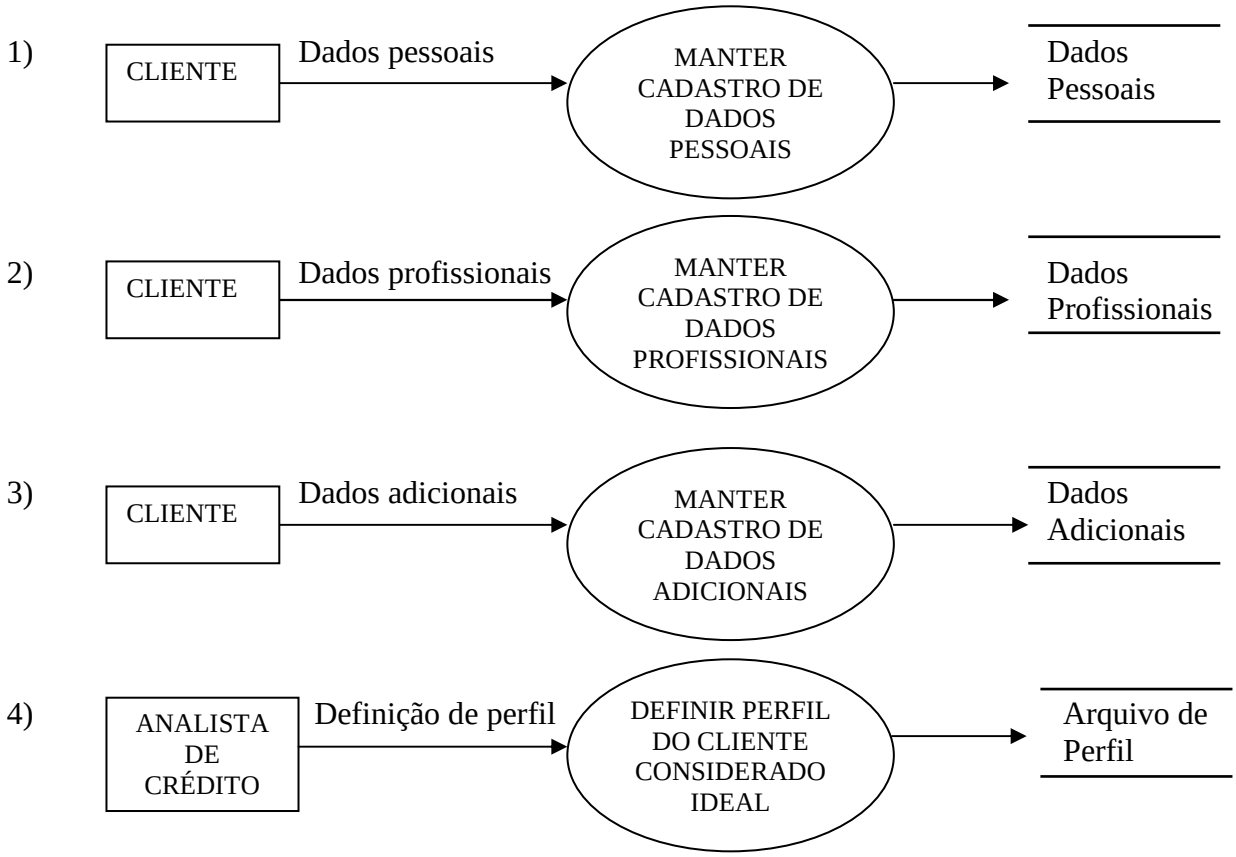


Os eventos do protótipo são:

1. Cliente solicita cadastramento.
2. Analista define perfil do cliente ideal.
3. Analista de Crédito efetua análise dos dados cadastrais do Cliente.
4. Emissão do Relatório de Limite de Crédito.

No modelo comportamental desenvolve-se o diagrama de fluxo de dados, dicionário de dados e modelo entidade x relacionamento (MER). O MER é um modelo conceitual de dados que representa, através de um diagrama, as associações existentes entre as entidades de dados, ou seja, demonstra relação existente entre os conjuntos de dados definidos a partir do levantamento realizado. Para o protótipo que pretende-se desenvolver, as figuras 15 e 16 mostram o diagrama de fluxo de dados (DFD), o MER e o dicionário de dados elaborados.

Figura 15: Modelo Comportamental - diagrama de fluxo de dados (DFD).



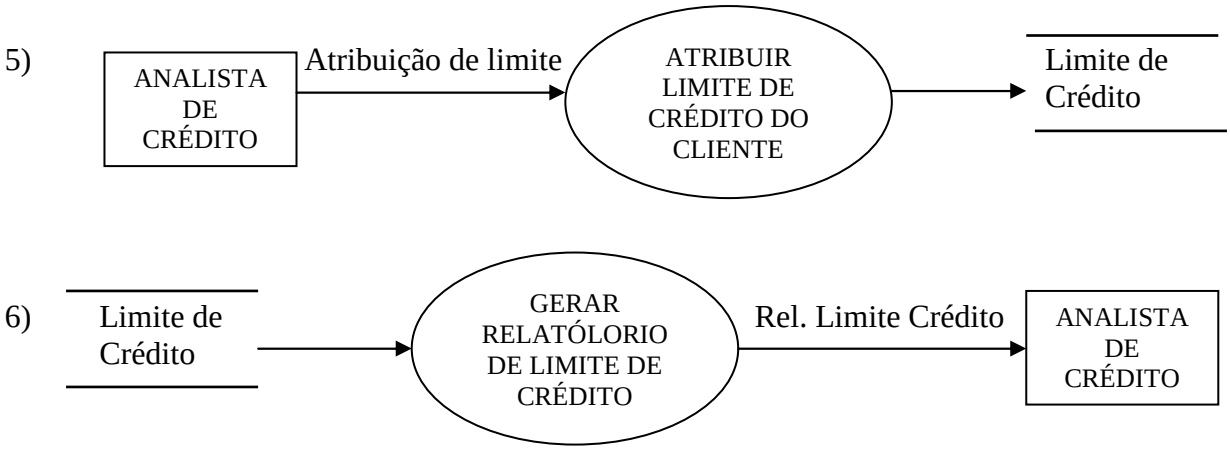
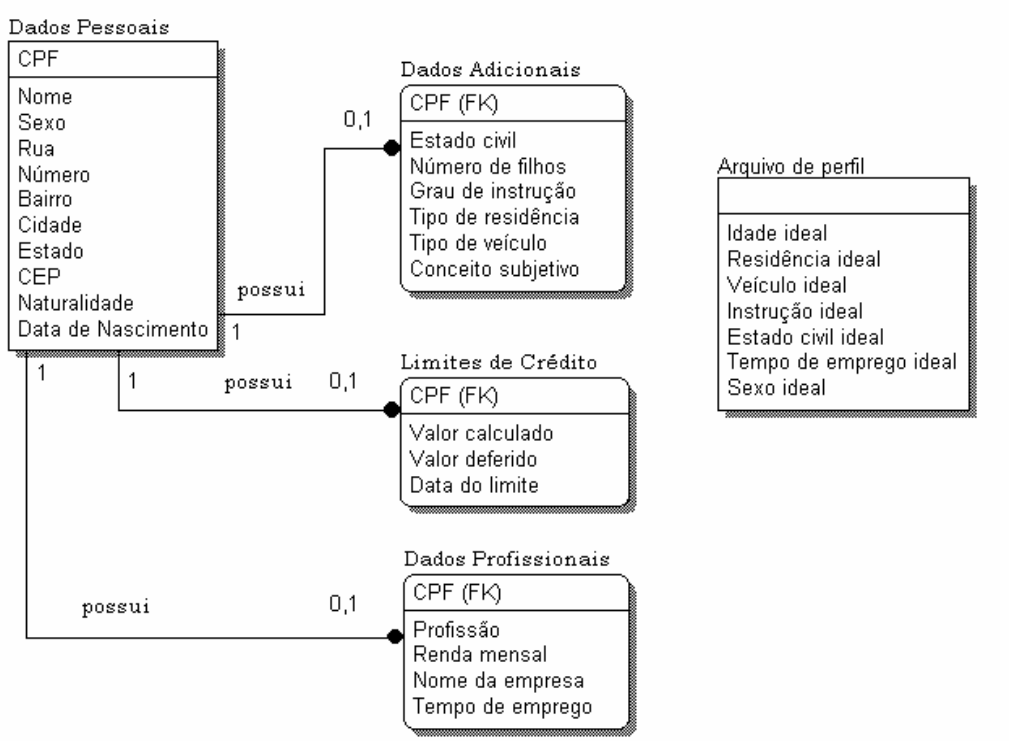


Figura 16: Modelo Entidade x Relacionamento.



Dicionário de dados:

Dados Pessoais = CPF + { nome + sexo + rua + número + bairro + cidade + estado + CEP +

naturalidade + data de nascimento}

Dados Profissionais = CPF + { profissão + renda mensal + nome da empresa + tempo de emprego}

Dados Adicionais = CPF + { estado civil + número de filhos + grau de instrução + tipo de residência + tipo de veículo + conceito subjetivo}

Atribuição de Limite = {CPF + valor deferido}

Relatório de Limite de Crédito = CPF + nome + data do limite + valor calculado + valor deferido}

A partir do MER são definidas as tabelas que comporão o projeto lógico do protótipo, definidas na forma da entidades identificadas, descritas a seguir:

a) Tabela Dados Pessoais: CPF, nome do cliente, sexo, rua, número, bairro, cidade, estado, CEP, naturalidade e data de nascimento;

b) Tabela Dados Profissionais: CPF, profissão, salário, nome da empresa e tempo de emprego;

c) Tabela Dados Adicionais: CPF, estado civil, número de filhos, grau de instrução, tipo de residência, tipo de veículo e conceito subjetivo;

d) Tabela Limite de Crédito: CPF, valor calculado, valor deferido e data do limite.

e) Arquivo de Perfil: idade ideal, residência ideal, veículo ideal, instrução ideal, estado civil ideal, tempo de emprego ideal e sexo ideal.

6.3 PLATAFORMA DE DESENVOLVIMENTO

O protótipo foi desenvolvido para a plataforma PC em ambiente Windows, utilizando-se um microcomputador com processador Celeron de 333 Mhz, e 64 Mb de memória RAM. O aplicativo final requer no mínimo processador de 100 Mhz, porém, com queda de performance. Para implementar o protótipo optou-se pelo ambiente de programação visual Borland Delphi, em sua versão 4.0, pela facilidade de se conseguir literatura e por já possuir

algum conhecimento sobre ela. O ambiente de programação Delphi possui algumas características merecedoras de destaque, como: "abordagem baseada em formulários e orientada a objetos, compilador extremamente rápido, suporte a banco de dados, integração com a programação em Windows e sua tecnologia de componentes" [CAN98].

6.4 AQUISIÇÃO DO CONHECIMENTO

Para a aquisição do conhecimento necessário para o desenvolvimento do protótipo, foram feitas entrevistas com funcionários do Banco do Brasil S.A., especialistas em análise de crédito. Destas entrevistas foram obtidas informações (capítulo 5) que, aliadas a dados obtidos junto à literatura especializada, deram base ao desenvolvimento do protótipo no que tange à aquisição do conhecimento.

Para dar entrada dos conhecimentos necessários, parte-se do menu principal (janela - Creditor) onde existe a chamada para o cadastro de informações dos clientes. As figuras 17 e 18 apresentam o menu principal e o menu de cadastro.

Figura 17: Menu principal (Creditor).

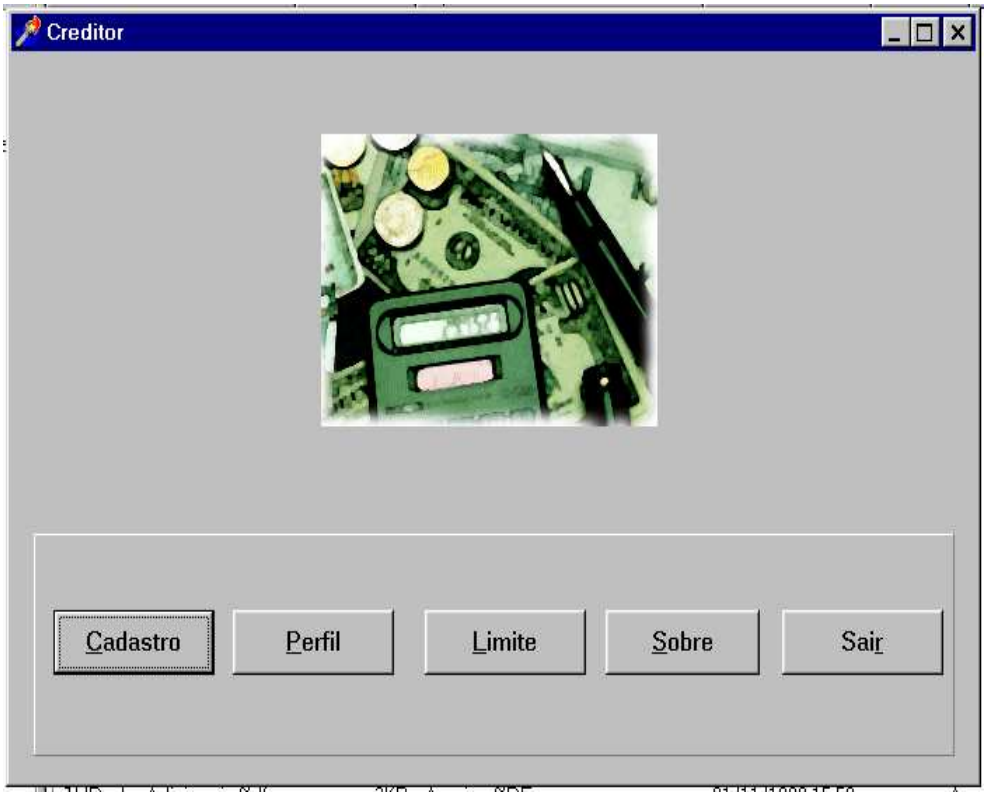
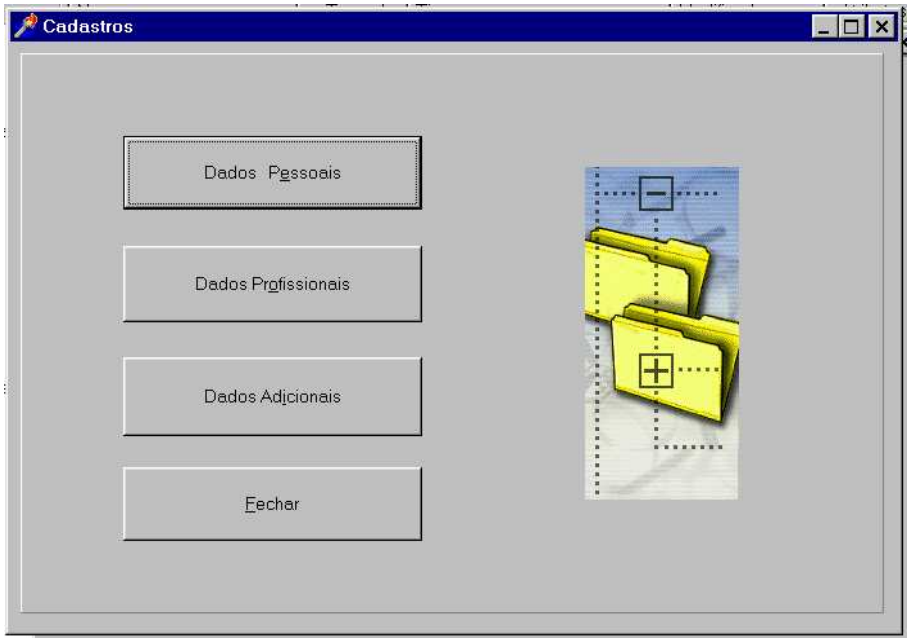


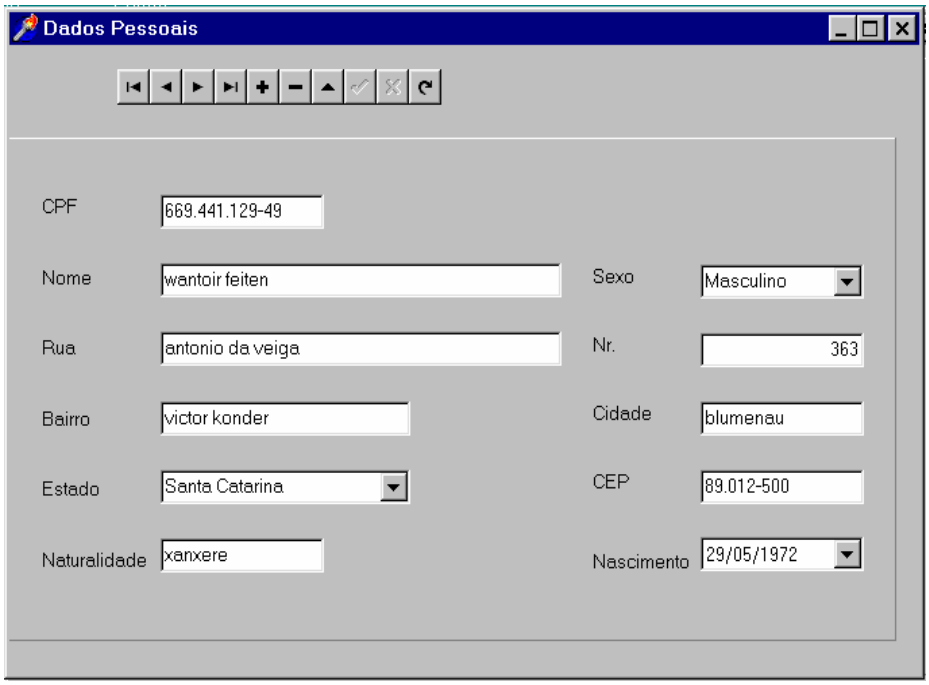
Figura 18: Cadastro.



A ordem na aquisição do conhecimento é dada pela ordem dos botões na tela, de cima para baixo:

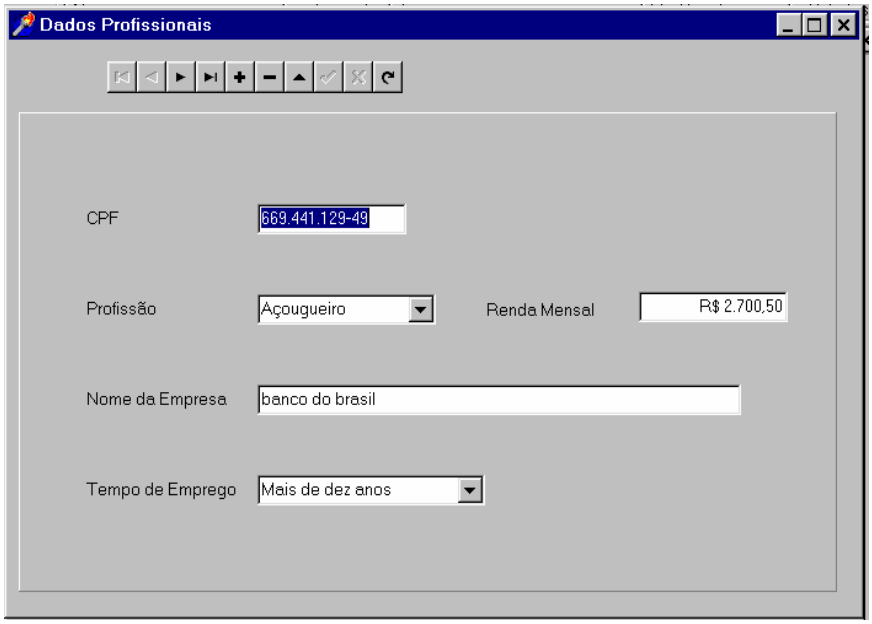
a) **Dados Pessoais:** os dados pessoais do cliente consistem em informações acerca de sua identificação e local de residência (figura 19);

Figura 19: Dados Pessoais.



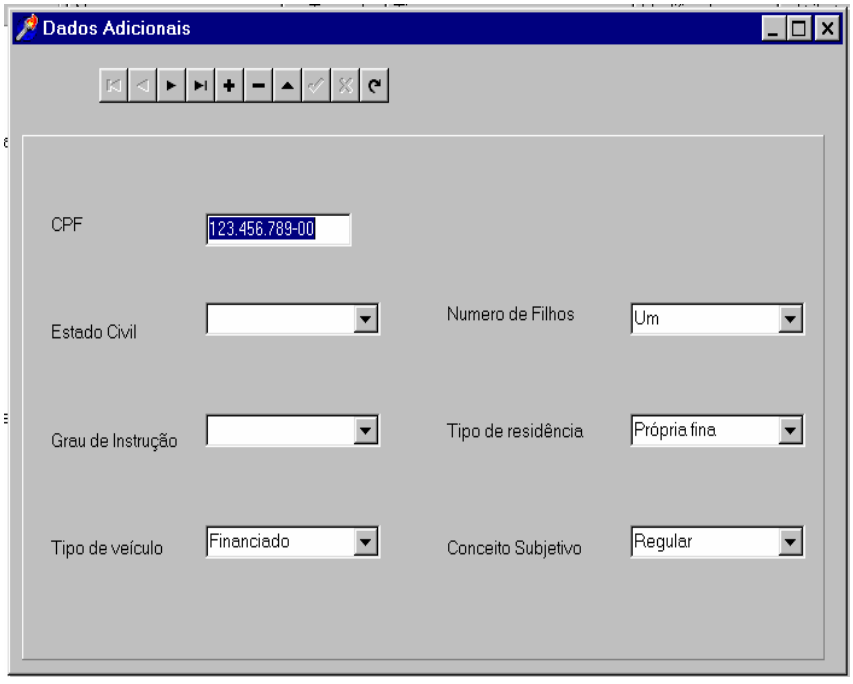
b) **Dados Profissionais:** os dados profissionais do cliente consistem em informações acerca de sua ocupação profissional (figura 20);

Figura 20: Dados Profissionais.



c) **Dados Adicionais:** os dados adicionais do cliente consistem em informações complementares, não relacionadas diretamente com sua ocupação profissional (figura 21);

Figura 21: Dados Adicionais.



6.5 REPRESENTAÇÃO DO CONHECIMENTO

A forma de representação do conhecimento escolhida para o sistema especialista é a forma de tabelas. Com a utilização do ambiente de programação Delphi 4.0, optou-se por montar uma estrutura de dados organizada em tabelas do tipo Paradox 7, fazendo uso do Database Desktop 7, utilitário que compõe o Delphi. As tabelas definidas no protótipo são: DadosPessoais.DB, Dados Profissionais.DB, DadosAdicionais.DB e Limites.DB. As três primeiras tabelas referem-se aos dados cadastrais dos clientes, adquiridos conforme relatado no item anterior. A tabela Limites.DB será descrita mais a frente, quando for relatado o processo de definição dos limites de crédito de cada cliente.

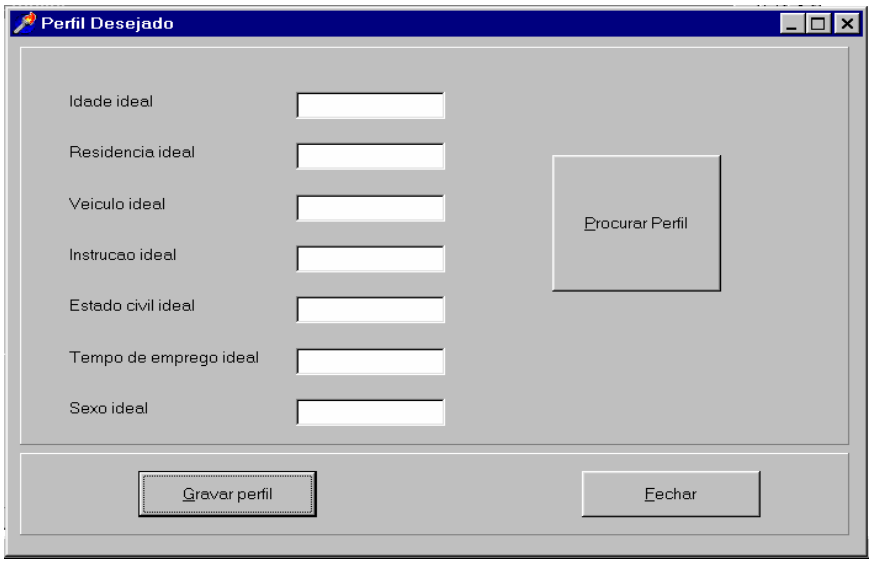
6.6 LIMPEZA DE DADOS

Para que o especialista realize a análise de cada cliente e determine um limite de crédito para este, é necessário que seja definido um perfil de cliente a ser tomado como base. Para definir este perfil, os especialistas utilizam-se geralmente da análise discriminante, descrita no capítulo 6.4. Como alternativa à este processo, foi implementada a etapa do processo de KDD denominada limpeza de dados.

A implementação baseou-se em pesquisas na base de informações, através de instruções em SQL (*Structured Query Language* - linguagem de consulta estruturada), que é uma linguagem de consulta e manipulação de bancos de dados. Estas pesquisas tem por finalidade identificar alguns dados dos clientes cujo item de afinidade é o conceito subjetivo "bom". Os resultados das consultas à base de dados são tratados de diversas formas, de acordo com a variável que se está definindo. Para o quesito idade é feita uma média entre as idades das pessoas selecionadas. Para os demais quesitos (residência, veículo, grau de instrução, estado civil, tempo de emprego e sexo), é verificado qual situação é encontrada com maior freqüência.

O processo é disparado na tela "Perfil Desejado", acessada através do botão "Perfil" da tela principal, quando o botão "Procurar Perfil" é acionado. (figura 22).

Figura 22: Perfil Desejado.

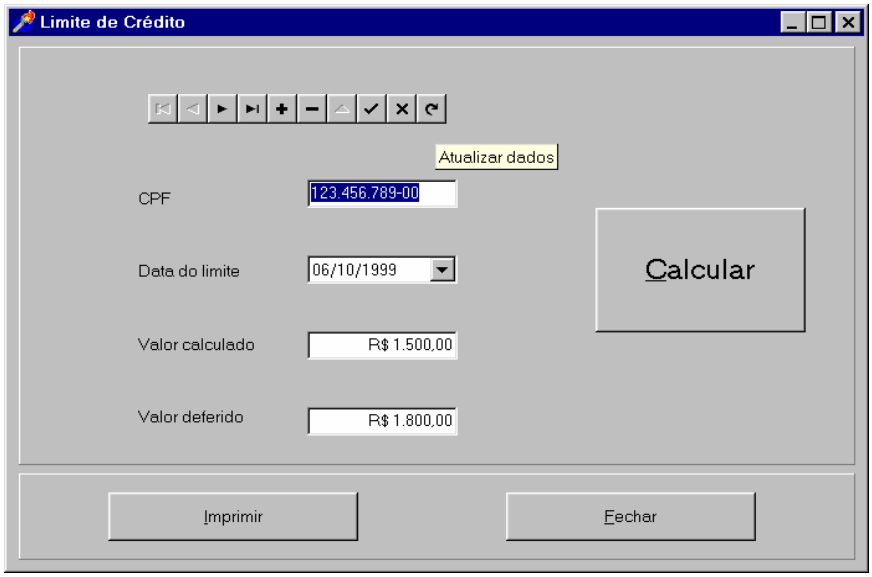


As informações definidas pelo processo executado podem ser gravadas no arquivo "Perfil.dat" acionando-se o botão "Gravar perfil".

6.7 MODELAGEM DIFUSA

Partindo do menu principal, através do botão "Perfil", aciona-se o módulo de definição do limite de crédito do cliente (figura 23). Neste módulo são realizadas iterações sobre a base de conhecimento, com o uso de lógica difusa, confrontando informações desta base com as informações do cadastro do cliente.

Figura 23: Limite de crédito.



Os dados dos limites de crédito já definidos podem ser recuperados digitando-se o CPF do cliente no campo próprio e acionando-se em seguida o botão "Recuperar". Se não houver limite definido, pode-se calcular o limite acionando-se o botão "Calcular". Após informado o valor deferido para o cliente, pelo especialista, este pode ser gravado na tabela Limites.DB, referida no capítulo 7.3. Estes dados também podem ser impressos.

6.7.1 CONJUNTOS DIFUSOS

Para calcular o limite de crédito foram definidos conjuntos difusos das informações representadas pelos requisitos **Salário**, **Instrução**, **Idade**, **Tempo de emprego** e **Filhos**, conforme descrito a seguir:

a) Para o requisito **Salário** foram criados três conjuntos difusos:

   – 1º conjunto é constituído dos valores que iniciam em zero e terminam em 500. Conjunto: (0,00 a 500,00) reais.

   – 2º conjunto é constituído dos valores que iniciam em 501 e terminam em 3.000. Conjunto: (501,00 a 3.000,00) reais.

   – 3º conjunto é constituído dos valores superiores a 3.001. Conjunto: (3.001,00 ao infinito) reais.

b) Para o requisito **Instrução**, o conjunto difuso foi construído tendo por base o número de anos necessários para se alcançar cada grau de formação. Para cada grau foi definida uma variável linguística, associada ao número de anos necessários para seu atingimento. Conjunto: (analfabeto, primário, ginásio, segundo grau, terceiro grau, pós-graduação, mestrado, doutorado).

c) No caso do requisito **Idade**, o conjunto difuso foi construído a partir da informação do perfil definido, conforme explicado no tópico 6.6. Esta informação é tomada como valor máximo inteiro positivo, sendo que o menor valor inteiro positivo sempre será 0 (zero). O limite superior é definido pela variável "Idade" do perfil do cliente ideal definido.

d) No requisito **Tempo de Emprego**, o conjunto difuso foi construído definido-se cinco intervalos de tempo associados à variáveis linguísticas. Conjunto: (até um ano, um a dois anos, três a cinco anos, seis a dez anos, mais de dez anos).

e) Para o requisito **Filhos**, a construção do conjunto difuso foi construído definido-se cinco variáveis linguísticas que são associadas ao número de filhos, iniciando-se em zero e chegando-se ao limite superior estabelecido como maior do que três. Conjunto: (nenhum, um, dois, três, mais de três).

Para as demais características não é possível criar um conjunto difuso porque as informações estão em um estado único. Como exemplo temos o caso do sexo, que só pode possuir dois estados: masculino e feminino.

6.7.2 FUNÇÕES DE PERTINÊNCIA

Como descrito no capítulo 4.2, função de pertinência é uma função matemática que tem como resultado o grau de pertinência de cada elemento de um conjunto difuso. Para cada conjunto difuso criado, citados no capítulo anterior, foi criada uma ou mais funções de pertinência, descritas a seguir, com informações obtidas junto a especialistas:

a) **Salário:** Existem neste caso três funções de pertinência (apêndice 1), uma para cada faixa de valor definida. A primeira para salários inferiores a R$ 500,00, a segunda para valores entre R$ 500,00 e R$ 3.000,00, e a terceira para valores acima de R$ 3.000,00.

- Função de pertinência 1: FP = (300/30)
- Função de pertinência 2: FP = (300/15)
- Função de pertinência 3: FP = (300/10)

b) **Instrução:** Se o grau de instrução do cliente for igual ao do perfil definido, é atribuído o grau de pertinência igual a 100, caso contrário, é aplicada a função abaixo, onde o valor linguístico é atribuído de acordo com a variável linguística. Para cada grau de instrução, identificados através das variáveis linguísticas, são

definidos os valores lingüísticos, atribuindo-se um valor numérico, correspondente ao número de anos necessários para aquela formação.

FP = (valor linguístico /tempo total de graduação)*80.

c) **Idade:** São duas funções de pertinência, sendo uma para quando a idade do cliente for inferior à idade do perfil, outra para quando tiver idade superior, e a terceira para quando tiver idade igual, quando o valor de pertinência será 100.

- Função de pertinência 1: FP = ((Idade cliente- Idade perfil)/ Idade perfil)*100)

- Função de pertinência 2: FP = ((Idade perfil- Idade cliente)/ Idade cliente)*100)

d) **Tempo de Emprego:** Quando o tempo de emprego do cliente for igual ao do perfil definido, é atribuído o grau de pertinência igual a 100, caso contrário, é aplicada a função abaixo, onde o valor linguístico é atribuído de acordo com a variável linguística. Para cada tempo de emprego, identificados através das variáveis linguísticas, são definidos os valores lingüísticos, atribuindo-se um valor numérico.

FP = (valor linguístico /10)*80.

e) **Filhos:** Para cada variável linguística criada é atribuído um valor liguístico. Quando o valor linguístico for igual a dois, é atribuído o grau de pertinência igual a 100, caso contrário, é aplicada a função de pertinência abaixo.

FP = 100 - ((valor linguístico/10)*80).

Para as demais variáveis existem somente duas funções de pertinência. A primeira possui valor de pertinência igual a 1,0, quando o cliente possui a característica desejada. A segunda atribui valor de pertinência igual a 0,0.

6.7.3 MÁQUINA DE INFERÊNCIA

A inferência é realizada através da aplicação das funções de pertinência aplicadas conforme seleção, de acordo com as informações do perfil desejado, manipulando os

conhecimentos armazenados na base de conhecimentos. A máquina de inferência foi programada com instruções do tipo:

SE <informação> menor que <condição>
ENTÃO dispara função de pertinência;

O processo de inferência manipula os conhecimentos cruzando os dados armazenados do perfil desejado com os dados do cliente, através das funções de pertinência. Quando não há função de pertinência, é verificado somente se a condição procurada existe ou não na base de dados do cliente.

Como resultado é encontrado o grau de pertinência entre o cliente analisado e o perfil do cliente definido conforme explicado no tópico 6.6. Com este grau de pertinência, foi utilizada uma fórmula que concluirá qual a sugestão de Limite de Crédito a ser proposta (apêndice 1).

6.8 TESTES REALIZADOS

Para testar o protótipo da forma mais abrangente possível, foi efetuado o cadastramento de diversos clientes fictícios, com as mais diversas características, tentando compreender a maior quantidade de variações. Para certificar outras funções do módulo de cadastramento, foram testadas também as opções de exclusão, alteração e navegação pelas informações das tabelas.

Foram realizadas várias definições de perfil, em diversos momentos do processo de cadastramento, como forma de observar as alterações das características definidas, de acordo com o aumento da base de dados. Para cada perfil definido, foram escolhidos aleatoriamente diversos clientes cadastrados, e definidos seus limites de crédito.

Para exemplificar o processo de inferência que leva à definição do limite de crédito de um cliente, será usado um perfil exemplo (tabela 3), e demonstradas as inferências realizadas para um cliente exemplo (tabela 4). A seguir é descrito o processo de inferência realizado com os dados de exemplo, e o cálculo do valor a ser sugerido ao analista como Limite de Crédito a ser definido para o cliente em análise.

Tabela 3: Perfil exemplo

Idade ideal	23 anos
Residência ideal	De familiares
Veículo ideal	Próprio
Instrução ideal	Mestrado
Estado civil ideal	Solteiro
Tempo de emprego ideal	Dois a cinco anos
Sexo ideal	Feminino

Tabela 4: Dados cadastrais do cliente exemplo

Idade	22 anos
Residência	Própria
Veículo	Próprio
Instrução	Terceiro grau
Estado civil	Solteiro
Tempo de emprego	Dois a cinco anos
Sexo	Masculino
Número de filhos	Nenhum
Salário	R$ 2.900,00

Regra 1
SE Idade do Cliente (22) = Idade do Perfil (23)
ENTÃO Valor de Pertinência = 100
SENÃO SE Idade do Cliente (22) > Idade do Perfil (23)
ENTÃO Valor de Pertinência = ((Idade do Cliente - Idade do Perfil) /Idade do Perfil)*100
SENÃO Valor de Pertinência = ((Idade do Perfil (23) - Idade do Cliente (22)) /Idade do Cliente (22))*100

Valor de Pertinência Idade = 5

Regra 2
SE Residência do Cliente (Própria) = Residência do Perfil (De familiares)
ENTÃO Valor de Pertinência = 100
SENÃO Valor de Pertinência = 0

Valor de Pertinência Residência = 0

Regra 3
SE Veículo do Cliente (Próprio) = Veículo do Perfil (Próprio)
ENTÃO Valor de Pertinência = 100
SENÃO Valor de Pertinência = 0

Valor de Pertinência Veículo = 100

Regra 4
SE Estado Civil do Cliente (Solteiro) = Estado Civil do Perfil (Solteiro)
ENTÃO Valor de Pertinência = 100
SENÃO Valor de Pertinência = 0

Valor de Pertinência Estado Civil = 100

Transformação das variáveis lingüísticas de instrução em valores
SE Instrução do Cliente (Terceiro grau) = 'Terceiro grau'
ENTÃO Valor Lingüístico = 16;

Regra 5
SE Instrução do Cliente (Terceiro grau) = Instrução do Perfil (Mestrado)
ENTÃO Valor de Pertinência = 100
SENÃO Valor de Pertinência = (Valor Lingüístico (16)/24)*80

Valor de Pertinência Instrução = 53

Transformação das variáveis lingüísticas de tempo de emprego em valores
SE Tempo de Emprego do Cliente (Dois a cinco anos) = 'Dois a cinco anos'
ENTÃO Valor Lingüístico = 6;

Regra 6
SE Tempo de Emprego do Cliente (Dois a cinco anos) = Tempo de Emprego do Perfil (Dois a cinco anos)
ENTÃO Valor de Pertinência = 100
SENÃO Valor de Pertinência = (Valor Lingüístico/10)*80

Valor de Pertinência Tempo de Emprego = 100

Regra 7
SE Sexo do Cliente (Masculino) = Sexo do Perfil (Feminino)
ENTÃO Valor de Pertinência = 100
SENÃO Valor de Pertinência = 0

Valor de Pertinência Sexo = 0

Transformação das variáveis lingüísticas de número de filhos em valores
SE Número de Filhos do Cliente (Nenhum) = 'Nenhum'
ENTÃO Valor Lingüístico = 2;

Regra 8
SE Número de Filhos do Cliente (Nenhum) = 'Nenhum'
ENTÃO Valor de Pertinência = 100

SENÃO Valor de Pertinência = 100 - ((Valor Lingüístico/10)*80)

Valor de Pertinência Número de Filhos = 100

Regra 9

SE Salário do Cliente (2.900) < 500
ENTÃO Valor de Pertinência = (300/30)
SENÃO SE Idade do Cliente (2.900) < 3000
  ENTÃO Valor de Pertinência = (300/15)
  SENÃO Valor de Pertinência = (300/10)

Valor de Pertinência Salário = 20

Cálculo do Limite de Crédito

Limite de Crédito Sugerido = ((VP_Idade + VP_Resid + VP_Veiculo
+ VP_Instruc + VP_EstCivil + VP_TpEmpr
+ VP_Sexo + VP_Filhos) /800) * ((VP_Salario
* Salário do Cliente) / 100)) * 12);

Limite de Crédito Sugerido = (((5 + 0 + 100 + 100 + 53 + 100 + 0 + 100) /800)
* ((20 * 2.900) / 100) * 12);

Limite de Crédito Sugerido = R$ 3.985,00

7 CONCLUSÕES E SUGESTÕES

7.1 CONCLUSÕES

A proposta do presente trabalho foi de demonstrar a utilização da tecnologia dos sistemas especialistas, unindo a teoria dos conjuntos difusos e técnicas de *data mining*, aplicados na área de análise de crédito para pessoas físicas, objetivando auxiliar o profissional responsável por esta tarefa a definir um valor de Limite de Crédito.

O sistema especialista no domínio do conhecimento no qual foi construído demonstrou que os objetivos pretendidos foram alcançados, mostrou-se muito útil a união da teoria dos conjuntos difusos e técnicas de *data mining*. A soma desta tecnologias demonstrou a possibilidade de utilização de sistemas especialistas na área de análise de crédito para pessoas físicas, na forma de uma ferramenta de apoio à decisão, liberando o especialista para outras atividades correlacionadas.

Constatou-se que alguns problemas do mundo real tem revelado a necessidade de tratar dados imprecisos e qualificativos. Mostra-se neste campo, uma crescente aplicação da lógica difusa. As perspectivas neste área são promissoras já que sistemas especialistas podem fundir-se a técnicas de *data mining* e à teoria dos conjuntos difusos para solucionarem problemas cada vez mais complexos.

7.2 LIMITAÇÕES

O protótipo construído possui algumas limitações, como geralmente acontece em trabalhos desenvolvidos em um espaço curto de tempo como o definido para este estudo. Pode-se citar as seguintes limitações:

a) uso de poucas variáveis na definição do perfil do cliente ideal desejado;

b) não implementação de módulo de manutenção das regras, acarretando necessidade de alteração do código fonte do protótipo;

c) não foi realizado estudo comparativo sobre as técnicas de *data mining* a fim de identificar qual se mostraria mais adequada ao projeto.

7.3 SUGESTÕES PARA TRABALHOS FUTUROS

Para as tecnologias apresentadas neste estudo, inúmeros caminhos poderão se abrir, mostrando um vasto número de aplicações possíveis. No caso específico deste trabalho, poderia ser feito um estudo mais aprofundado sobre todo o processo de análise de crédito.

Poderia ser feito um estudo mais detalhado sobre as técnicas de *data mining*, especificamente sobre análise de seleção estatística, aprimorando o processo aqui implementado.

Quanto ao uso de lógica difusa associada em um sistema especialista, poderia ser realizado um estudo comparando os diversos métodos de desfusificação existentes.

APÊNDICE 1 - REGRAS DE INFERÊNCIA

Regra 1
SE Idade do Cliente = Idade do Perfil
ENTÃO Valor de Pertinência = 100
SENÃO SE Idade do Cliente > Idade do Perfil
ENTÃO Valor de Pertinência = ((Idade do Cliente - Idade do Perfil) /Idade do Perfil)*100
SENÃO Valor de Pertinência = ((Idade do Perfil - Idade do Cliente) /Idade do Cliente)*100

Regra 2
SE Residência do Cliente = Residência do Perfil
ENTÃO Valor de Pertinência = 100
SENÃO Valor de Pertinência = 0

Regra 3
SE Veiculo do Cliente = Veiculo do Perfil
ENTÃO Valor de Pertinência = 100
SENÃO Valor de Pertinência = 0

Regra 4
SE Estado Civil do Cliente = estado Civil do Perfil
ENTÃO Valor de Pertinência = 100
SENÃO Valor de Pertinência = 0

Transformação das variáveis lingüísticas de instrução em valores
SE Instrução do Cliente = 'Analfabeto'
ENTÃO Valor Lingüístico = 0;

SE Instrução do Cliente = 'Primário'
ENTÃO Valor Lingüístico = 5;

SE Instrução do Cliente = 'Ginásio'
ENTÃO Valor Lingüístico = 8;

SE Instrução do Cliente = 'Segundo grau'
ENTÃO Valor Lingüístico = 11;

SE Instrução do Cliente = 'Terceiro grau'
ENTÃO Valor Lingüístico = 16;

SE Instrução do Cliente = 'Pós-graduação'
ENTÃO Valor Lingüístico = 18;

SE Instrução do Cliente = 'Mestrado'
ENTÃO Valor Lingüístico = 21;

SE Instrução do Cliente = 'Doutorado'
ENTÃO Valor Lingüístico = 24;

Regra 5
SE Instrução do Cliente = Instrução do Perfil
ENTÃO Valor de Pertinência = 100
SENÃO Valor de Pertinência = (Valor Lingüístico/24)*80

Transformação das variáveis lingüísticas de tempo de emprego em valores
SE Tempo de Emprego do Cliente = 'Até um ano'
ENTÃO Valor Lingüístico = 2;

SE Tempo de Emprego do Cliente = 'Um a dois anos'
ENTÃO Valor Lingüístico = 4;

SE Tempo de Emprego do Cliente = 'Dois a cinco anos'
ENTÃO Valor Lingüístico = 6;

SE Tempo de Emprego do Cliente = 'Cinco a dez anos'
ENTÃO Valor Lingüístico = 8;

SE Tempo de Emprego do Cliente = 'Mais de dez anos'
ENTÃO Valor Lingüístico = 10;

Regra 6
SE Tempo de Emprego do Cliente = Tempo de Emprego do Perfil
ENTÃO Valor de Pertinência = 100
SENÃO Valor de Pertinência = (Valor Lingüístico/10)*80

Regra 7
SE Sexo do Cliente = Sexo do Perfil
ENTÃO Valor de Pertinência = 100
SENÃO Valor de Pertinência = 0

Transformação das variáveis lingüísticas de número de filhos em valores
SE Número de Filhos do Cliente = 'Nenhum'
ENTÃO Valor Lingüístico = 2;

SE Número de Filhos do Cliente = 'Um'
ENTÃO Valor Lingüístico = 4;

SE Número de Filhos do Cliente = 'Dois'
ENTÃO Valor Lingüístico = 6;

SE Número de Filhos do Cliente = 'Três'
ENTÃO Valor Lingüístico = 8;

SE Número de Filhos do Cliente = 'Mais de três'
ENTÃO Valor Lingüístico = 10;

Regra 8
SE Número de Filhos do Cliente = 'Nenhum'
ENTÃO Valor de Pertinência = 100
SENÃO Valor de Pertinência = 100 - ((Valor Lingüístico/10)*80)

Regra 9
SE Salário do Cliente < 500
ENTÃO Valor de Pertinência = (300/30)
SENÃO SE Idade do Cliente < 3000
  ENTÃO Valor de Pertinência = (300/15)
  SENÃO Valor de Pertinência = (300/10)

Cálculo do Limite de Crédito
Limite de Crédito Sugerido = ((VP_Idade + VP_Resid + VP_Veiculo + VP_Instruc
+ VP_EstCivil + VP_TpEmpr + VP_Sexo + VP_Filhos)
/800) * ((VP_Salario * Salário do Cliente) / 100)) * 12);

REFERÊNCIAS BIBLIOGRÁFICAS

[ALM92] ALMEIDA, Hamilton. Políticas econômicas serão iguais até 95. **Zero Hora**, Porto Alegre, 24 mai 1992.

[AVI98] ÁVILA, Bráulio Coelho. Data Mining. In: **VI ESCOLA REGIONAL DE INFORMÁTICA DA SBC**. Curitiba : Champagnat, 1998. p. 87-106.

[BER97] BERRY, Michael J. A.; LINOFF, Gordon. **Data mining techniques.** USA : Wiley Computer Publishing, 1997.

[CAN98] CANTU, Marco. **Dominando o Delphi 4 - A Bíblia.** São Paulo : Makron Books, 1998.

[FAY96] FAYYAD, Usama M... [et all]. **Advances in knowledge discovery and data mining.** Mento Park : AAAI : MIT, 1996.

[FIG98] FIGUEIRA, Rafael Medeiros Andrade. **Miner: um software de inferência de dependências funcionais.** Rio de Janeiro, 1998. Trabalho de Conclusão de Curso – Instituto de Matemática, Universidade Federal do Rio de Janeiro.

[HAR88] HARMON, Paul; KING, David. **Sistemas Especialistas.** Rio de Janeiro : Editora Campus, 1988.

[HAR98] HARRISON, Thomas H. **Intranet data warehouse.** São Paulo : Berkeley Brasil, 1998.

[HEI95] HEINZLE, Roberto. **Protótipo de uma ferramenta para criação de sistemas especialistas baseados em regras de produção.** Florianópolis : UFSC, 1995. Dissertação de Mestrado, Universidade Federal de Santa Catarina, Programa de Pós-Graduação em Engenharia de Produção e Sistemas.

[LAP93] LAPOLLI, Flávio Rubens. **Sistema especialista difuso para controle de estações de tratamento de esgotos pelo processo de iodos ativados.** Florianópolis : UFSC, 1993. Dissertação de Mestrado, Universidade Federal

de Santa Catarina, Programa de Pós-Graduação em Engenharia de Produção e Sistemas.

[LEM76] LEME, Ruy Aguiar da Silva. **Projeção da demanda.** São Paulo: Fundação Vanzolini, 1976.

[LEV88] LEVINE, Robert I.; DRANG, Diane E.; EDELSON, Barry. **Inteligência artificial e sistemas especialistas.** São Paulo: McGraw-Hill, 1988.

[PAC91] PACHECO, Roberto C.S.. **Tratamento de imprecisão em sistemas especialistas.** Florianópolis : UFSC, 1991. Dissertação de Mestrado, Universidade Federal de Santa Catarina, Programa de Pós-Graduação em Engenharia de Produção e Sistemas.

[PER95] PEREIRA, Cledy Gonçalvez. **Um sistema especialista com técnicas difusas para os limites da agência.** Florianópolis : UFSC, 1995. Dissertação de Mestrado, Universidade Federal de Santa Catarina, Programa de Pós-Graduação em Engenharia de Produção e Sistemas.

[PET73] PETERS, William Stanley; SUMMERS, George W.. **Análise estatística e processo decisório.** Rio de Janeiro: Fundação Getúlio Vargas em convêncio com Instituto Nacional do Livro - MEC e Editora da Universidade de São Paulo, 1973.

[RAB95] RABUSKE, Renato Antônio. **Inteligência artificial.** Florianópolis: Editora da UFSC, 1995.

[RAU96] RAUTENBERG, Sandro. **Um protótipo de sistema especialista difuso para definição de salários por habilidades**. Blumenau, 1996. Trabalho de Conclusão de Curso (Bacharelado em Ciências da Computação) - Centro de Ciências Exatas e Naturais, FURB.

[RIB87] RIBEIRO, Horácio da Cunha e Souza. **Introdução aos sistemas especialistas.** Rio de Janeiro - São Paulo : Livros Técnicos e Científicos Editora, 1987.

[ROS95] ROSS, Timothy Jack. **Fuzzy logic with Eengineering applications.** Norwell : McGraw-Hill, 1995.

[SEC96] SECURATO, José Roberto. **Decisões financeiras em condições de risco**. São Paulo : Atlas, 1996.

[SOL81] SOLOMON, Erza, PRINGLE, John J.. **Introdução à administração financeira.** São Paulo : Atlas, 1981.

[WEL94] WELSTEAD, Stephen T. **Neural network and fuzzy logic applications in C/C++.** New york: John Willey & Sons, 1994.