UNIVERSIDADE FEDERAL DO MARANHÃO
CURSO DE PÓS-GRADUAÇÃO EM ENGENHARIA DE ELETRECIDADE

ANDERSON ARAÚJO CASANOVA

MINERAÇÃO DE DADOS:
ALGORITMO DA CONFIANÇA INVERSA

São Luís
2005

ANDERSON ARAÚJO CASANOVA

MINERAÇÃO DE DADOS:
ALGORITMO DA CONFIANÇA INVERSA

Dissertação apresentada ao curso de Pós–Graduação em Engenharia Elétrica da Universidade Federal do Maranhão como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação.

Orientador: Professor Dr. Sofiane Labidi.

São Luís

2005

Casanova, Anderson Araújo

Mineração de dados: algoritmo da confiança inversa / Anderson Araújo Casanova. – São Luís, 2005.

74f.

Dissertação (Mestrado em Ciência da Computação) – Universidade Federal do Maranhão, 2005.

1. Algoritmo 2. Algoritmo da confiança inversa 3. Lógica nebulosa 4. Similaridade I. Título

CDU 004.421

ANDERSON ARAÚJO CASANOVA

MINERAÇÃO DE DADOS:
ALGORITMO DA CONFIANÇA INVERSA

Dissertação apresentada ao curso de Pós-Graduação em Engenharia Elétrica da Universidade Federal do Maranhão como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação.
Área de concentração: Ciência da Computação.

Aprovada em / /

BANCA EXAMINADORA

_____

Prof. Dr. Sofiane Labidi (Orientador)
Universidade Federal do Maranhão

_____

Prof. Dr. Zair Abdelouahab
Universidade Federal do Maranhão

_____

Prof. Dr. Pedro Porfírio Muniz Farias
Universidade de Fortaleza

“Não se pode ensinar tudo a alguém, pode-
se apenas ajudá-lo a encontrar por si
mesmo”.
- *Galileu Galilei*

Para meus pais

AGRADECIMENTOS

A DEUS por tudo, principalmente pela família que tenho.

A meus pais, Angela e Jairo Casanova, alicerces da minha vida. Mãe sem você nada disso seria possível.

Aos meus avós que ajudaram na minha formação.

A minha namorada, Mirella, pela compreensão e pelo apoio.

A minha irmã, Andreza e sobrinhos Vitor e Vinícius.

Ao professor Dr. Sofiane Labidi pela orientação para realização deste trabalho.

A Antonio Luna e família, e Fernando e família por terem sido amigos e conselheiros.

Ao amigo Ewaldo Eder que esteve junto durante toda essa caminhada.

Aos amigos da FAMA que me ensinaram e me ajudaram em mais essa etapa da minha vida.

Aos amigos, Alfredo, Emanoel e Fabiano.

Ao professor Ociran pela ajuda quanto ao referencial teórico.

A professora Liratelma pela revisão textual.

Aos funcionários do Hospital Universitário Presidente Dutra pela paciência e pelo bom atendimento.

Aos funcionários da Coordenadoria de Pós-Graduação e em especial ao Alcides, pelos bons serviços oferecidos e que foram fundamentais para a realização deste trabalho.

Aos meus alunos com quem aprendo mais a cada dia.

SUMÁRIO

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE GRÁFICOS

LISTA DE SIGLAS

ACI – Algoritmo da Confiança Inversa.

CI – Confiança Inversa.

HC-FMUSP – Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo.

HUUFMA – Hospital Universitário da Universidade Federal do Maranhão.

KDD – *Knowledge Discovery in Databases*.

VI – Valor Intervalo.

SUS – Sistema Único de Saúde.

UDP – Unidade Presidente Dutra.

UMI – Unidade Materno Infantil.

RESUMO

Este trabalho apresenta estudos que culminaram no desenvolvimento de um algoritmo de mineração de dados que, faz extração de conhecimento e que possibilita um melhor aproveitamento das informações coletadas. Decisões baseadas em informações imprecisas e com falta de critérios podem fazer com que recursos, de qualquer tipo, sejam mal aplicados. A informação necessária que tornem a aplicação dos recursos mais justa e eficiente, e que facilitem o trabalho tanto dos usuários de um determinado serviço quanto aos que prestam o serviço, devem ser baseadas considerando a grande variedade de critérios estabelecidos. A tomada de decisão deve ser com base na avaliação dos mais variados tipa de dados e analisada por especialistas que julguem quais as necessidades, para que os critérios de busca do conhecimento sejam definidos. O Algoritmo da Confiança Inversa – ACI realiza mineração de dados utilizando a técnica de regras de associação e propõe uma nova medida que amplia a dimensão das informações extraídas através de cinco regras fixas. O ACI também classifica e associa itens similares, utilizando o conceito da lógica nebulosa (*fuzzy logic*), através de parâmetro estabelecido pelo usuário. O ACI foi aplicado no centro cirúrgico do HUUFMA – Hospital Universitário da Universidade Federal do Maranhão visando à extração de conhecimento (padrões).

Palavras-chave: mineração de dados, regras de associação, lógica nebulosa, similaridade e algoritmo da confiança inversa.

ABSTRACT

This work presents studies that culminated in the development of a data mining algorithm that extracts knowledge in a more efficient way and allows for a better use of the collected information. Decisions based on imprecise information and a lack of criteria can cause the relatively few resources available to be poorly applied, burdening taxpayers and consequently the state. This much-needed information which allows for the fairest and most efficient application of available resources and which would facilitate the work of the users as well as those who render the services should be based upon consideration of the great variety of established criteria. The making of a decision should be based upon the evaluation of the most varied types of data and be analyzed by specialists who can judge which are true needs, so that the criteria for the search of knowledge may be defined. The Algorithm of Inverse Confidence - ACI accomplishes data mining using the technique of association rules, and it proposes a new measure that enlarges the dimension of extracted information through five fixed rules. ACI also classifies and associates items, using the concept of the fuzzy logic, through parameters established by the user. ACI was applied in the surgical center of HUUFMA - Academical Hospital of the Federal University of Maranhão - envisioning the extraction of knowledge (standards).

Keywords: data mining, association rules, fuzzy logic, similarity, and algorithm of the inverse confidence.

1 INTRODUÇÃO

A computação, há muito tempo, vem evoluindo tecnologias que estão contribuindo para melhoria no tratamento de informações, fazendo com que se tenha um maior controle dos acontecimentos e processos do dia-a-dia de qualquer instituição. Na área de saúde (que foi alvo da aplicação do ACI), a tecnologia da informação já se faz presente em muitos locais, mas ainda não o bastante para fazer com que todas as informações importantes possam ser coletadas, em meio a enorme quantidade de dados que são gerados diariamente.

Nos dias de hoje, toda instituição necessita utilizar seus recursos com maior prudência (seja esse recurso financeiro ou humano), para um melhor atendimento às necessidades da sociedade. No que se refere à saúde pública, os problemas causados pela má administração dos recursos destinados às instituições de saúde pública, são muitos. As causas que levam a culminação desses problemas são as mais variadas. Dentre eles podemos citar algumas como: utilização indevida de materiais, falta de pessoal capacitado para realização dos trabalhos, ausência de informações precisas necessárias para tomada de decisões importantes, processos de trabalho mal definidos, fraudes, entre outros motivos que circulam o cotidiano da saúde pública brasileira.

O Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo (HC-FMUSP), o maior complexo hospitalar da América Latina, tem mais de 2.200 leitos ativos, e movimentação ambulatorial que chega a fazer mais de um milhão de consultas por ano. No complexo do HC-FMUSP circulam diariamente quase 35 mil pessoas e seu orçamento anual atinge a cifra de 250 milhões de dólares. O número total de prontuários ativos no HC-FMUSP está estimado entre um e dois milhões e cresce na razão de 50.000 novos prontuários por mês. Portanto, o arquivo do HC-FMUSP renova-se a cada dois a três anos. Se imaginarmos que o

volume de atendimento tem sido o mesmo por, pelo menos 20 anos, mais de 12 milhões de prontuários estiveram de posse da instituição neste período. Estes documentos encontram-se armazenados na forma tradicional de pastas com vários milhões de documentos individuais, a grande maioria dos quais manuscritos, quase sempre de maneira ilegível. Assim é que, por não ter um sistema de armazenamento eletrônico de dados médicos, instituições como o HC-FMUSP joga fora, literalmente, uma quantidade inimaginável de conhecimento médico institucional. (Massad, 2003).

1.1 Justificativa

Gerenciar instituições de grande porte sem algum tipo de suporte pode se tornar uma tarefa humanamente impossível, sem o auxílio das ferramentas de sistema de informação. Além do mais, deixar de utilizar ferramentas de extração de conhecimento nos dias atuais fará com que se deixe de tomar decisões importantes. Utilizando um algoritmo de mineração de dados, pretende-se extrair conhecimento e fazer com que decisões sejam tomadas baseadas em informações mais precisas.

1.2 Objetivo Geral

Propor e implementar um algoritmo de mineração de dados, utilizando os conceitos de regras de associação e lógica nebulosa, algoritmo esse que fará extração de informações nos dados que foram gerados em um período de tempo e que auxilie na tomada de decisão.

1.3 Objetivos Específicos

1. Apresentar os principais conceitos utilizados e definir as funcionalidades básicas sobre o funcionamento do algoritmo;
2. Definir um repositório de dados históricos e que contenha informações relevantes para a tomada de decisões;
3. Apresentar o algoritmo de mineração proposto, assim como o seu funcionamento;
4. Extrair informações úteis de um repositório utilizando o algoritmo proposto;
5. Demonstrar e analisar as informações extraídas da base de dados.

1.4 Organização da dissertação

A dissertação tem a seguinte organização: o primeiro capítulo apresentou a introdução, assim como justificativa, objetivos geral e específicos.

No segundo capítulo apresentaremos a revisão bibliográfica necessária para o embasamento da solução proposta, começando com a mineração de dados, seguido pelo capítulo três com as regras de associação e o capítulo quatro com os conceitos da lógica nebulosa (*fuzzy logic*) que fará o trabalho de acabamento final na coleta das regras.

O quinto capítulo apresenta o Algoritmo da Confiança Inversa – ACI e todo o seu funcionamento a extração das regras e a classificação das mesmas através da lógica nebulosa.

O sexto capítulo descreve o ambiente experimental, o HUUFMA – Hospital Universitário da Universidade Federal do Maranhão – todos os parâmetros utilizados no teste juntamente com os resultados obtidos com a aplicação do algoritmo.

No capítulo sete apresenta trabalhos relacionados ao trabalho proposto.

Por fim, no capítulo oito, são apresentadas as conclusões sobre este trabalho de pesquisa e as sugestões para trabalhos futuros.

2 MINERAÇÃO DE DADOS

Com o passar dos anos, desde que a informática entrou em nosso cotidiano, imensos volumes de informação são manipulados, ou seja, coletados e armazenados. O simples fato de armazenar e recuperar essas informações já traz um grande benefício, pois agora já não é mais necessário procurar informação em volumosos e ineficazes arquivos de papel. Contudo, apenas o fato de se recuperar informações tornou-se insuficiente. O processo de mineração de dados propicia a procura de padrões que auxiliem nas decisões do dia a dia (Navega, 2002). Neste capítulo serão apresentados os principais conceitos sobre a mineração de dados assim como algumas técnicas e algoritmos de mineração.

De acordo com Han (2000), a mineração de dados surgiu no início da década de oitenta e seguiu a passos largos durante a década de noventa podendo ser visto como resultado da evolução natural da tecnologia da informação. Ainda segundo Han (2000), a mineração de dados é um campo multidisciplinar, que envolve trabalhos nas áreas de tecnologia de banco de dados, inteligência artificial, aprendizagem de máquina, redes neurais, estatística, recuperação da informação, computação de alto desempenho e visualização de dados.

Quanto a Wang (2003), a mineração de dados é a extração da informação preditiva em grandes bases de dados.

Já para Rud (2001), mineração de dados é um termo que cobre uma vasta gama de técnicas utilizadas pelas mais variadas indústrias.

Para Bose (1999), mineração de dados é o processo de descobrir significantes correlações, padrões e tendências peneirando em grande quantidade de dados armazenados em *data warehouse*, utilizando não só tecnologias de reconhecimento de padrões, como também técnicas estatísticas e matemáticas. Mineração de dados provê um meio de extrair informações desconhecidas das crescentes bases de dados acessíveis em *data warehouse* e

cria vantagens competitivas para as organizações. As informações são retiradas do *data warehouse* utilizando técnicas de análise avançadas tais como de redes neurais, heurística, raciocínio indutivo e lógica nebulosa.

Para Han (2000), muitas pessoas tratam o termo mineração de dados como um sinônimo para o termo "*Descoberta do Conhecimento em Base de Dados"- (Knowledge Discovery in Databases – KDD).*

Alternativamente, outros vêem a mineração de dados como simplesmente uma etapa essencial no processo da descoberta de conhecimento em bases de dados. A descoberta de conhecimento é um processo descrito como na figura 1, e consiste em uma seqüência iterativa das seguintes etapas:

- Limpeza dos dados: nessa etapa, rotinas são utilizadas para tratar valores nulos, e corrigir dados inconsistentes;
- Integração dos dados: onde múltiplas fontes de dados podem ser combinadas. Durante essa etapa os seguintes aspectos são considerados: a integração de esquema, a detecção de redundâncias (duplicação) tanto de atributos quanto de tuplas e a existência de conflitos de dados;
- Seleção de dados: seleciona na base de dados os dados relevantes para análise;
- Transformação dos dados: transformação ou consolidação dos dados em formatos apropriados para o processo de mineração de dados;
- Mineração de dados: aplicação de técnicas específicas para extração de padrões de acordo com o tipo de conhecimento a ser minerado;
- Avaliação dos padrões: identificação da importância dos padrões encontrados.
- Apresentação do conhecimento: uso de técnicas de visualização e representação do conhecimento para apresentar ao usuário o conhecimento obtido.

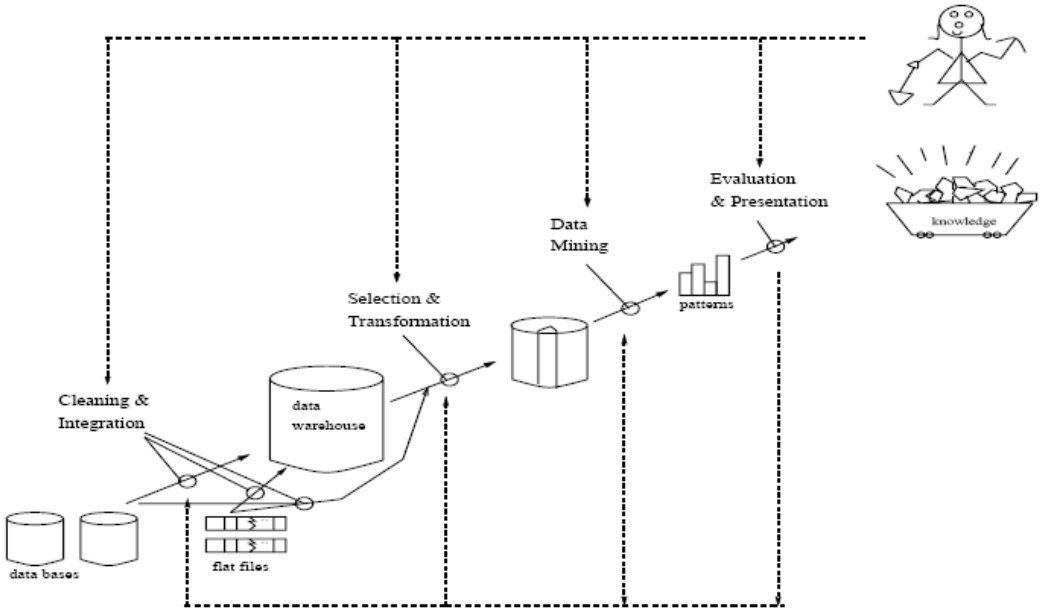

Figura 1. *Data Mining* e o processo de descoberta do conhecimento – KDD

Fonte: Han, 2000.

Na prática, Kantardzic (2003), descreve dois objetivos primários do *data mining*: a predição e a descrição. A predição envolve o uso de algumas varáveis ou campos na série de dados para predizer valores desconhecidos ou futuros de outras variáveis de interesse. Por outro lado, a descrição focaliza encontrar padrões de descrição que podem ser interpretados pelos seres humanos.

A mineração de dados tornou-se muito útil na década passada no que se refere à aquisição de mais informações, compreender melhor o funcionamento de um determinado negócio e encontrar maneiras novas de inseri-lo em novos mercados. A figura 2 mostra uma visão histórica da mineração de dados.

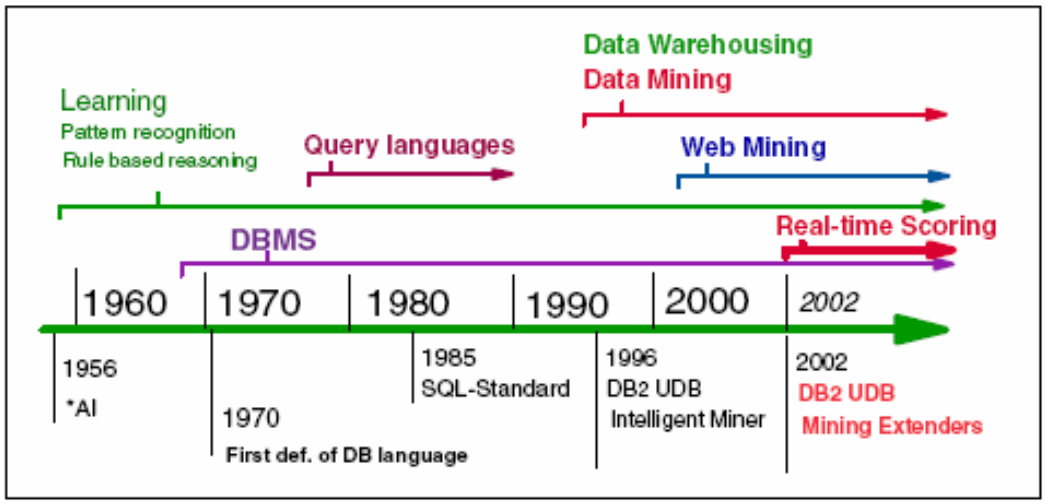

Figura 2. Visão histórica da mineração de dados.

Fonte: Baragoin, 2002

As técnicas de mineração de dados mais usuais são: árvores de decisão, redes neurais artificiais, algoritmos genéticos, lógica nebulosa e regras de associação. A técnica de regras de associação utilizada no Algoritmo da Confiança Inversa será abordada na próxima seção 3.

2.1 Técnicas de mineração de dados

Tem-se percebido um grande crescimento, tanto na elaboração e aperfeiçoamento das técnicas de mineração de dados quanto na utilização destas nas mais variadas áreas. A classificação pode ser representada pelas técnicas utilizadas ou de maneira mais abrangente, como proposto por Chen (1996), que um sistema de mineração de dados pode ser classificado de acordo com os seguintes critérios:

- Tipos de base de dados: os sistemas de mineração de dados podem ser classificados segundo o tipo da base de dados em que estão sendo executados, ou seja, se esse sistema é um minerador de dados relacional quando executado sobre uma base de dados relacional, ou um minerador de dados orientado a objetos se executado sobre uma base orientada a objetos;
- Tipos de conhecimento: existem dois modelos a considerar. Um preocupado com o conhecimento propriamente dito, incluindo regras de associação, regras de classificação, agrupamentos, e outro preocupado com o nível de abstração do conhecimento descoberto;
- Tipos de técnicas: a escolha da técnica está fortemente relacionada com o tipo de conhecimento que se deseja extrair ou com os dados nos quais se aplicam tais técnicas. Entretanto, nota-se uma visão mais genérica, em que as técnicas são caracterizadas em mineração baseada na generalização, em padrões e na estatística.

O progresso da área de extração de conhecimento de bases de dados e sua utilização nos mais variados domínios e pelas mais diversas organizações têm motivado o desenvolvimento de várias ferramentas comerciais além da elaboração de muitos protótipos de pesquisa. O processo de extração de conhecimento de bases de dados é facilitado consideravelmente se for usada uma ferramenta que ofereça suporte para uma variedade de técnicas, com diferentes algoritmos e voltadas para várias tarefas de mineração de dados.

O desenvolvimento de ferramentas comerciais de mineração de dados tem como objetivo principal fornecer aos tomadores de decisão das organizações, que são usuários geralmente não especialistas em mineração de dados, ferramentas intuitivas e amigáveis. É interessante que estas ferramentas ofereçam suporte às várias etapas do processo de mineração de dados assim como disponibilizem apoio para diversas técnicas e tarefas.

No quadro 01, Rezende (2003), apresenta algumas ferramentas comerciais e protótipos disponíveis no mercado e suas respectivas características. No quadro *específico* representa uma ferramenta para apoiar uma tarefa específica que não possui a generalidade e flexibilidade encontrada nos pacotes. Assim, por serem mais restritas, as ferramentas específicas tendem a ser mais simples e fáceis de serem entendidas.

NOME	TÉCNICAS DISPONÍVEIS	FABRICANTE SITE	TIPO DE APLICATIVO
PolyAnalyst	Classificação, regressão, regras de associação, clustering, sumarização e modelagem de dependências.	Megaputer Intelligence www.megaputer.com	Pacote
Magnum Opus	Regras de associação	Rule Quest www.rulequest.com	Específico
XpertRule Miner	Classificação, regras de associação e clustering.	Attar Software Ltd. www.attar.com	Pacote
DataMite	Regras de associação	Dr Philip Vasey através do LPA prolog.	Específico
Microsoft Data Analyzer 2002	Classificação e clustering	Microsoft Corporation www.microsoft.com	Pacote
Oracle 9i Data Mining	Classificação e regras de associação	Oracle Corp. www.oracle.com	Pacote
Darwin	Classificação, regressão e clustering.	Oracle Corp. www.oracle.com	Pacote
MineSet	Classificação, regressão, regras de associação e clustering.	Silicon Graphics Inc. www.sgi.com	Pacote
WEKA	Classificação, regressão e regras de associação.	University of Waikato www.cs.waikato.ac.nz	Pacote

Intelligent Miner	Regras de associação, padrões seqüenciais, classificação, clustering, sumarização e modelagem de dependência.	IBM Corp. www.ibm.com	Pacote
MLC++	Classificação, regressão e clustering.	Silicon Graphics Inc. www.sgi.com/tech/mlc	Biblioteca
See5	Classificação	Rule Quest www.rulequest.com	Específico
Cubist	Regressão	Rule Quest www.rulequest.com	Específico
Clementine	Classificação, regras de associação, clustering e padrões seqüenciais.	SPSS Inc. www.spss.com	Pacote
Data-Miner Software Kit	Classificação e regressão.	Data-Miner PTy LTd www.data-miner.com	Específico

Quadro 01. Algumas ferramentas para mineração de dados.
Fonte: Rezende, 2003

3 REGRAS DE ASSOCIAÇÃO

Uma importante área de pesquisa em mineração de dados trata da descoberta de regras de associação, que descrevem relações de associação entre diferentes atributos (Mitra, 2003). Esta seção provê a introdução para regras de associação.

Uma regra de associação é uma expressão da forma

$$X \rightarrow Y,$$

onde X e Y são conjuntos de itens. O significado de tal regra é que transações da base de dados que contém X tendem a conter Y também. O conjunto de itens que aparece à esquerda da seta (representado por X) é chamado de *antecedente* da regra. Já o conjunto de itens que aparece à direita da seta (representado por Y) é o *conseqüente* da regra. Assim, uma regra de associação tem o seguinte formato:

$$\text{Antecedente} \rightarrow \text{Conseqüente}$$

A cada regra são associados dois fatores: *suporte* e *confiança*. Para uma regra de associação $X \rightarrow Y$, o suporte indica a porcentagem de registros em que aparecem X e Y simultaneamente, sobre o total de registros. Já a confiança indica a porcentagem de registros que contém X e Y, sobre o total de registros que possuem X.

Um conjunto de itens é chamado de **itemset** e seu suporte é a porcentagem das transações que contêm todos os itens do **itemset**. Um **itemset** é dito freqüente quando o seu suporte é maior ou igual a um valor de suporte mínimo definido pelo usuário (Escovar, 2004).

O problema de descobrir todas as regras de associação pode ser decomposto em duas etapas:

- Encontrar todos os conjuntos de itens (**itemsets**) que apresentam suporte maior que o suporte mínimo estabelecido pelo decisor. Os **itemsets** que atendem a este quesito são denominados **itemsets** freqüentes; e
- Utilizar os **itemsets** freqüentes obtidos para gerar as regras de associação do banco de dados.

Essa técnica pode ser usada em diversas aplicações, como a análise do carrinho de compras (*market basket analysis*). Neste tipo de aplicação há o interesse em descobrir associações entre os itens comprados pelos consumidores, ou seja, quais produtos são comprados juntos com outros produtos. Um exemplo típico de uma regra de associação que pode ser extraída de uma base de dados de um supermercado é a seguinte: quando um consumidor compra pão ele também compra leite (pão $\rightarrow$ leite). Essas informações podem ser utilizadas, por exemplo, para dispor os itens que são freqüentemente comprados juntos nas prateleiras, de maneira a encorajar a venda dos mesmos (Barioni, 2002).

4 LÓGICA NEBULOSA

O termo *fuzzy* em língua inglesa pode ter diversos significados, de acordo com o contexto de interesse, mas o conceito básico deste adjetivo passa sempre pelo vago, indistinto, incerto. As tentativas de tradução para o português ainda não são uma unanimidade: nebuloso e difuso são os exemplos mais populares na área de engenharia. Sistemas de apoio à decisão, algoritmos para aproximação de funções e sistemas de controle baseados em lógica *fuzzy* estão entre as formas mais populares da utilização desses conceitos (Rezende, 2003).

A Lógica Nebulosa baseia-se na teoria de conjuntos nebulosos (*Fuzzy Sets*), e teve seus conceitos e princípios introduzidos por Zadeh na década de 60 (Escovar apud Zadeh, 1978a). Ela trata, matematicamente, informações imprecisas usualmente empregadas na comunicação humana, permitindo inferir uma resposta aproximada para uma questão baseada em um conhecimento que é inexato, incompleto ou não totalmente confiável. É nesta natureza da informação que reside a nebulosidade.

Enquanto na lógica booleana, usualmente empregada em computação, são definidos apenas dois valores possíveis — verdadeiro (1) ou falso (0) — a Lógica Nebulosa é multivalorada (ou seja, há um conjunto de valores possíveis) e nesse aspecto ela pode ser considerada uma extensão da primeira. (Escovar, 2004).

Na lógica nebulosa os valores são expressos lingüisticamente (verdade, muito verdade, não verdade, falso, muito falso), onde cada termo lingüístico é interpretado como um subconjunto nebuloso do intervalo unitário.

Outras características da lógica nebulosa podem ser resumidas da seguinte maneira: nos sistemas lógicos binários, os predicados são exatos (par, ímpar, maior que, menor que), ao passo que na lógica nebulosa os predicados são nebulosos (alto e baixo). Nos sistemas lógicos

clássicos, o modificador mais utilizado é a negação, enquanto que na lógica nebulosa uma variedade de modificadores é possível (muito, mais ou menos). Estes modificadores são essenciais na geração dos termos lingüísticos, tais como: muito alto, mais ou menos perto, etc. Deste modo a decisão não se resume a um "sim" ou um "não", mas também tem decisões abstratas, sendo a lógica nebulosa uma técnica inteligente que fornece um mecanismo para manipular informações imprecisas (Canoas apud Zadeh, 1965).

O quadro 02 sintetiza algumas das principais características, vantagens e desvantagens da lógica nebulosa.

CARACTERÍSTICAS	VANTAGENS	DESVANTAGENS
A Lógica Nebulosa está baseada em palavras e não em números, ou seja, os valores verdades são expressos lingüisticamente. Por exemplo: quente, muito frio, verdade, longe, perto, rápido, vagaroso, médio;	O uso de variáveis lingüísticas nos deixa mais perto do pensamento humano;	Necessitam de mais simulação e testes.
Possui vários modificadores de predicados, tais como: muito, mais ou menos, pouco, bastante, médio, etc.	Requer poucas regras, valores e decisões.	Dificuldade de estabelecer regras corretamente.
Manuseia todos os valores entre 0 e 1, tomando estes como limites apenas.	Simplifica a solução de problemas e a aquisição da base de conhecimento.	Não há uma definição matemática precisa.

Quadro 02. Características, vantagens e desvantagens da lógica nebulosa.
Fonte: Camargos, 2003.

4.1 Conjuntos Nebulosos

A função característica de um conjunto clássico pode assumir somente os valores 0 ou 1, determinando dessa forma quem são os membros e os não-membros desse conjunto. Essa função pode ser generalizada de forma que ela possa assumir valores em um determinado intervalo, e o valor assumido indica o *grau de pertinência* do elemento no conjunto em questão. Essa função é chamada de *função de pertinência*. A função de pertinência de um conjunto nebuloso *A* é denotada por μA, desta forma:

$$\mu A: X \rightarrow [0,1].$$

A função de pertinência mapeia os elementos de um conjunto clássico *X* em números reais no intervalo [0,1]. Assim, um conjunto nebuloso *A* é caracterizado por uma função de pertinência $\mu A(x)$, que associa a cada elemento do conjunto um número real no intervalo [0,1]. Desta forma, o valor de $\mu A(x)$ representa o grau de pertinência do elemento *x* no conjunto *A*. Quanto maior o valor de $\mu A(x)$, maior o grau de pertinência de *x* no conjunto *A*. As funções de pertinência podem ser representadas através de vários tipos de funções, a escolha do tipo de função depende da aplicação no qual será utilizada.

4.2 Regras de Associação Nebulosa

Regras de associação quantitativas requerem a criação de intervalos apropriados para cada atributo. Entretanto, esses intervalos podem freqüentemente não ser concisos e significativos o bastante para que especialistas humanos descubram o conhecimento não trivial. Conjuntos *fuzzy* podem ser usados para representar os intervalos, gerando, dessa forma

regras de associação nebulosa. A atribuição de termos lingüísticos significativos aos conjuntos nebulosos torna as regras mais compreensíveis. (Mitra, 2003)

4.3 Variáveis Lingüísticas

Uma variável lingüística é definida como uma entidade utilizada para representar de modo impreciso, e portanto, lingüístico, um conceito de um dado problema. Ela admite como valores apenas expressões lingüísticas (freqüentemente chamadas de termos primários), como "frio", "muito grande", "aproximadamente alto", etc. Estes valores contrastam com os valores assumidos por uma variável numérica, que admite apenas valores precisos, ou seja, números.

Um termo primário de uma dada variável lingüística pode ser representado por um conjunto *fuzzy* existente no universo de discurso no qual esta variável está definida. Assim, cada conjunto *fuzzy* definido neste universo é associado a um conceito lingüístico que classifica ou define um valor impreciso para a variável em questão. Para um dado elemento x no universo de discurso, o valor de pertinência $\mu_A(x)$ representa o quanto este elemento satisfaz o conceito representado pelo conjunto *fuzzy* A.

Os temos primários definidos para uma dada variável lingüística formam a sua estrutura de conhecimento, chamada de partição *fuzzy* desta variável. Na figura 3 é mostrado um exemplo de partição *fuzzy* de uma variável lingüística chamada "Temperatura". O universo de discurso utilizado é um segmento da escala Celsius de temperatura, entre 0 e 50 graus.

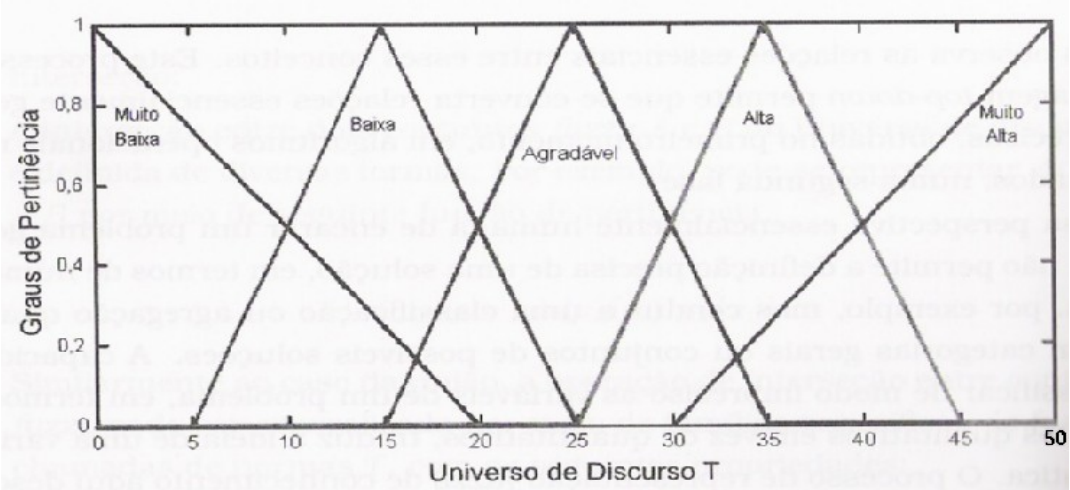

Figura 3. Partição *fuzzy* de uma variável lingüística representando a temperatura.

Fonte: Rezende, 2003.

A forma de utilização das variáveis lingüísticas depende da definição das propriedades sintáticas e semânticas que vão reger o comportamento do sistema de conhecimento *fuzzy*. As propriedades sintáticas irão definir o formato em que serão armazenadas informações lingüísticas *fuzzy*. Elas proporcionam a criação de uma Base de Conhecimento contendo sentenças estruturadas, sistematizando os processos de armazenamento, busca e processamento dos dados existentes.

Por outro lado, as propriedades semânticas vão especificar de que modo é extraído e processado o conhecimento, armazenado na forma de declarações condicionais *fuzzy*, ou regras de produção *fuzzy*, contida na estrutura definida pelas propriedades sintáticas (Rezende, 2003).

5 ALGORITMO DA CONFIANÇA INVERSA

Nesta seção será abordada a medida proposta no trabalho, assim como o funcionamento do ACI em todos os detalhes.

5.1 A Confiança Inversa

A *Confiança Inversa – CI* é a medida que corresponde a porcentagem de vezes que ocorre X e Y simultaneamente sobre o total de registros que possuem Y. A *CI* do **itemset** X $\rightarrow$ Y é a Confiança do **itemset** Y $\rightarrow$ X. Assim como a *CI* do **itemset** Y $\rightarrow$ X é a Confiança de X $\rightarrow$ Y. O fato de se utilizar uma medida, para extração de regras, que tem como denominador o antecedente X, que é a *Confiança*, e não haver nenhuma medida que utilizasse o conseqüente Y como denominador, motivou a criação da *CI* e quais os resultados que viriam a ser obtidos com a sua aplicação em uma base de dados para extração de conhecimento.

Para os testes iniciais foi criada uma base de dados, apresentada no quadro 03, com 50 registros, em que cada registro representa uma pessoa fictícia através de dois campos: um para a cidade natal e o outro para a titulação.

REGISTRO	CIDADE	TITULAÇÂO	REGISTRO	CIDADE	TITULAÇÃO
01	Imperatriz	Médio	26	São Luís	Doutor
02	Imperatriz	Médio	27	São Luís	Doutor
03	Imperatriz	Médio	28	São Luís	Doutor
04	Imperatriz	Médio	29	São Luís	Doutor
05	Imperatriz	Médio	30	São Luís	Doutor
06	Imperatriz	Médio	31	Fortaleza	Médio
07	Imperatriz	Médio	32	Fortaleza	Especialista
08	Imperatriz	Especialista	33	Fortaleza	Especialista
09	Imperatriz	Mestre	34	Fortaleza	Especialista
10	Imperatriz	Mestre	35	Fortaleza	Especialista
11	Belém	Médio	36	Fortaleza	Especialista
12	Belém	Médio	37	Fortaleza	Especialista
13	Belém	Especialista	38	Fortaleza	Mestre
14	Belém	Mestre	39	Fortaleza	Doutor
15	Belém	Mestre	40	Fortaleza	Doutor
16	Belém	Mestre	41	Natal	Médio
17	Belém	Mestre	42	Natal	Especialista
18	Belém	Mestre	43	Natal	Especialista
19	Belém	Doutor	44	Natal	Mestre
20	Belém	Doutor	45	Natal	Mestre
21	São Luís	Médio	46	Natal	Mestre
22	São Luís	Especialista	47	Natal	Mestre
23	São Luís	Especialista	48	Natal	Mestre
24	São Luís	Mestre	49	Natal	Mestre
25	São Luís	Doutor	50	Natal	Doutor

Quadro 03. Registros utilizados nos testes iniciais para testes da *CI*.

Nos testes foram escolhidos **itemsets** em que o conseqüente é fixo e o antecedente varia em todas as possibilidades que X pode assumir. Tabelas 01, 02 e 03. A seguir apresentam-se os resultados obtidos aplicando a *CI* para os **itemsets** escolhidos.

Tabela 01. Primeiro grupo de **itemsets** e os resultados obtidos aplicando a *CI*.

ITEMSET CIDADE → TITULAÇÃO	***NÚMERO DE OCORRÊNCIAS DE X***	***NÚMERO DE OCORRÊNCIAS DE Y***	***VALOR DA CI***
Imperatriz → Médio	10	7	58,33%
Belém → Médio	10	2	16,67%
São Luís → Médio	10	1	8,33%
Fortaleza → Médio	10	1	8,33%
Natal → Médio	10	1	8,33%
Total	50	12	99,99%

De acordo com os resultados obtidos através da *CI*, duas observações devem ser ressaltadas:

- Para o **itemset** Imperatriz → Médio obteve-se CI = 58,33% e observa-se na tabela 01 que a soma das pessoas que têm ensino médio da cidade de Imperatriz é maior do que a soma das pessoas que têm ensino médio de todas as outras cidades.
- Para os outros **itemsets** a soma das pessoas que têm nível médio da sua respectiva cidade é menor do que a soma das pessoas que têm nível médio das outras cidades.

De acordo com os resultados obtidos através da CI para os itemsets da tabela 02 o itemset São Luís → Doutor apresentou CI = 54,55%.

Tabela 02. Segundo grupo de **itemsets** e os resultados obtidos aplicando a *CI*.

ITEMSET ***CIDADE → TITULAÇÃO***	***NÚMERO DE OCORRÊNCIAS DE X***	***NÚMERO DE OCORRÊNCIAS DE Y***	***VALOR DA CI***
Imperatriz → Doutor	10	0	0,0%
Belém → Doutor	10	2	18,18%
São Luís → Doutor	10	6	54,55%
Fortaleza → Doutor	10	2	18,18%
Natal → Doutor	10	1	9,09%
Total	50	11	100%

Assim como ocorreu com o para o **itemset** Imperatriz → Médio na tabela 01, as mesmas observações podem ser feitas para o **itemset** São Luís → Doutor da tabela 02:

- A quantidade de doutores da cidade de São Luís é maior que a soma de doutores das outras cidades;
- Para os outros **itemset**s a soma das pessoas que têm doutorado da sua respectiva cidade é menor do que a soma das pessoas que têm doutorado das outras cidades.

De acordo com os resultados obtidos através da CI para os itemsets da tabela 03, a seguinte observação deve ser citada: para o itemset Fortaleza → Especialista observamos que a quantidade de especialistas de Fortaleza é igual à soma dos especialistas das outras cidades.

Tabela 03. Terceiro grupo de **itemsets** e os resultados obtidos aplicando a *CI*.

ITEMSET CIDADE → TITULAÇÃO	***NÚMERO DE OCORRÊNCIAS DE X***	***NÚMERO DE OCORRÊNCIAS DE Y***	***VALOR DA CI***
Imperatriz → Especialista	10	1	8,33%
Belém → Especialista	10	1	8,33%
São Luís → Especialista	10	2	16,67%
Fortaleza → Especialista	10	6	50%
Natal → Especialista	10	2	16,67%
Total	50	12	100%

Com os resultados obtidos para os **itemsets** testados, agora pode-se apresentar um significado para o valor obtido pela *CI*. Os valores que a *CI* pode assumir, as regras e os respectivos significados para um **itemset** são demonstrados no quadro 04 a seguir.

VALORES DA *CI*%	**REGRA**	**SIGNIFICADO**
100	1	Só existe ocorrência de y em x.
>50 e <100	2	A soma das ocorrências de y quando ocorre x é maior do que a soma das ocorrências de y para as outras possibilidades de x.
50	3	A soma das ocorrências de y quando ocorre x é igual a soma das ocorrências de y para as outras possibilidades de x.
>0 e < 50	4	A soma das ocorrências de y quando ocorre x é menor do que a soma das ocorrências de y para as outras possibilidades de x.
0	5	Não existem ocorrências de y em x.

Quadro 04. Valores, Regras e o significado da *CI*.

5.2 Funcionamento

Os algoritmos de mineração baseados em regras de associação buscam por padrões que possuam um mínimo de freqüência em uma base de dados. Dessa forma, se pode gerar um número de regras muito grande, ou muito pequeno, dependendo dos parâmetros utilizados na busca dessas informações. Além do mais, os algoritmos existentes trabalham com duas medidas (*Suporte* e *Confiança*) que não retiram todas as possíveis informações sobre os dados analisados.

O algoritmo realiza busca de padrões. Utiliza a técnica de regras de associação para encontrar os itemsets e lógica nebulosa para classificação de dados do tipo categórica, ou seja, tipo de algoritmo que trata os itens como simples seqüências de caracteres sem analisar o seu significado semântico.

O usuário formula hipóteses e executa testes em um banco de dados, para validar ou refutar tais hipóteses. O ACI, idealizado pelo próprio autor da dissertação, faz extração de dados guiada pelo usuário e, além das duas medidas usuais, propõe uma nova medida que, como resultado, apresenta uma regra dentre cinco possíveis, para auxiliar na coleta de conhecimento. Utilizando a lógica nebulosa que além de julgar a regra, que pode ser classificada como regra ruim, regular, boa, ótima ou excelente. Também procura regras similares de acordo com um critério pré-estabelecido.

O objetivo do ACI pode ser descrito dessa forma: dados

- Um conjunto de transações (*n*);
- A quantidade mínima do **itemset** procurado (*xy*);
- Um *Suporte* mínimo (*supmin*);
- Uma *Confiança* mínima (*confmin*);

- O **itemset$_i$** para o qual desejamos procurar regras, onde *i* é o número que corresponde à seqüência dos **itemsets** propostos pelo usuário;
- O Valor Intervalo – *VI* que é o maior valor que uma regra pode distanciar de outra regra, para ser classificada como não similar.

Obter todas as regras de associação que possuam:

- *xy* ≥ **itemset** ;
- *Confiança*$_i$ ≥ *confmin*;
- *Suporte*$_i$ ≥ *supmin*.

E encontrar também:

- A relação de Y para um ou todos os valores de X, *CI*$_i$;
- A classificação da regra (ruim, regular, boa, ótima e excelente);
- **Itemsets** propostos pelo usuário similares e a similaridade entre esses **itemsets** dentro do intervalo *VI* (não similar, pouco similar, quase similar e similar).

O ACI varre o banco de dados a procura de todas as ocorrências do **itemset** e encontra o *Suporte*, *Confiança* e a *CI*. (Figura 4, linhas 1 a 8). Caso a quantidade do **itemset$_i$** procurado seja menor que o valor de *xy*, o *Suporte* seja menor que o valor de *supmin* e a *Confiança* seja menor que o valor de *confmin* o algoritmo então descarta o **itemset$_i$** procurado e pede ao usuário um novo **itemset** (**itemset**$_{i+1}$) para uma nova pesquisa (Figura 1 linhas 9 e 10). Caso os parâmetros de *xy*, *supmin* e *confmin* sejam satisfeitos o programa então irá encontrar a regra a qual o **itemset** pesquisado pertence. A partir do valor da *CI* encontra-se a regra e o significado a qual o **itemset** pertence (Figura 5, linhas 1 e 10).

```
1)  para t ← 1 até n faça        // n é o número total de transações.
2)      cont x
3)      cont y
4)      cont itemset_i
5)      fim (para)
6)  Suporte_i ← xy * 100/n
7)  Confiança_i ← xy * 100/x
8)  CI_i ← xy * 100/y
9)  se (xy > itemset_i) e (supmin > Suporte_i) e (confmin > Confiança_i) então
10)     entrar com outro itemset (itemset_i+1) e retornar para linha 1
11) senão
12)     // encontrar a regra para o itemset proposto.
```

Figura 4. Busca do **itemset** proposto e determinação das medidas $Suporte_i$, $Confiança_i$ e CI_i.

```
1)      se (CI_i = 100) então          // encontrar a regra para o itemset_i proposto
2)              regra 1
3)      se (CI_i > 50) e (CI_i < 100) então
4)              regra 2
5)      se (CI_i = 50) então
6)              regra 3
7)      se (CI_i > 0) e (CI_i < 50) então
8)              regra 4
9)      se (CI_i = 0) então
10)             regra 5
```

Figura 5. Seleção da regra de acordo com o valor da CI_i.

As regras da *CI* são obtidas sem que se tenha a necessidade de se fazer outra varredura nos registros invertendo o **itemset** no qual estamos procurando uma regra. O ACI se baseia na *CI* para descobrir a relação entre o número de ocorrências do conseqüente (Y) para o antecedente (X).

O tópico 5.3 a seguir, descreve como o ACI classifica um **itemset** proposto pelo usuário.

5.3 Classificação de um Itemset

Com a utilização da lógica nebulosa, pode-se classificar **itemsets** através da utilização de termos lingüísticos, para melhor compreensão dos resultados. Os termos lingüísticos (ruim, regular, boa, ótima e excelente) para classificação de um **itemset** foram utilizados para expressar o conhecimento em nível da linguagem natural.

Utilizando o conceito dos conjuntos nebulosos, o ACI classifica o resultado da *CI* para o **itemset** proposto em regra ruim, regular, boa, ótima ou excelente. As funções de pertinência para cada um desses conjuntos são, respectivamente, **(μ_{Ruim} (CI_i), $\mu_{Regular}$ (CI_i), μ_{Boa} (CI_i), $\mu_{Ótima}$ (CI_i) e $\mu_{Excelente}$ (CI_i))** exibidas no quadro 05.

1	Quando $CI_i \leq 15$
$\mu_{Ruim}(CI_i) = (25 - CI_i)/10$	Quando $CI_i > 15$ e $CI_i < 25$
0	Quando $CI_i \geq 25$
0	Quando $CI_i \leq 15$ e $CI_i \geq 45$
$\mu_{Regular}(CI_i) = (CI_i - 15)/10$	Quando $CI_i > 15$ e $CI_i < 25$
$\mu_{Regular}(CI_i) = (45 - CI_i)/10$	Quando $CI_i > 35$ e $CI_i < 45$
1	Quando $CI_i \geq 25$ e $CI_i \leq 35$

0	Quando $CI_i \leq 35$ e $CI_i \geq 65$
$\mu_{Boa}(CI_i) = (CI_i - 35)/10$	Quando $CI_i > 35$ e $CI_i < 45$
$\mu_{Boa}(CI_i) = (65 - CI_i)/10$	Quando $CI_i > 55$ e $CI_i < 65$
1	Quando $CI_i \geq 45$ e $CI_i \leq 55$
0	Quando $CI_i \leq 55$ e $CI_i \geq 85$
$\mu_{Ótima}(CI_i) = (CI_i - 55)/10$	Quando $CI_i > 55$ e $CI_i < 65$
$\mu_{Ótima}(CI_i) = (85 - CI_i)/10$	Quando $CI_i > 75$ e $CI_i < 85$
1	Quando $CI_i \geq 65$ e $CI_i \leq 75$
0	Quando $CI_i \leq 75$
$\mu_{Excelente}(CI_i) = (CI_i - 75)/10$	Quando $CI_i > 75$ e $CI_i < 85$
1	Quando $CI_i \geq 85$

Quadro 05. Função de pertinência para os conjuntos regra ruim, regular, boa, ótima e excelente.

O conjunto *x*, que assume os valores da CI_i de 0 a 100% é mapeado, através da função de pertinência, em conjuntos nebulosos que assumem os valores entre [0,1]. A representação gráfica dos conjuntos nebulosos para CI_i é apresentada na figura 6.

Os valores que definem o conjunto nebuloso da confiança inversa, para as respectivas funções de pertinência, podem ser redefinidos de acordo com o propósito o qual o ACI está sendo utilizado.

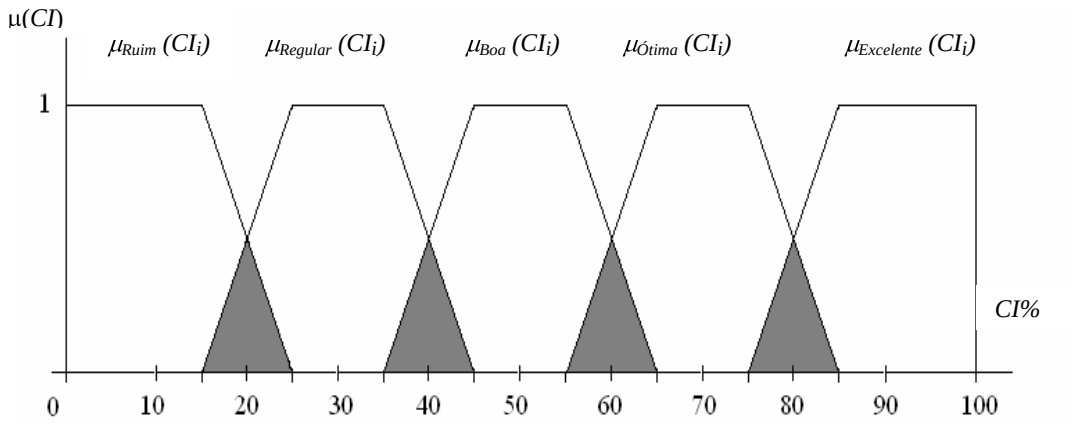

Figura 6. Representação gráfica dos conjuntos nebulosos $\mu_{Ruim}(CI_i)$, $\mu_{Regular}(CI_i)$, $\mu_{Boa}(CI_i)$, $\mu_{Ótima}(CI_i)$ e $\mu_{Excelente}(CI_i)$.

Após o **itemset** proposto pelo usuário ser classificado (figuras 7, 8, 9, 10 e 11), o algoritmo procura por outros **itemset**s que possuam um valor aproximado ao **itemset** primeiramente proposto. O próximo **itemset** a ser pesquisado pelo ACI também é proposto pelo usuário (**itemset**$_{i+1}$). Após entrar com outro **itemset,** o ACI executa para o novo **itemset** as linhas das figuras 4 e 5 para encontrar os valores de *Suporte*, *Confiança* e *CI* que será o alvo da comparação entre os **itemsets**. O tópico 5.3.2 descreve como o ACI encontra similaridade entre **itemsets**, ou seja, similaridade entre os valores da *CI* dos **itemset**$_i$ e **itemset**$_{i+1}$.

1) **Se** $CI_i <= 15$ **então**
2) Regra Ruim
3) **Se** $(CI_i > 15)$ **e** $(CI_i < 25)$ **então**
4) $(25 - CI_i)/10$
5) **Se** $CI_i >= 25$ **então**
6) **Itemset$_i$** não é classificado como regra Ruim

Figura 7. Classificação de um **itemset** no conjunto nebuloso $\mu_{Ruim}(CI_i)$.

1) **Se** $(CI_i <= 15)$ **e** $(CI_i >= 45)$ **então**
2) **Itemset$_i$** não é classificado como regra Regular
3) **Se** $(CI_i > 15)$ **e** $(CI_i < 25)$ **então**
4) $(CI_i - 15)/10$
5) **Se** $(CI_i > 35)$ **e** $(CI_i < 45)$ **então**
6) $(45 - CI_i)/10$
7) **Se** $(CI_i >= 25)$ **e** $(CI_i <= 35)$ **então**
8) Regra Regular

Figura 8. Classificação de um **itemset** no conjunto nebuloso $\mu_{Regular}(CI_i)$.

1) **Se** ($CI_i >= 65$) **e** ($CI_i <= 35$) **então**
2) **Itemset$_i$** não é classificado como regra Boa
3) **Se** ($CI_i > 35$) **e** ($CI_i < 45$) **então**
4) $(CI_i - 35)/10$
5) **Se** ($CI_i > 55$) **e** ($CI_i < 65$) **então**
6) $(65 - CI_i)/10$
7) **Se** ($CI_i >= 45$) **e** ($CI_i <= 55$) **então**
8) Regra Boa

Figura 9. Classificação de um **itemset** no conjunto nebuloso $\mu_{Boa}(CI_i)$.

1) **Se** ($CI_i >= 85$) **e** ($CI_i <= 55$) **então**
2) **Itemset$_i$** não classificado como regra Ótima
3) **Se** ($CI_i > 55$) **e** ($CI_i < 65$) **então**
4) $(CI_i - 55)/10$
5) **Se** ($CI_i > 75$) **e** ($CI_i < 85$) **então**
6) $(85 - CI_i)/10$
7) **Se** ($CI_i >= 65$) **e** ($CI_i <= 75$) **então**
8) Regra Ótima

Figura 10. Classificação de um **itemset** no conjunto nebuloso $\mu_{Ótima}(CI_i)$.

1) **Se** $CI_i <= 75$ **então**
2) **Itemset$_i$** não é classificado como regra Excelente
3) **Se** ($CI_i > 75$) **e** ($CI_i < 85$) **então**
4) $(CI_i - 75)/10$
5) **Se** $CI_i >= 85$ **então**
6) Regra Excelente

Figura 11. Classificação de um **itemset** no conjunto nebuloso $\mu_{Excelente}(CI_i)$.

5.4 Similaridade entre Itemsets

Dependendo da área de conhecimento, diferentes critérios para descoberta de conhecimento podem ser utilizados. A seguir seguem algumas explicitações a respeito da similaridade.

5.4.1 Similaridade

No processamento de imagens, mais especificamente na segmentação, a seleção de critérios de similaridade depende não apenas do problema em consideração, mas também dos dados (imagem) disponíveis. Por exemplo, a análise de imageamento por satélite, para levantamento de terrenos, depende fortemente do uso de cor. Esse problema poderia ser muito mais difícil de ser tratado com a utilização apenas de imagens monocromáticas (Url 2).

As técnicas mais empregadas para classificação de padrões são baseadas no conceito de similaridade de padrões, onde um objeto é identificado como sendo ***x*** se suas características coincidirem, o mais próximo possível, das características de ***x*** (Costa, 2001).

A medida de similaridade é a formalização de uma determinada filosofia de julgamento de semelhança, através de um modelo matemático concreto. No raciocínio baseado em casos pode-se formalizar o conceito de similaridade de três formas diferentes (Url 3):

- Similaridade como predicado;
- Similaridade como relação de preferência;
- Similaridade como medida.

A primeira idéia concebe similaridade como entre dois objetos ou fatos, que existem ou não existem. A segunda, pressupõe a idéia de uma similaridade maior ou menor, enquanto o terceiro enfoque postula a quantificação da extensão dessa semelhança.

Trabalhos envolvendo o tema similaridade podem ser encontrados em (Azevedo, 2003) e (Angeles, 2003).

5.4.2 Função Similaridade **($\mu_{Similaridade}(D)$)**

O ACI foi testado com um intervalo de similaridade escolhido de forma arbitrária. Para que possam ser aplicados diferentes critérios de similaridade, o ACI possibilita a mudança da variável *VI*, que define o intervalo de similaridade entre **itemsets**.

A similaridade proposta pelo ACI procura por **itemsets** que possuam valores aproximados, não levando em consideração o valor semântico dos **itemsets** comparados.

Com o agrupamento de vários **itemsets** para uma mesma regra, através da lógica nebulosa, evita-se a perda de **itemsets** que possuam valores da *CI* muito aproximados ao valor da *CI* do **itemset** inicialmente proposto.

A função de pertinência que julga se outros **itemsets** são similares **($\mu_{Similaridade}(D)$)** ao **itemset** proposto inicialmente é somente para os **itemsets** que tenham

$$\left|CI_{\ i} - CI_{\ i+1}\right| < VI \qquad (1)$$

calculamos

$$D = \left|CI_i - CI_{i+1}\right| \qquad (2)$$

e

$$\mu_{Similarida\ de}(D) = \left|\frac{D}{VI} - 1\right|. \qquad (3)$$

Onde:

- CI_i é o valor da confiança inversa para o **itemset** proposto;
- CI_{i+1} é o valor da confiança inversa para o **itemset** seguinte;
- D é o módulo entre a diferença de CI_i e CI_{i+1};
- VI é o máximo valor que uma regra pode distanciar de outra regra, para ser classificada como não similar.

A definição de $D \geq 0$ acontece pelo fato de que pode ocorrer de $CI_{\ i+1} > CI_{\ i}$, ou seja, a diferença entre **itemsets** pode ocasionar um valor negativo.

Para redefinição da variável *VI* como um subconjunto nebuloso [0, 1] dentro do conjunto nebuloso de classificação da regra, é utilizado o conceito do conjunto de *Cantor*. Segundo Lima (1976) o conjunto de *Cantor K* é um subconjunto fechado do intervalo [0, 1], obtido como complementar de uma reunião de intervalos. Dessa forma, *VI* é redefinido como um subconjunto nebuloso [0, 1] dentro do conjunto nebuloso de classificação da regra.

A função pode ser ajustável, basta alterar a variável *VI*, desta forma o subconjunto nebuloso é redimensionado de forma que se pode aumentar ou diminuir o intervalo que torna um **itemset** similar ou não similar a outro **itemset**. Definido o subconjunto nebuloso, para o intervalo de similaridade *VI*, o algoritmo aplica (1) para certificar que as regras estejam dentro

do intervalo estipulado pelo usuário. Caso a diferença entre CI_i e CI_{i+1} seja maior ou igual ao valor de *VI*, então os **itemsets** são classificados como *não similares*, ou seja, o resultado da **($\mu_{Similaridade}(D)$)** é igual à zero (Figura 12, linhas 1 e 11).

No caso em que (1) é satisfeita, o algoritmo encontra o valor para *D* na equação (2), então calcula-se a função similaridade que é o quociente de *D* por *VI* menos 1 (Figura 12, linhas 2 e 3). Após encontrar o valor da **($\mu_{Similaridade}(D)$)** o ACI apresenta a similaridade entre os **itemsets** propostos (Figura 12, linhas 4 a 9).

Desta forma, pode-se encontrar **itemsets** similares ao **itemset** procurado ou demonstra a não similaridade entre **itemsets**.

O quadro 06 mostra a classificação, apresentada pelo algoritmo, para dois **itemsets,** de acordo com o resultado da **($\mu_{Similaridade}(D)$)**. Quanto mais próximo de um for o resultado da função, mais similares se tornam dois **itemsets**.

($\mu_{Similaridade}(D)$) = 1	Similar
0,5< = **($\mu_{Similaridade}(D)$)** < 1	Quase Similar
0 < **($\mu_{Similaridade}(D)$)** < 0,5	Pouco Similar
($\mu_{Similaridade}(D)$) = 0	Não Similar

Quadro 06. Classificação para o conjunto nebuloso similaridade.

Por fim o ACI apresenta os **itemsets** que possuem alguma similaridade, e a respectiva classificação, com o **itemset** proposto.

//Similaridade

1) **Se** $|CI_i - CI_{i+1}| < VI$ **então**

2) $D = |CI_i - CI_{i+1}|$

3) $\mu_{Similaridade}(D) = \left|\frac{D}{VI} - 1\right|$

4) **Se** $(\mu_{Similaridade}(D)) = 1$ **então**

5) **Itemsets** Similares

6) **Se** $(\mu_{Similaridade}(D)) < 1$ e $(\mu_{Similaridade}(D)) >= 0{,}5$ **então**

7) **Itemsets** Quase Similares

8) **Se** $(\mu_{Similaridade}(D)) < 5$ e $(\mu_{Similaridade}(D)) > 0$ **então**

9) **Itemsets** Pouco Similar

10) **senão**

11) **Itemsets** Não Similares

Figura 12. Similaridade entre **itemsets**.

6 APLICAÇÃO DO ALGORITMO

O algoritmo teve como alvo nos testes iniciais o HUUFMA, mais especificamente o centro cirúrgico. Esta seção apresenta informações a respeito do HUUFMA e os resultados obtidos na aplicação do ACI.

6.1 Ambiente Experimental

O HUUFMA é um hospital de ensino, pesquisa e extensão, que destina 100% dos seus leitos aos usuários do Sistema Único de Saúde (SUS) - sua fonte de financiamento. O Hospital Universitário é formado por um conjunto de duas Unidades: Unidade Materno Infantil e Unidade Presidente Dutra.

A Unidade Materno Infantil – UMI – oferece assistência integral à mulher e à criança, buscando garantir aos usuários um atendimento humanizado. Na Unidade Presidente Dutra – UPD – são oferecidos os Serviços Assistências em Clínica Médica, Clínica Cirúrgica, Transplantes, Hemodinâmica, UTI Geral e Cardíaca, Litotripsia, Terapia Renal Substitutiva e outros (URL 01). O quadro 07 apresenta alguns dados do HUUFMA.

Os serviços oferecidos pelo HUUFMA são:

- Serviço de Cardiologia;
- Central de notificação, captação e distribuição de órgãos;
- Comissões;
- Serviço de Litotripsia e Urodinâmica;

- Serviço de Nefrologia;
- Serviço de Neonatologia;
- Neurocirurgia;
- Serviço de Cirurgia da Obesidade;
- Serviço de Oftalmologia;
- Serviço de Otorrinolaringologia;
- Serviço de Cirurgia.

O algoritmo foi aplicado no banco de dados do centro cirúrgico do HUUFMA, UPD. As tabelas estavam estruturadas no Microsoft Access®. A base de dados do HUUFMA continha 29562 registros, com data a partir do dia 22/02/1998. A figura 13 apresenta o modelo lógico de dados da aplicação. A seção 7.2 apresenta alguns dos resultados obtidos com a aplicação do ACI no banco de dados do HUUFMA.

Nome do Estabelecimento	Hospital Universitário
Razão Social	Universidade Federal do Maranhão
Tipo de Unidade	Hospital Geral
Esfera Administrativa	Federal
Natureza Jurídica Institucional	Administração Direta – MEC
Número Total de Leitos	515
Número de Salas de Cirurgia	14
Número de Especialidades Médicas	26

Quadro 07. Dados do HUUFMA.

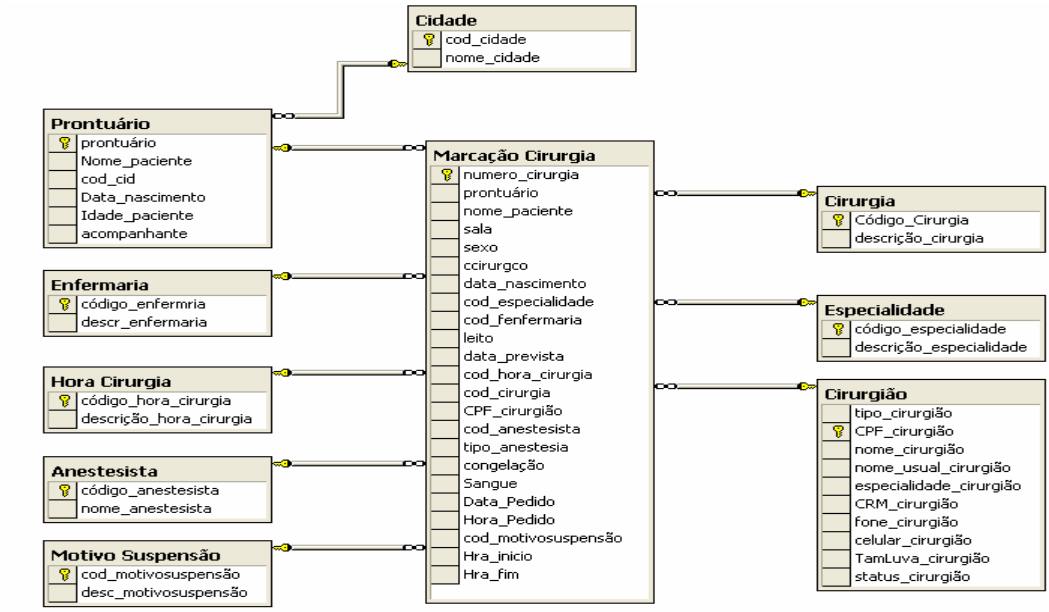

Figura 13. Modelo lógico de dados do centro cirúrgico do HUUFMA.

6.2 Resultados Preliminares

Os itens utilizados nos testes foram escolhidos arbitrariamente, ou seja, não foi utilizado nenhum critério de um especialista ou funcionário do HUUFMA.

O quadro 08 apresenta alguns dos campos usados nos testes e suas respectivas descrições. Como o objetivo era testar quais as regras que seriam extraídas através da medida da *CI*, foram utilizados valores de suporte e confiança baixos e não foram aplicadas às medidas *suporte* e *confiança* os conceitos da lógica nebulosa.

CAMPO	CONTEÚDO DOS CAMPOS
Anestesia	Local, Geral, Raque, Bloqueio e Peridural.
Cirurgia	Tireoidectomia
Clínica	Bucomaxila, Endoscopia, Plástica, Pediatria, Vascular, Obstetrícia, Oftalmologia, Neurologia, Ortopedia, Cirurgia Geral, Proctologia, Urologia, Ginecologia, Cabeça e Pescoço, Otorrinolaringologia e Cardíaca.
Motivo Suspensão	1 – Ausência de médico; 2 – Paciente não internado; 3 – Decisão médica; 4 – Falta de condições do paciente; 5 – Falta de condições do setor; 6 – Falta de leito na UTI; 7 – Falta de sangue; 8 – Outros; 9 – Falta de material.
Sexo	M – masculino e F – Feminino

Quadro 08. Campos utilizados no teste e as respectivas descrições.

A variável *VI* que define o intervalo de similaridade, foi definida de modo arbitrário com o valor 20%, ou seja, também não foi utilizado critério para escolha do valor intervalo.

6.2.1 Classificação dos Itemsets

A tabela 12 apresenta os **itemsets** pesquisados para os testes iniciais do ACI. Em seguida as seções 6.2.2 a 6.2.4 mostram os resultados obtidos para cada **itemset** para a classificação, assim como os respectivos comentários. A seção 7.2.4 mostra o resultado obtido para os **itemsets** para a similaridade entre regras.

Itemset	X	Y
Itemset 1	Sexo (Masculino)	Anestesia (Local)
Itemset 2	Clinica (Cirurgia Geral)	Motivo Suspensão (Dois)
Itemset 3	Sexo (Feminino)	Cirurgia (Tireoidectomia)

Quadro 09. **Itemsets** pesquisados nos testes iniciais.

6.2.2 Classificação para o Itemset 1

Sexo (Masculino) $\Rightarrow$ Anestesia (Local)

- *Suporte*: 4,99%
- *Confiança*: 8,80%
- *CI*: 63,48%
- Regra: A soma das ocorrências de *y* quando ocorre *x* é maior do que a soma das ocorrências de *y* para as outras possibilidades de *x*.
- Classificação: $\mu_{Boa}(CI) = 0{,}15$ $\mu_{Ótima}(CI) = 0{,}85$

Significado da regra para o **itemset 1**: O uso da anestesia local é mais freqüente no sexo masculino com uma considerável diferença. Com o valor da *CI* maior do que 50% mostra que a anestesia local é mais aplicada no sexo masculino do que no sexo feminino, que é o outro possível conteúdo de ***x***. Aplicando a função de pertinência da classificação temos o valor 0, 84, ou seja, o **itemset 1** está mais próximo de ser classificado como uma regra ótima do que como uma regra boa. O gráfico 1 apresenta os valores do *Suporte, Confiança e CI* e a figura 14 apresenta o valor da *CI* do **itemset 1** no conjunto nebuloso.

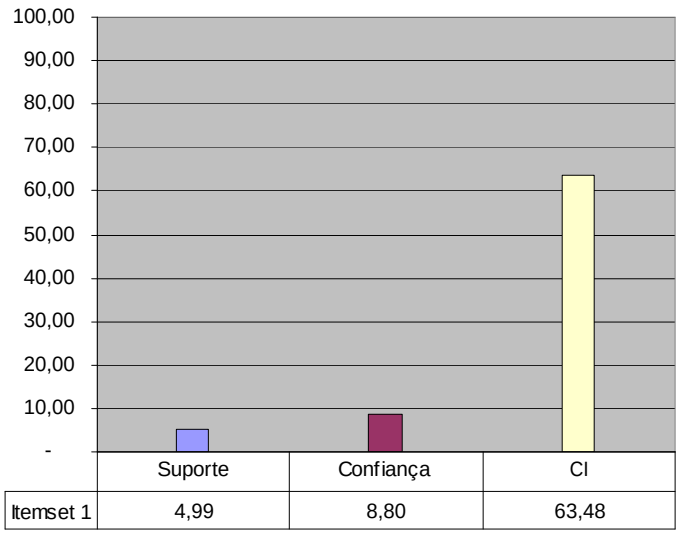

Gráfico 01. Valores de *Suporte*, *Confiança* e *CI* para o **itemset 1**.

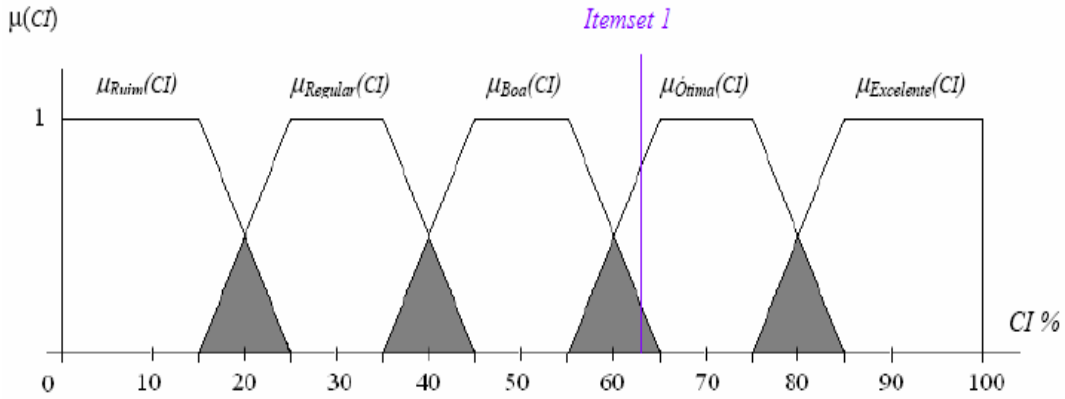

Figura 14. Valor da *CI* do **itemset 1** no conjunto nebuloso da classificação.

6.2.3 Classificação para o Itemset 2

Clínica (Cirurgia Geral) ⇒ Motivo Suspensão (Dois)

- *Suporte*: 2,56%
- *Confiança*: 7,77%
- *CI*: 40,15%
- Regra: A soma das ocorrências de *y* quando ocorre *x* é menor do que a soma das ocorrências de *y* para as outras possibilidades de *x*.
- Classificação: $\mu_{Regular}$ **(CI)** = 0,48 μ_{Boa} ***(CI)*** **=** 0,52

Significado da regra para o **itemset 2**: a soma das ocorrências dos outros motivos de suspensão de cirurgia ocorrem mais na cirurgia geral do que o motivo dois, que significa "paciente não internado". O motivo dois é o mais encontrado para ocorrência de Cirurgia Geral, porque possui o valor da *CI* maior do que quando se procura outra especialidade e motivo de suspensão número dois. Aplicando a função de pertinência para regra dois, obtemos o valor 0, 52, ou seja, o **itemset 2** pode ser classificado como regra boa, mas está no limiar de também ser classificada como regra regular. O gráfico 2 apresenta os valores do **itemset 2,** e a figura 15 apresenta o valor da *CI* do **itemset 2** no conjunto nebuloso.

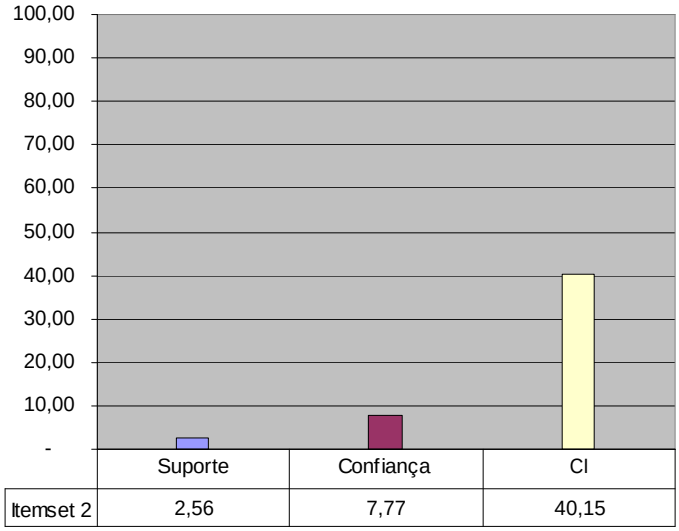

Gráfico 02. Valores de Suporte, Confiança e Confiança Inversa para o **itemset 2**.

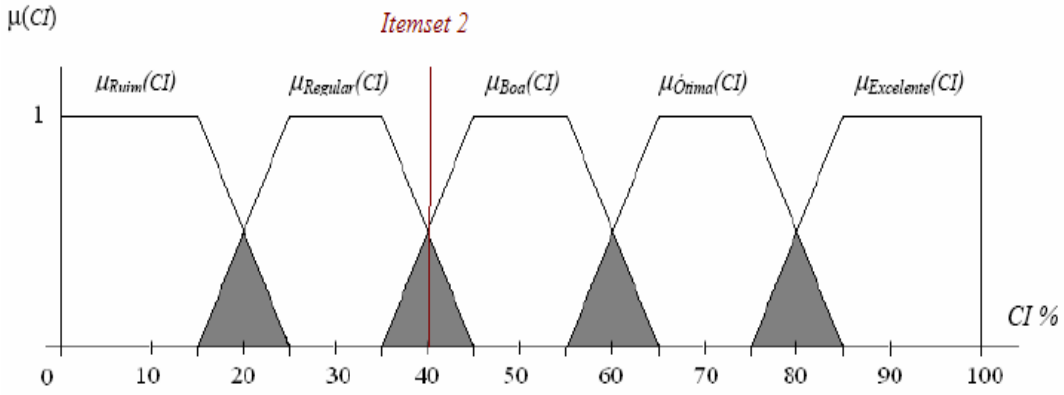

Figura 15. Valor da CI do **itemset 2** no conjunto nebuloso da classificação.

6.2.4 Classificação para o Itemset 3

Sexo (Feminino) $\Rightarrow$ Cirurgia (Tireoidectomia)

- *Suporte*: 0,75%
- *Confiança*: 1,73%
- *CI*: 87,74%
- Regra: A soma das ocorrências de *y* quando ocorre *x* é menor do que a soma das ocorrências de *y* para as outras possibilidades de *x*.
- Classificação: $\mu_{Excelente}\ (CI) = 1$

Significado da regra para o **itemset 3**: as cirurgias de Tireoidectomia são mais freqüentes no sexo feminino do que no sexo masculino, o que é comprovado na bibliografia da área médica segundo Valenti (1967). A *CI* para este **itemset** atingiu um valor bastante elevado e que poderá facilmente se transformar em regra com baixa margem de erro. De acordo com a função de pertinência, o **itemset 3** é classificado como regra excelente. O gráfico 3 apresenta os valores do **itemset 3,** e a figura 16 apresenta o valor da *CI* do **itemset 3** no conjunto nebuloso.

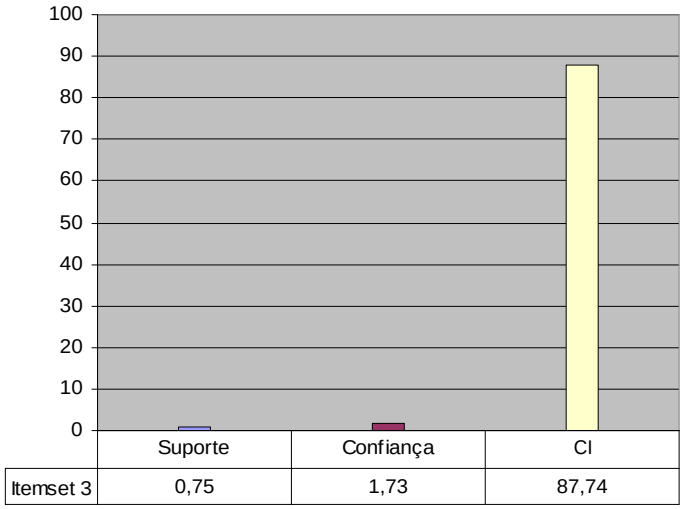

Gráfico 03. Valores de *Suporte*, *Confiança* e *CI* para o **itemset 3**.

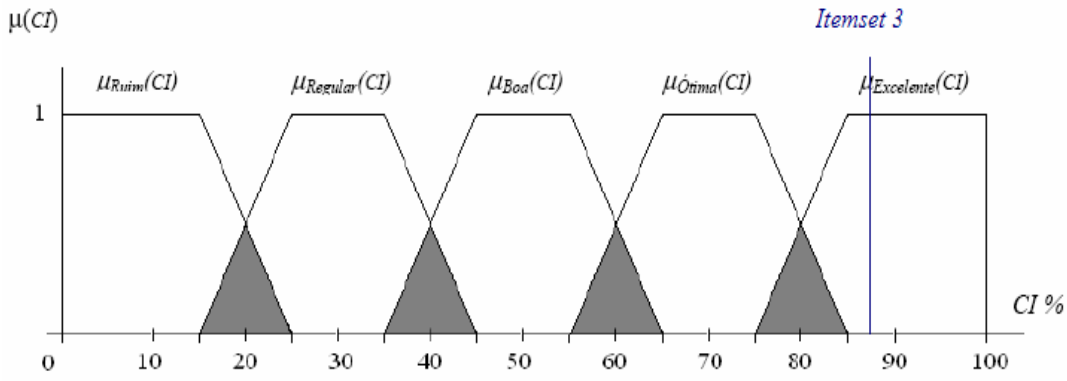

Figura 16. Valor da CI do **itemset** 3 no conjunto nebuloso da classificação.

A figura 17 apresenta os **itemsets 1**, **2** e **3** no conjunto nebuloso de classificação da regra.

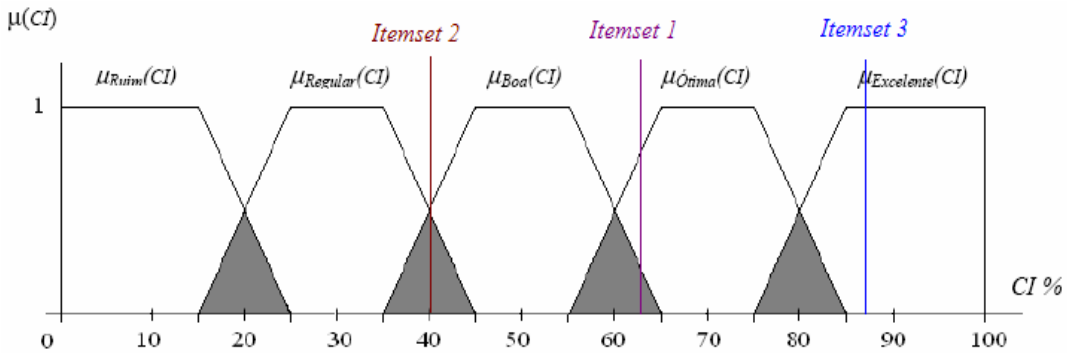

Figura 17. **Itemset**s **1**, **2** e **3** no conjunto nebuloso da classificação.

6.3 Similaridade entre Itemsets

A figura 18 exibe os **itemsets** (*1*, *2* e *3*) propostos pelo usuário e os respectivos intervalos de similaridades. As linhas em vermelho, transversais ao eixo da *CI*, exibem a distância do valor intervalo utilizado para testar a similaridade entre os **itemsets** propostos. As regras que possuem *CI* em algum dos intervalos terão alguma similaridade com o respectivo **itemset**. Observa-se que nenhum dos **itemsets** propostos chega a atingir uma das linhas do *VI* de outro **itemset**, para que fosse classificado pela **($\mu_{Similaridade}$ *(D)*)**.

Aplicando a função **($\mu_{Similaridade}$ *(D)*)** para os resultados obtidos para os **itemsets 1**, **2** e **3**, obtemos a seguinte informação: os **itemsets** foram classificados como "**não similares**" por que a diferença entre as regras é maior do que o intervalo de similaridade proposto.

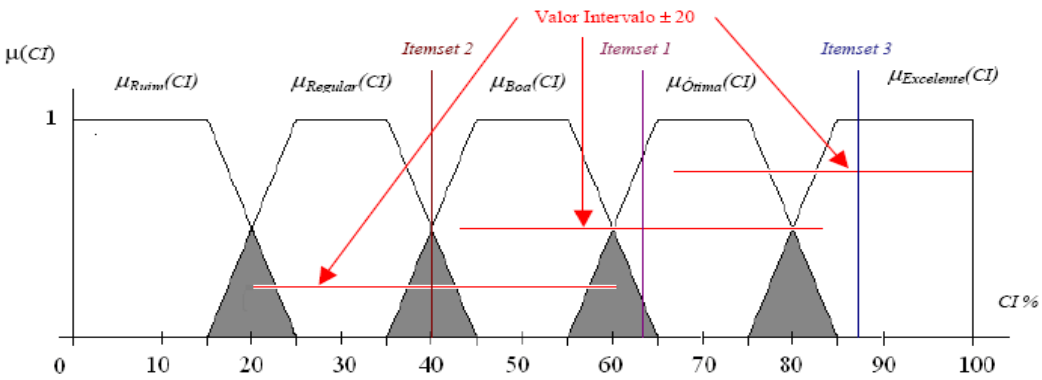

Figura 18. **Itemsets 1**, **2** e **3** no conjunto nebuloso da classificação e os respectivos intervalos de similaridade.

Similares aos **itemsets** propostos foram encontrados dois **itemsets**. O **itemset 4** Sexo (Masculino) ⇒ Turno (Matutino), cirurgias realizadas no turno matutino e o **itemset 5** Clínica (Cirurgia Geral) ⇒ Origem Paciente (Um). Origem Paciente um significa: paciente interno.

Os valores encontrados para o **itemset 4** são:

- *Suporte*: 32,94%;
- *Confiança*: 58,09%;
- *CI*: 57,83%.

O gráfico 4 apresenta os valores de *Suporte*, *Confiança* e *CI* par o **itemset 4**.

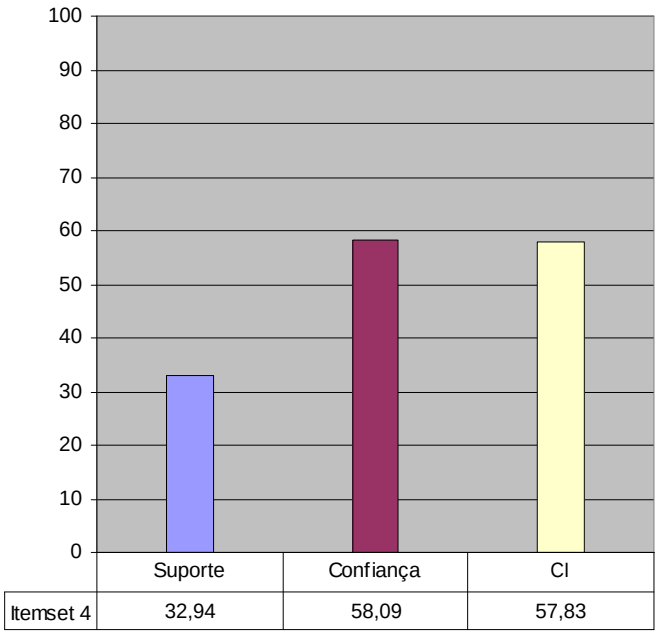

Gráfico 04. Valores de *Suporte*, *Confiança* e *CI* para o **itemset 4**.

O resultado da aplicação da **($\mu_{Similaridade}$ *(D)*)** para os **itemsets 1** e **4** é de 0,72. Dessa forma, de acordo com a tabela 1 da seção 6.2, são classificados como “quase similares”.

O resultado da aplicação da **($\mu_{Similaridade}$ *(D)*)** para os **itemsets 2** e **4** é de 0,12. Dessa formam, de acordo com a tabela 1 da seção 6.2, são classificados como “pouco similares”.

A figura 19 apresenta o **itemset 4** comparado aos **itemsets 1** e **2**. A letra “D” representa a diferença entre os valores da *CI* dos **itemsets** usada no cálculo do resultado da **($\mu_{Similaridade}$ *(D)*)**.

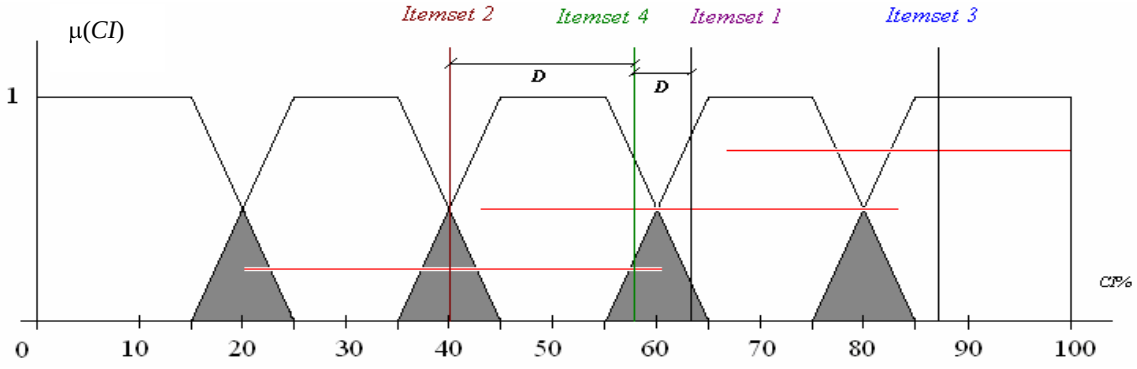

Figura 19. Comparação do **itemset 4** com os **itemsets 1, 2** e **3**.

Os valores encontrados para o **itemset 5** são:

- *Suporte*: 23,99%;
- *Confiança*: 72,80%;
- *CI*: 33,74%.

O gráfico 5 apresenta os valores de *Suporte*, *Confiança* e *CI* par o **itemset 5**.

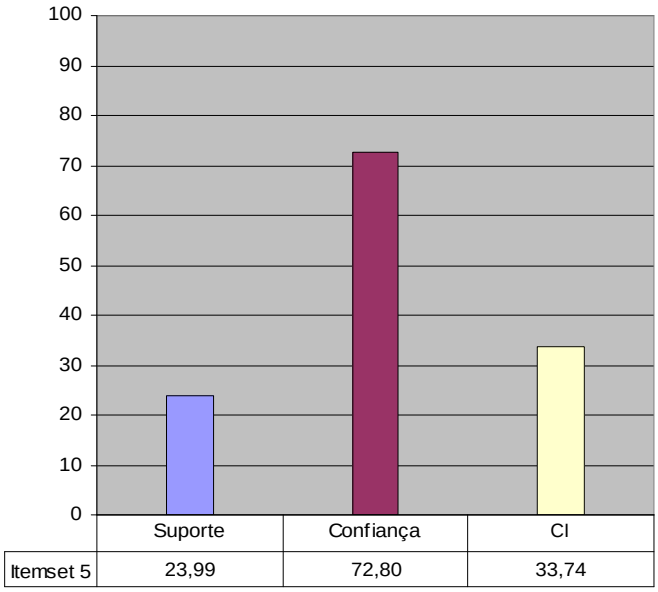

Gráfico 05. Valores de *Suporte*, *Confiança* e *CI* para o **itemset 5**.

O resultado da aplicação da **($\mu_{Similaridade}$ *(D)*)** para os **itemsets 2** e **5** é de 0,68. Dessa forma, de acordo com a tabela 1 da seção 6.2, são classificados como “quase similar”.

A figura 20 apresenta o **itemset 5** comparado ao **itemsets 5**. Assim como na figura anterior, a letra “*D*” representa a diferença entre os valores da *CI* dos **itemsets** usada no cálculo do resultado da **($\mu_{Similaridade}$ *(D)*)**.

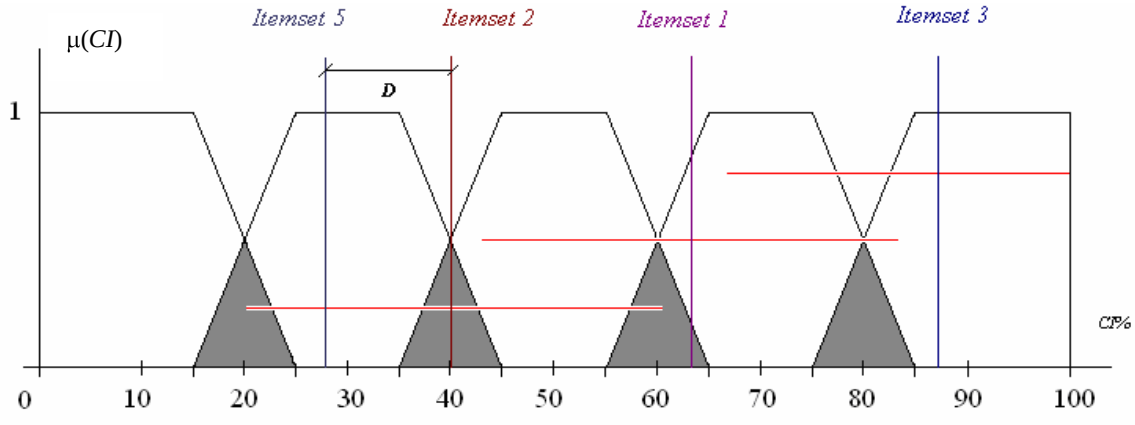

Figura 20. Comparação do **itemset** 5 com os **itemset**s 1, 2 e 3.

O **itemset 4,** de acordo com a $\mu_{Boa}(CI)$, tem valor de 0,72. Está mais próximo de ser classificado como regra boa, enquanto que o **itemset 1** está mais próximo de ser classificado como regra ótima. Já o **itemset 5,** de acordo com a $\mu_{Regular}(CI)$, tem resultado 1 e é classificado como regra regular.

7 TRABALHOS RELACIONADOS

Em face aos algoritmos de mineração existentes, que utilizam técnicas de regras de associação, é feito um comparativo entre o algoritmo proposto e um algoritmo clássico quando se trata de regras de associação, o Apriori. Escovar (2004) e Kantardzic (2003).

7.1 Declaração Formal do Problema

Seja um banco de dados contendo informações sobre os pesquisadores atuantes no Brasil. Almejasse descobrir associações importantes entre esses dados, onde:

- I = {i1, i2,..., im} é um conjunto de literais, denominados itens. São as características e atributos dos pesquisadores. Por exemplo, I = {idade, sexo,..., área de atuação, artigos publicados};
- T é um conjunto de certos itens de um pesquisador, tal que $T \subseteq I$;
- D é uma tabela representando todas as características e atributos de todos os pesquisadores; e
- X, Y são conjuntos de itens específicos dos pesquisadores, tal que $X \subseteq T$ e $Y \subseteq T$.

Uma regra de associação é uma implicação da forma $X \Rightarrow Y$, onde $X \subset I$, $Y \subset I$ e $X \cap Y = \varnothing$. A regra $X \Rightarrow Y$ pertence a D com confiança c se c% dos registros em D que contém X também contém Y. A regra $X \Rightarrow Y$ tem suporte s em D se s% dos registros em D contém $X \cup Y$. Então, dado uma tabela D, o objetivo é descobrir as regras de associação interessantes.

O problema de descobrir todas as regras de associação pode ser decomposto em duas etapas:

1. Encontrar todos os conjuntos de itens (**itemset**s) que apresentam suporte maior que o suporte mínimo estabelecido pelo decisor. Os **itemsets** que atendem a este quesito são denominados **itemsets** freqüentes; e
2. Utilizar os **itemsets** freqüentes obtidos para gerar as regras de associação do banco de dados.

O desempenho geral da mineração de regras de associação é determinado pela primeira etapa, a qual exige sucessivas buscas na base de dados. Em se encontrando o conjunto dos **itemsets** freqüentes, as regras de associação correspondentes podem ser diretamente identificadas.

Algoritmos que realizem a contagem eficiente dos grandes **itemsets** são a chave para o sucesso dos métodos de mineração em grandes bancos de dados (Romão, 2001).

7.2 Funcionamento do APRIORI

O algoritmo Apriori é um dos algoritmos mais conhecidos quando o assunto é mineração de regras de associação em grandes bancos de dados centralizados. Ele encontra todos os conjuntos de itens freqüentes, denominados **itemsets** freqüentes (L_k).

O algoritmo principal (Apriori) faz uso de duas funções: a função Apriori_gen, para gerar os candidatos e eliminar aqueles que não são freqüentes, e a função Genrules, utilizada para extrair as regras de associação.

O primeiro passo do algoritmo Apriori é realizar a contagem de ocorrências dos itens para determinar os **itemsets** freqüentes de tamanho unitário (1-**itemset**s freqüentes). Os passos posteriores, k, consistem em duas fases. Primeiro, os **itemsets** freqüentes L_{k-1}, encontrados no passo anterior (k-1), são utilizados para gerar os conjuntos de itens potencialmente freqüentes, os **itemsets** candidatos (C_k). O procedimento para geração de candidatos é descrito no parágrafo seguinte. Na seqüência, é realizada uma nova busca no banco de dados, contando-se o suporte de cada candidato em C_k.

A geração dos **itemsets** candidatos, de antemão, toma como argumento L_{k-1}, o conjunto de todos (k - 1) **itemsets** freqüentes. Para tal, utiliza-se a função Apriori_gen, que retorna um superconjunto de todos os k-**itemset**s freqüentes. A intuição por trás desse procedimento é que, se um **itemset** X tem suporte mínimo, todos os seus subconjuntos também terão. A função, em um primeiro estágio, une L_{k-1} com L_{k-1}. No estágio seguinte, são eliminados os **itemset**s $c_k \in C_k$, desde que um dado (k-1)-subset de c_k não pertença a L_{k-1}.

O último passo é a descoberta das regras de associação, obtida através da função Genrules. A geração de regras, para qualquer **itemset** freqüente, significa encontrar todos os subsets não vazios de l. Assim, para todo e qualquer subset a, produz-se uma regra $a \Rightarrow (l - a)$ somente se a razão (suporte (l)/suporte(a)) é ao menos igual à confiança mínima estabelecida pelo usuário (Romão, 2001).

O quadro 10 a seguir faz comparações em alguns pontos chaves em relação ao ACI.

ALGORITMOS / CARACTERÍSTICAS	APRIORI	ACI
Utilizam **itemset** definido pelo usuário na busca de conhecimento?	Não	Sim
Apresentam tendência de Y em relação a X sem precisar inverter o **itemset** para uma nova busca?	Não	Sim
Possui uma medida que tem um número de regras fixas?	Não	Sim
Utiliza lógica nebulosa para colocar o conhecimento extraído ao nível de linguagem natural?	Não	Sim

Quadro 10. Comparativo entre o ACI e o Apriori.

Observações:

a. Gerar regras de interesse do usuário, baseadas em uma nova medida que não são tratadas pelos outros algoritmos e que possam ser importantes e auxiliar na extração de conhecimento;
b. Ter uma quantidade de regras fixas tornando a análise dos resultados mais concisa e eficiente, ao invés de um conjunto indefinido de regras;
c. Auxiliado pela lógica nebulosa, agrupar regras, com base na similaridade, evitando a perda de informações nos caso em que regras estejam em faixa de valores próximos, mas com classificações diferentes.

7.3 Medida *Lift*

Lift (Passari, 2004) e (Url 2) mede quanto uma regra melhora a previsão de um resultado do que simplesmente assumindo o resultado. A melhoria é definida, matematicamente, como a freqüência observada para uma regra dividida pela freqüência esperada, dadas as freqüências de cada um dos itens.

Em comparação com a *CI*, as seguintes observações podem ser citadas:

a. *Lift* mede quanto uma regra melhora uma previsão e a *CI* classifica um **itemset**;
b. *Lift* trabalha com o **itemset** X $\rightarrow$ Y enquanto que a *CI* trabalha com **itemset** Y $\rightarrow$ X. Isso significa que as duas medidas são importantes, pois extraem diferentes tipos de informações.

c. *Lift* depende da confiança para apresentar uma informação a cerca de uma regra, ou seja, depende de uma outra medida. A *CI* não depende de nenhuma outra medida, assim como as medidas clássicas *Suporte* e *Confiança*.

8 CONCLUSÕES E TRABALHOS FUTUROS

A finalidade dessa dissertação foi a de apresentar o Algoritmo da Confiança Inversa. Foi apresentada a aplicação do algoritmo em um banco de dados, com os registros dos pacientes do centro cirúrgico do HUUFMA. Foram enfatizadas, a técnica de mineração de dados, regras de associação e a lógica nebulosa, para criação e atribuição de termos lingüísticos aos valores encontrados pelo algoritmo.

A extração de informações feita pelo ACI foi bastante proveitosa. O algoritmo foi aplicado no HUUFMA apenas nos testes iniciais, podendo ser utilizado em bases de dados em outros setores do próprio hospital assim como também em instituições que possuam outro contexto de trabalho com uma base de dados histórica.

A seguir, é apresentado, de forma resumida, o resultado obtido com a aplicação do algoritmo.

Dificuldades

Alguns registros do banco de dados do centro cirúrgico do HUUFMA estavam incompletos, o que limitou a escolha quanto aos campos utilizados nos testes iniciais. Alguns campos necessitaram de padronização quanto a seus valores, por apresentarem conteúdos de mesmo significado com caracteres diferentes.

A Medida da Confiança Inversa

Assim como o *Suporte* e a *Confiança*, quanto maior o valor da *CI* maiores são as chances do conhecimento extraído vir a se tornar regra. A medida da *CI* complementa as informações extraídas para os **itemset** pesquisados.

Algoritmo da Confiança Inversa

O ACI tem as seguintes características:

1. Não gera um conjunto muito grande de regras, ao invés disso, o ACI procura agregar o máximo de **itemsets** para a mesma regra, pois possui cinco regras fixas que podem ser empregadas para vários **itemsets** pesquisados;
2. O algoritmo realiza pesquisa somente com **itemsets** de interesse do usuário;
3. Pode gerar regras tanto de $X \rightarrow Y$ quanto de $Y \rightarrow X$. Neste último mesmo quando os valores de *suporte* e *confiança* sejam baixos, o que é o maior diferencial do ACI em relação aos outros algoritmos;
4. Aplica - se a lógica nebulosa para classificação e busca de **itemsets** similares. Dessa forma aproveitam-se regras com proximidade de valores classificados pelo ACI.

Classificação das Regras

A classificação dos **itemsets** em um conjunto de cinco regras torna mais prática a interpretação dos valores obtidos. O agrupamento dos **itemsets** em regra ruim, regular, boa, ótima ou excelente, tornou mais interessante a análise dos resultados obtidos, pois além do resultado numérico tem-se também a associação deste valor a uma variável lingüística para uma melhor interpretação dos resultados.

Similaridade entre Regras

Aplicando a **($\mu_{Similaridade}$ *(CI)*)** aos **itemsets** pesquisados tem-se que os valores que demonstram o quão próximo um **itemset** está em relação a outro **itemset** baseado em um intervalo (*VI*) estipulado pelo usuário. Esse intervalo que pode ser redefinido de acordo com critérios específicos de especialistas na área a qual o ACI está sendo aplicado para descoberta de conhecimento.

A partir de agora pode-se aproveitar regras de acordo com um grau de similaridade. Desta forma além de classificar os **itemsets** de acordo com o valor da *CI*, agora se aproveita-se ao máximo, qualquer proximidade entre os mesmos.

Análise de um Itemset Específico

O resultado obtido pelo ACI para o **itemset** Sexo (Feminino) $\Rightarrow$ Cirurgia (Tireoidectomia) demonstrou um valor alto para a *CI*, valor esse que corresponde a uma realidade comprovada na bibliografia da área médica (tópico 6.2.4 Classificação para o Itemset 3). Se a análise do **itemset** fosse baseada somente nos valores obtidos com *Suporte* e *Confiança*, não seria possível apresentar o alto grau de incidência de cirurgias de Tireoidectomia em pacientes do sexo feminino.

Propostas de trabalhos futuros

- Testar e validar o algoritmo, utilizando outra base de dados que apresentem um contexto diferente da base de dados utilizadas nos testes preliminares;
- Comparar os resultados do ACI com resultados de outros algoritmos de mineração;
- Aprimorar o ACI com outras tecnologias para maior interoperabilidade entre diversas bases de dados;
- Definir um critério para a atribuição de termos lingüísticos para os conjuntos nebulosos apresentados nessa dissertação;
- Aplicar a função similaridade em questão ao contexto dos **itemsets** pesquisados;
- Implementar o ACI de forma que a procura de **itemset**s similares não necessite da intervenção do usuário, ou seja, o algoritmo encontre e, posteriormente, apresente todos os **itemsets** que estiverem dentro do intervalo de similaridade (*VI*) estipulado pelo usuário.

REFERÊNCIAS

ANGELES, Pablo. **Estudo de Tochas de Plasma Através da Teoria da Similaridade.** (Tese de Mestrado. Instituto de Física). Universidade Estadual de Campinas. 2003.

AZEVEDO, Carlos., PLASTINO, Alexandre., VASCONCELOS, Ana. **PROCSIMO: uma Ferramenta de Procura de Similaridade entre Operons.** Universidade Federal Fluminense. Niterói. 2003.

BARAGOIN, Corinne., CHAN, Ronnie., GOTTSCHALK, Helena., MEYER, Gregor., PEREIRA, Paulo., VERHEES, Jaap. **Enhance Your Business Aplications: Simple Integration of Advance Data Mining Functions**. First Edition. December. 2002.

BARIONI, Maria., **Visualização de Operações de Junção de Sistemas de Base de Dados para Mineração de Dados. (**Dissertação de Mestrado – Instituto de Ciências Matemáticas e Computação). USP. São Carlos. 2002

BOSE, Ranjit., SUGUMARAN, Vijayan., **Application of Intelligent Agent Technology for Managerial Data Analysis and Mining**. The DATA BASE for Advances in Information Systems – Winter. 1999.

CAMARGOS, Fernando., **Lógica Nebulosa: uma abordagem filosófica e aplicada**. Departamento de Informática e de Estatística. UFSC. 2003.

CANÔAS, Ana. C., **Aplicação de Lógica Nebulosa na Análise das Redes de Energia Elétrica**. (Dissertação de Mestrado – Departamento de Sistemas de Energia Elétrica). UNICAMP. Campinas. 2003.

CHEM, M., HAN, J., YU, P., **Data Mining: An Overview from Database Perspective**. 1997.

COSTA, Silvia, M., F., **Classificação e Verificação de Impressões Digitais**. (Dissertação de Mestrado – Departamento de Sistemas Eletrônicos). Escola Politécnica da Universidade de São Paulo. 2001.

ESCOVAR, G., Eduardo. **Algoritmo SSDM para Mineração de Dados Semanticamente Similares**. (Dissertação de Mestrado – Departamento de Computação) UFSCar, São Carlos. 2004.

HAN, Jiawei., KAMBER, Micheline., **Data Mining: Concepts and Techniques.** Manuscript based on a forthcoming book by Jiawei Han and Micheline Kamberer, Morgan Kaufmann Publishers. 2000

KANTARDZIC, Mehmed., **Data Mining: Concepts, Models, Methods and Algorithms**. Ed. Wiley. 2003.

LIMA, E., Lages., **Curso de Análise**. Vol 1. Instituto de Matemática Pura e Aplicada. 1976.

MASSAD, E., MARIN, H. de F., AZEVEDO, R. Soares de. **O Prontuário Eletrônico do Paciente na Assistência, Informação e Conhecimento Médico.** Núcleo de Informática em Enfermagem. Universidade Federal de São Paulo. São Paulo. 2003.

MITRA, Sushmita., ACHARYA, Tinku., **Data Mining Multimedia, Soft Computing, and Bioinformatics**. Ed. Wiley. 2003.

NAVEGA, S., **Princípios Essenciais do Data Mining**. Publicado nos Anais do Infoimagem, Cenadem, Novembro. Agosto. 2002

PASSARI, A., **Exploração de Dados Atomizados para Previsão de Vendas no Varejo Utilizando Redes Neurais**. (Dissertação – Mestrado Departamento de Administração) USP. 2004.

PYLE, Dorian., **Business Modeling and Data Mining.** Morgan Kaufmann Publishers. San Francisco. CA. USA. 2003.

REZENDE, S. Oliveira., **Sistemas Inteligentes: Fundamentos e Aplicação.** 1ª Edição. Editora Manole. Barueri. SP. 2003.

ROMÃO, Wesley., NIEDERAUER, Carlos., MARTINS., Alejandro., TCHOLAKIAN, Aran., PACHECO, Roberto., BARCIA, Ricardo. **Extração D Regras D Associação em C&T: O Algoritmo Apriori.** Centro Tecnológico. UFSC – Universidade Federal de Santa Catarina. 2001.

RUD, P. Olivia. **Data Mining Cookbook**. Ed. Wiley. 2001.

VALENTI, P., **Medicina Interna: Compendio Práctico de Patologia Médica y Terapéutica Clínica**. Séptima Edición. Editorial Marin. Barcelona. 1967.

WANG, J. **Data Mining – Opportunities and Challenges**. Idea Group Publishing. 2003.

URLs

(Url 1) Hospital Universitário – HUUFMA. Acessado em 15 de janeiro de 2005. http://www.huufma.br/site/web/apresentacao.html

(Url 2) Segmentação. Técnicas de Segmentação de Imagens para Recuperação de Informações Visuais. Acessado em 20 de janeiro de 2005. http://atlas.ucpel.tche.br/~vbastos/segmenta.htm

(Url 3) Reconhecimento de Padrões – Raciocínio Baseado em Casos. Acessado em 30 de janeiro de 2005. http://www.inf.ufsc.br/~awangenh/RP/cbr.html