UNIVERSIDADE DO EXTREMO SUL CATARINENSE - UNESC

CURSO DE CIÊNCIA DA COMPUTAÇÃO

MAIARA COSTA MOTTA

AVALIAÇÃO DE DESEMPENHO DOS ALGORITMOS PARA CLASSIFICAÇÃO EM DATA MINING, ID3 E C4.5, NA SHELL ORION

CRICIÚMA, JULHO DE 2010

MAIARA COSTA MOTTA

AVALIAÇÃO DE DESEMPENHO DOS ALGORITMOS PARA CLASSIFICAÇÃO EM DATA MINING, ID3 E C4.5, NA SHELL ORION

Trabalho de Conclusão de Curso apresentado para obtenção do Grau de Bacharel em Ciência da Computação da Universidade do Extremo Sul Catarinense.

Orientadora: Prof.ª MSc. Merisandra Côrtes de Mattos

CRICIÚMA, JULHO DE 2010

MAIARA COSTA MOTTA

Avaliação de Desempenho dos Algoritmos para Classificação em Data Mining, ID3 e C4.5, na Shell Orion

Submetido ao corpo docente do Curso de Ciência da Computação da Universidade do Extremo Sul Catarinense como um dos requisitos para obtenção do grau de Bacharel em Ciência da Computação.

Profa. MSc. Ana Claudia Garcia Barbosa
Coordenadora do Curso de Ciência da Computação

Banca Examinadora:

Prof. MSc. Merisandra Côrtes de Mattos (UNESC)
Orientador

Prof. Msc. Cristian Cechinel (Universidade Federal do Pampa, Bagé/RS)

Prof. MSc. Kristian Madeira (UNESC)

RESUMO

A crescente utilização de bases de dados para armazenamento de informações por grandes empresas e organizações, impulsionou o interesse na descoberta de conhecimento e padrões contidos nesses dados. O processo de extração de conhecimento de bases de dados tendo como principal etapa o *data mining*, auxilia na análise desses dados e na tomada de decisão, utilizando para isso ferramentas computacionais, que empregam métodos e tarefas de *data mining*. Nesta pesquisa foi desenvolvida a avaliação de desempenho da tarefa de classificação pelos algoritmos ID3 e C4.5, na *Shell* Orion Data Mining Engine. No desenvolvimento da avaliação de desempenho foram empregados métodos estatísticos como: validação cruzada; cálculo do fator de confiança; matriz confusão e o cálculo do erro médio. Também realizou-se a validação do conhecimento, apresentando-se os resultados obtidos em formas de regras ao especialista do domínio de aplicação, que pode também validar a *Shell* Orion quanto a funcionalidade e usabilidade, esta avaliação foi efetuada em forma de entrevista com questionário. A classificação foi realizada a partir de uma base de dados contendo registros de pacientes com sepse da Unidade de Tratamento Intensivo do Hospital das Clinicas de Porto Alegre, Rio Grande do Sul. A avaliação dos algoritmos da *Shell* Orion obteve resultados bastante satisfatórios, principalmente pela isenção de taxas de erro de classificação, o que pode ser verificado também por meio da entrevista com o especialista que avaliou a ferramenta, confirmando a utilidade e relevância dos algoritmos implementados.

Palavras-Chave: *Data Mining*; *Shell* Orion *Data Mining Engine*, Classificação, Avaliação de Desempenho, Validação do Conhecimento.

ABSTRACT

The increasing use of databases for information storage by large companies and organizations has motivated an interest in knowledge discovery and data standards contained in these. The process of extracting knowledge from databases has in the main stage, the data mining process. It assists in data analysis and decision making, using computational tools that employ methods and data mining tasks. This research was conducted to evaluate performance during the classification's tasks by ID3 and C4.5 algorithms, this implementations are inside the software called Shell Orion Data Mining Engine. During the developing the performance evaluation statistical methods were employed such as: cross validation; calculation of confidence factor; confusion matrix and calculation of average error. Also was conducted the validation of the knowledge, presenting the results in forms of rules to the domain's specialist of application. The specialist could also validate the functionality and usability of Shell Orion tool. This assessment was made in an interview with questionnaire. The classification was made from a database containing records of patients with sepsis in the Unidade de Tratamento Intensivo do Hospital das Clínicas of Porto Alegre, Rio Grande do Sul. The Shell Orion algorithms' evaluation obtained satisfactory results, mainly for tax exemption of misclassification, which can also be verified through interviews with the specialist who evaluated the tool, confirming the usefulness and relevance of the implemented algorithms.

Keywords: Data Mining; Shell Orion Data Mining Engine, Classification, Performance Evaluation, Knowledge Validation.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CART	*Classification and Regression Trees*
DCBD	Descoberta de Conhecimento em Base de Dados
DM	*Data Mining*
IBGE	Instituto Brasileiro de Geografia e Estatística
KDD	*Knowledge Discovery in Databases*
SGBD	Sistema Gerenciador de Banco de Dados
TCC	Trabalho de Conclusão de Curso
UNESC	Universidade do Extremo Sul Catarinense
UTI	Unidade de Terapia Intensiva

SUMÁRIO

1 INTRODUÇÃO

O desenvolvimento da tecnologia da informação possibilitou as empresas e instituições uma maior capacidade de armazenamento dos seus dados, mediante isso, teve-se a necessidade de especialistas realizarem a análise desses dados a fim de proporcionar um ganho de informação para estas instituições, bem como auxiliar no processo de tomada de decisão. No entanto, essas bases ao longo dos anos tornaram-se muito volumosas o que dificultou e transformou a análise por especialistas humanos muito trabalhosa (REZENDE, 2005).

Assim, novas metodologias e ferramentas computacionais fizeram-se primordiais para a análise destes dados. Surgiu então, o conceito de Descoberta de Conhecimento em Bases de Dados, que visa a identificação de padrões e relações. Esse processo é composto pelas etapas de pré-processamento, *data mining* e pós-processamento. A etapa de *data mining* é considerada a principal, pois ela é efetivamente a responsável pela identificação dessas relações, envolvendo métodos e algoritmos que evidenciam padrões e vinculações entre as variáveis registradas na base de dados. Esta etapa do DCBD integra conceitos de banco de dados, estatística e aprendizado de máquina (GOLDSHMDT; PASSOS, 2005).

Na aplicação de *data mining* empregam-se ferramentas computacionais, denominadas de *Shell* que auxiliam na execução de tal processo. Essas ferramentas em sua maioria são comerciais, considerando isso, existem pesquisas acadêmicas de diversas universidades que têm como objetivo o desenvolvimento de uma aplicação gratuita. O Grupo de Pesquisa em Inteligência Computacional Aplicada, do curso de Ciência da Computação da Universidade do Extremo Sul Catarinense (UNESC), tem em desenvolvimento desde o ano de 2005 o projeto da *Shell* Orion *Data Mining Engine*. A *Shell* Orion aplica diferentes tarefas de *data mining* e possui já implementados métodos das tarefas de associação, classificação e

clusterização. Cada algoritmo tem sido desenvolvido por um acadêmico da UNESC em seu Trabalho de Conclusão de Curso.

Esta ferramenta tem sido utilizada para análise de diferentes tipos de bases de dados, como por exemplo, as oriundas do Curso de Medicina da UNESC: prevalência da asma e rinite em adolescentes escolares de Criciúma; registros referentes a sepse[1] em pacientes da Unidade de Terapia Intensiva do Hospital de Clínicas de Porto Alegre, entre outras. Além da aplicação da *Shell* Orion em diferentes projetos de pesquisa que envolvem os cursos de Ciência da Computação, Engenharia Ambiental e Ciências Ambientais, para análise de dados referentes ao monitoramento de bacias hidrográficas da região carbonífera (Bacia do Rio Urussanga e Araranguá).

A aplicação destas diferentes bases de dados têm gerado resultados satisfatórios, porém existe uma tendência natural de crescimento da quantidade de conhecimento armazenado, bem como de se validar as relações e os padrões que vêm sendo descobertos quando aplicam-se bases de dados específicas na *Shell* Orion. Salienta-se que esta validação é importante para verificar se o conhecimento descoberto está correto e atende as expectativas do especialista do domínio de aplicação para que se possa confiar nas relações obtidas (SOMMERVILLE, 2003).

A validação de bases de conhecimento depende da técnica utilizada na implementação, o que dificulta o estudo dessas bases. Na maioria dos casos a validação pode ser realizada quanto ao seu comportamento por meio de uma base de testes, que são problemas/casos submetidos ao sistema e que devem ser analisados por meio de métodos e modelos estatísticos, que foram definidos no decorrer da pesquisa mediante os estudos realizados (NASSAR, 2007).

[1] Doença desenvolvida a partir de uma infecção sistêmica grave, conhecida como infecção generalizada (CECIL; GOLDMAN; AUSIELLO, 2005).

Esta pesquisa apresenta uma metodologia para avaliar o desempenho e validar o conhecimento descoberto pela aplicação do método de classificação e dos algoritmos de *data mining,* ID3 e C4.5 implementado na *Shell* Orion, aplicando-se para análise uma base de dados.

1.1 OBJETIVO GERAL

Demonstrar as formas de avaliação de desempenho em *Data Mining*.

1.2 OBJETIVOS ESPECIFICOS

Os objetivos do presente trabalho são:

a) compreender o *data mining* e o processo de avaliação de desempenho;
b) identificar métodos e modelos estatísticos para avaliação de desempenho em aplicações de *data mining*;
c) selecionar os métodos e modelos a serem utilizados na realização da pesquisa;
d) apresentar os métodos e modelos estatísticos selecionados;
e) utilizar uma base de dados para validar as relações e padrões descobertos pelo *data mining* nesta base;
f) **e**mpregar a *Shell* Orion *Data Mining Engine* na avaliação realizada;
g) oferecer uma alternativa para avaliação de desempenho em *data mining*.

1.3 JUSTIFICATIVA

A informação se difere do conhecimento, pois informações são constituídas de dados organizados, agrupados e categorizados em modelos para ter significado, enquanto o conhecimento é capaz de habilitar ações corretas, constituindo-se no uso produtivo da informação. O conhecimento é algo intangível, aplicado aos fatos e idéias adquiridos por estudos, experiências, observações ou investigações. Assim, o conhecimento tem se tornado fundamental, pois apresenta um grande potencial competitivo para as organizações (CARVALHO, 2005).

Estas sempre enfrentaram o problema de armazenamento dos dados, que anteriormente era feito em papel. O surgimento e o desenvolvimento de tecnologias que possibilitaram a digitalização da informação, ocasionaram a necessidade do desenvolvimento de novas ferramentas para gerenciamento e armazenamento. No decorrer dos anos essas bases de dados tornaram-se muito robustas, o que dificultou a análise e extração de conhecimento manualmente. A partir daí surgiram técnicas de descoberta de conhecimento baseando-se em ferramentas computacionais que utilizam abordagens de inteligência computacional e estatística.

Na análise desses dados são empregados métodos estatísticos com o objetivo de identificar padrões nessas bases. Contudo, esse processo possui limitações como a necessidade de pessoas especializadas, os diferentes tipos de informações contidas, bem como, o grande volume de dados. Considerando esses aspectos, realiza-se o processo de Descoberta do Conhecimento em Bases de Dados que transforma informações em conhecimento utilizando técnicas/ferramentas que possibilitam apresentar e analisar dados. Neste contexto, a etapa de *data mining* é importante, pois efetivamente é ela que encontra

padrões e relações entre os dados, possibilitando-se a gestão do conhecimento (REZENDE, 2005).

Cada tarefa de *data mining* possui diferentes métodos de implementação dependendo do conhecimento que se deseja extrair, esses métodos estão implementados em diferentes ferramentas, denominadas *Shells*, utilizadas para auxiliar na descoberta de padrões e relações relevantes. A *Shell* Orion *Data Mining Engine* é um projeto do Grupo de Pesquisa em Inteligência Computacional Aplicada, do curso de Ciência da Computação na Universidade do Extremo Sul Catarinense (CASSETTARI JUNIOR, 2008).

Essa ferramenta usa bases de dados principalmente na área da saúde,o uso dessas bases tem gerado resultados satisfatórios, mas existe a necessidade de validar o conhecimento descoberto pela aplicação dessas bases, com intuito de torná-las mais confiáveis quanto as respostas obtidas, pois resultados com conhecimento incorreto ou irrelevante podem dificultar a localização daqueles que são úteis. Além disso, também é importante ressaltar que a percepção de uma incorreção pode gerar desconfiança da correção dos demais conhecimentos obtidos.

A *Shell* Orion é uma ferramenta gratuita desenvolvida em ambiente acadêmico, sendo importante dar continuidade ao projeto de desenvolvimento. Como ela já possui implementados diversos algoritmos de *data mining* que possibilitam a descoberta de conhecimento em bases de dados, é necessário que se valide o conhecimento descoberto e se avalie o conhecimento de alguns algoritmos presentes na Orion, a fim de se analisar a confiabilidade e eficácia da ferramenta que encontra-se em desenvolvimento.

Esse processo envolve o domínio de procedimentos de avaliação de desempenho, para isso a estatística oferece métodos e modelos que permitem analisar a sensibilidade do conhecimento obtido. A análise dos resultados depende do paradigma adotado no desenvolvimento do sistema. A *Shell* Orion é baseada em paradigmas de aprendizado de

máquina, por isso tem a vantagem de descobrir conhecimento, relações e padrões em grandes bases de dados de modo eficaz.

1.4 ESTRUTURA DO TRABALHO

Este trabalho possui em sua totalidade cinco capítulos. A partir do primeiro capítulo, apresenta-se a introdução, os objetivos gerais e específicos, assim como a justificativa.

No Capítulo 2 é apresentado conceitos e as etapas da descoberta de conhecimento em bases de dados, as tarefas os métodos e os conceitos de *data mining*, bem como a apresentação da *Shell* orion *Data Mining Engine*, sua interface e algoritmos implementados.

O Capítulo 3 aborda a avaliação do desempenho em data mining, e os métodos utilizados para a avaliação e validação.

Os trabalhos correlatos a esta pesquisa encontram-se no Capítulo 4.

O desenvolvimento no trabalho proposto e as etapas compreendidas na realização da avaliação do desempenho, estão apresentadas no Capítulo 5, sendo seguido pela conclusão.

2 DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS

O desenvolvimento da computação juntamente com a Tecnologia da Informação possibilitou que empresas e instituições armazenassem bases de dados, cada vez mais robustas, podendo atingir proporções de gigabytes ou terabytes. Esses dados atraíram a atenção das indústrias nos últimos anos, pois nessas bases estão contidos conhecimentos úteis. Estes conhecimentos podem ser usados, por exemplo, como diferencial competitivo, na análise mercadológica e no controle de produção, entre outros (HAN; KAMBER, 2001, tradução nossa).

Existe uma diferença fundamental entre dados, informação e conhecimento. Os dados são elementos brutos, sem significado que podem ser captados e transformados em informação. As informações são esses dados processados e que passam a ter significado e contextos definidos, já o conhecimento é o uso útil dessa informação para o domínio da aplicação, podendo ser aplicado na prática (GOLDSCHMIDT; PASSOS, 2005).

A expansão das bases de dados tornou inviável a análise por um especialista humano, pois esse seria incapaz de imaginar todas as relações existentes entre esses dados. O processo de Descoberta de Conhecimento em Bases de Dados (DCBD) ou *Knowledge Discovery in Databases* (KDD), objetiva encontrar conhecimento a partir de dados armazenados em grandes bases de dados, sendo compreensível e podendo auxiliar em tomadas de decisões (REZENDE, 2005).

A DCBD é composta por várias etapas operacionais, tais como o pré-processamento, *data mining* e o pós-processamento, apresentados conforme Figura 1.

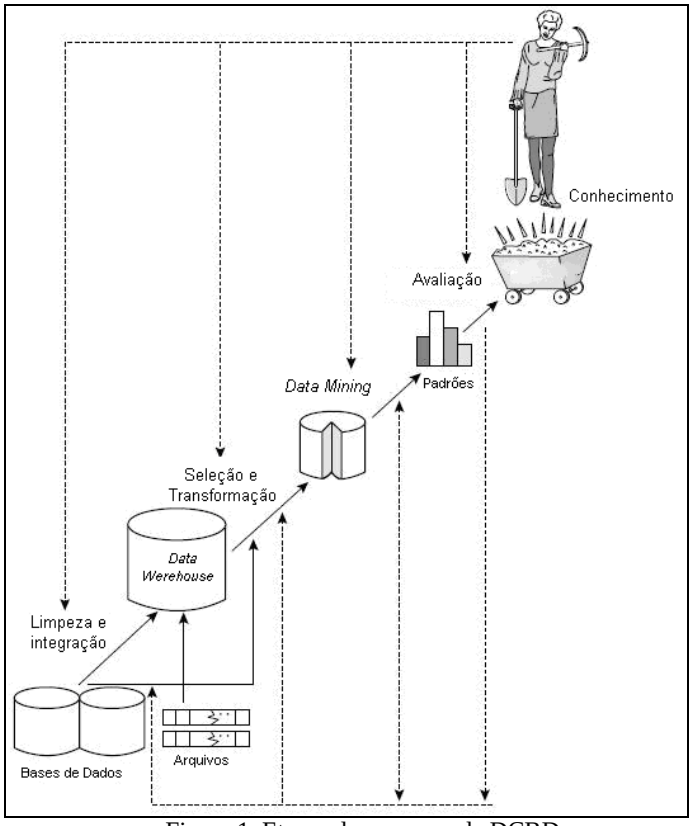


Figura 1. Etapas do processo de DCBD
Fonte: Extraído de HAN, J; KAMBER, M. (2001)

Na etapa de pré-processamento os dados brutos são tratados e preparados para o *Data Mining*, onde são descobertos padrões e relações entre esses dados, importantes no contexto da aplicação, para isso são utilizados algoritmos de *data mining*. No pós-processamento é analisada a sensibilidade dessas descobertas de conhecimento, assim como o tratamento visual (SILVA, 2002).

2.1 PROCESSO DE DESCOBERTA DO CONHECIMENTO EM BASES DE DADOS

No DCBD é preciso que se conheça bem o domínio da aplicação, os objetivos que se desejam atingir e a aplicação final do conhecimento extraído. É um processo não trivial que objetiva encontrar em dados, padrões que sejam válidos, novos, que possam ser úteis e compreensíveis (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996, tradução nossa).

Os dados armazenados, muitas vezes, não estão prontos para a aplicação de algoritmos de extração de padrões, por isso é indispensável processos como o de tratamento, limpeza e redução dos dados armazenados, cada etapa busca aperfeiçoar os resultados e prepará-los para a seguinte.

O pré-processamento inicia pela extração e integração, que seleciona os dados a serem analisados, podendo não estarem num mesmo formato ou local, então é necessário efetuar a integração para depois submeter ao processo de extração de padrões. Outro tratamento é a transformação que visa facilitar a aplicação dos algoritmos, e a limpeza que objetiva a exclusão de dados errôneos, em branco ou com ruídos. Na redução são selecionados os dados que realmente interessam no contexto da aplicação, diminuindo a quantidade para facilitar o processamento, esse processo pode ser feito por meio de amostragem[2] (REZENDE, 2005).

A etapa onde busca-se novos padrões relevantes, identificando relações e padrões em grandes bases de dados é a de *Data Mining* (DM). Esta é considerada a fase mais importante de DCBD, onde existem diversas tarefas que realizam a extração do conhecimento. A escolha dos algoritmos depende diretamente do tipo de problema a que ele vai ser aplicado (HAN; KAMBER, 2001, tradução nossa).

No pós-processamento onde são mostradas as saídas de dados, pode-se analisar os padrões encontrados na etapa de DM. Nessa fase é possível verificar a sensibilidade do conhecimento extraído. Envolve técnicas de interpretação, visualização e representação de conhecimento, que são apresentados ao usuário (GOLDSCHMIDT; PASSOS, 2005).

Existem três tipos de usuário, o especialista do domínio que tem amplo conhecimento sobre o campo da aplicação, o analista que é o responsável pela extração e execução do conhecimento e o usuário final que vai utilizar o conhecimento extraído. Esse

[2] Seleção de elementos aleatoriamente, que represente a população total da base de dados (REZENDE, 2005).

usuário pode muitas vezes ser o próprio especialista (REZENDE, 2005).

A área de DCBD é caracterizada por ser multidisciplinar, pois envolve conceitos e conhecimentos de Banco de Dados[3], Inteligência Artificial[4], Aprendizado de Máquina[5] e Estatística[6]. A relevância do conhecimento extraído depende diretamente da área de aplicação e está relacionada com a utilidade, originalidade e compreensão.

A principal etapa do processo de DCBD é a de *data mining*, essa etapa forma o centro do processo podendo muitas vezes ser confundida com ele.

2.2 *DATA MINING*

Juntamente com o crescimento do volume de dados que se pode armazenar em computadores, torna-se necessário a compreensão dos dados contidos nesses repositórios, que são ricos em informação e bastantes complexos. Esse fato tem causado interesse em diversos ramos de atuação, tais como Negócios, Ciência, Medicina e Engenharia. *Data mining* é uma metodologia que por meio de tarefas e métodos extrai conhecimento a partir dos dados (KANTARDZIC, 2003, tradução nossa).

A definição científica de DM é bastante discutida por pesquisadores que discordam sobre fundamentos e limites da tecnologia. Uma definição que deve ser considerada é "KDD é o processo não-trivial de identificar, em dados, padrões válidos, novos, potencialmente úteis e ultimamente compreensíveis" (Fayyad et al, 1996, tradução nossa).

Segundo Hand, Heikki e Smith (2001, tradução nossa) DM tem o objetivo de

[3] Banco de dados é um conjunto de informações que normalmente agrupa registros utilizáveis para um mesmo fim (HAM; KAMBER, 2001).
[4] Inteligência Artificial é a área da ciência da computação orientada ao entendimento, construção e validação de sistemas inteligentes (Rich; Knight, 1994).
[5] É um sub-campo da inteligência artificial dedicado ao desenvolvimento de algoritmos e técnicas que permitam ao computador aprender (HAM; KAMBER, 2001).
[6] Um conjunto de técnicas e métodos de pesquisa que entre outros tópicos envolve o planejamento do experimento a ser realizado, a coleta qualificada dos dados, a inferência, o processamento, a análise e a disseminação das informações (IBGE, 2010).

analisar as bases de dados (na maioria das vezes grandes) a fim de encontrar novas relações e resumir essas bases em fragmentos compreensíveis e úteis aos analistas. A partir dessas relações, e resumos que são encontrados quando aplicadas as ferramentas e tarefas de DM, surgem os modelos e padrões de conhecimento.

Data mining trata-se de um processo iterativo e interativo. Iterativo pois a evolução depende da descoberta, e interativo uma vez que é o resultado do esforço entre computadores e seres humanos. O resultado final depende da capacidade do especialista em definir seus objetivos e problemas aliado a capacidade de processamento de pesquisa dos computadores (KANTARDZIC, 2003, tradução nossa).

Existem ferramentas e tarefas de *data mining*, com diferentes modelagens, para encontrar conhecimento em bases de dados, essas estruturas se diferem pelos algoritmos, tarefas e métodos que são implementados.

2.2.1 Tarefas De *Data Mining*

As tarefas de DM podem ser classificadas como preditivas e descritivas. A preditiva utiliza campos ou variáveis com o objetivo de prever dados futuros, determinando novos atributos, como exemplo, pode-se citar a classificação, estimativa e as séries temporais. Enquanto a descritiva produz padrões novos e válidos visando a compreensão humana, são exemplos a clusterização e a associação (KANTARDZIC, 2003, tradução nossa).

No processo de aquisição do conhecimento é indispensável a escolha da tarefa a ser utilizada de acordo com o objetivo a ser atingido. Essas tarefas são:

a) **associação**: identifica padrões que se repetem frequentemente, levando a correlações entre os dados. Um exemplo seria a análise dos produtos que são vendidos juntos num supermercado (HAN; KAMBER, 2001, tradução nossa);

b) **classificação**: o objetivo da classificação é identificar um grupo a qual cada caso pertence. Um exemplo na área da telefonia, seria a realização de chamadas a longa distância, com base na duração da chamada e na freqüência que ela é realizada, poderia se oferecer uma promoção de minutos ao cliente (TWO CROWS CORPORATION, 1999);

c) **regressão**: na regressão são analisados dados numéricos por meio de modelos estatísticos. É possível analisar projeções feitas a partir de um dado inicial. Por exemplo, prever o volume de venda, preços e taxas de insucesso. Essa tarefa pode-se tornar difícil, por depender de complexas interações de múltiplas variáveis (TWO CROWS CORPORATION, 1999);

d) **clusterização**: no processo de clusterização os dados são agrupados de acordo com suas semelhanças, dessa forma dados de um mesmo *cluster* são muito parecidos entre si e dados de *clusters* diferentes são bastante diferentes (HAN; KAMBER, 2001, tradução nossa);

e) **previsão das séries temporais**: analisa os eventos ocorridos ao longo do tempo com base em um evento inicial. Por exemplo, um consumidor que adquiriu um microcomputador hoje, seis meses depois adquiriu um DVD-ROM. Assim, é possível fazer promoções direcionadas a esses consumidores (NAVEGA, 2002).

O conhecimento do especialista em sua área de atuação é de grande importância para a escolha adequada da tarefa de acordo com o domínio da aplicação, e consequentemente na obtenção de resultados mais eficientes e com maior qualidade (KANTARDZIC, 2003, tradução nossa). Essas tarefas podem conter diferentes métodos de implementação, cada método é desenvolvido por algoritmos específicos.

2.2.2 Métodos De *Data Mining*

Os métodos de *data mining* são implementações específicas, um método é implementado por meio de algoritmos específicos. Alguns métodos bastantes populares são:

a) **árvores de decisão**: são estruturas utilizadas quando se deseja dividir os dados em grupos menores, o topo da árvore contém o processo e os nós as variáveis que podem ser a saídas com as medidas a serem tomadas, usando para isso regras de decisão, os subgrupos tornam-se mais homogêneos entre si (CARVALHO, 2005);

b) **redes neurais**: as redes neurais artificiais usam modelos matemáticos baseados nos neurônios e no funcionamento do cérebro humano. Estas redes manipulam informação por meio de interações de unidades de processamento na rede neural artificial (GOLDSCHMIDT; PASSOS, 2005);

c) **lógica *fuzzy***: usa raciocínios aproximados, baseando-se em pertinência, é uma extensão da lógica clássica de conjuntos, que aceita valores no intervalo [0 e 1]. A lógica *fuzzy* possibilita trabalhar com conceitos imprecisos ou indeterminados (REZENDE, 2005);

d) **algoritmos genéticos**: baseados na teoria da evolução e na reprodução genética, usa a combinação de dois estados iniciais para gerar o próximo. É utilizado na resolução de problemas complexos de otimização, com muitas variáveis e restrições ou quando se tem um grande espaço de busca (RUSSELL;NORVIG, 2004).

As tarefas, métodos e algoritmos de DM encontram-se implementadas em aplicações denominadas *Shell.* Essas aplicações em sua grande maioria são comerciais. Pensando nisso o Grupo de Pesquisa em Inteligência Computacional Aplicada do Curso de

Ciência da Computação da UNESC, vem desenvolvendo ao longo dos anos por acadêmicos em seus trabalhos de conclusão de curso a *Shell* Orion *Data Mining Engine*.

2.3 *SHELL* ORION *DATA MINING ENGINE*

O projeto da *Shell* Orion *Data Mining Engine* teve início no ano de 2005, com a implementação das tarefas de classificação e associação. Desenvolvido pela Bacharel Diana Colombo Pelegrin, com o método de indução de árvores de decisão pelo algoritmo ID3 em seu Trabalho de Conclusão de Curso, intitulado: A Tarefa de Classificação e o Algoritmo ID3 para Indução de Árvores de Decisão na *Shell* de *Data Mining* Orion, e Diego Paz Casagrande com a implementação do menu principal e a tarefa de associação pelo algoritmo *Apriori*, com o trabalho: O Módulo da Tarefa de Associação Pelo Algoritmo *Apriori* no Desenvolvimento da *Shell* de *Data Mining* Orion (BORTOLOTTO, 2007).

Ao longo dos semestres foram desenvolvidas outras funcionalidades. Em 2007 a Bacharel Lidiane Rosso Raimundo desenvolveu o trabalho intitulado: O Algoritmo CART pelo Critério de Gini na Tarefa de Classificação da *Shell* Orion *Data Mining Engine*.

No mesmo ano o Bacharel Dênis Piazza Martins desenvolveu em seu trabalho a tarefa de clusterização, intitulado: O Algoritmo de Particionamento K-Means na Tarefa de Clusterização da *Shell* Orion *Data Mining Engine*. E o Bacharel Leandro Sehnem Bortolotto desenvolveu o trabalho intitulado: O Método de Redes Neurais pelo Algoritmo de Kohonen para Clusterização na *Shell* Orion *Data Mining Engine*.

Em 2008 o Bacharel José Márcio Cassettari Junior desenvolveu em seu TCC o trabalho intitulado: O Método de Lógica Fuzzy pelo Algoritmo Gustafson-Kessel na Tarefa de Clusterização da *Shell* Orion *Data Mining Engine*.

Em 2009 o Bacharel Daniel Perego desenvolveu em seu trabalho outro algoritmo da tarefa de clusterização, intitulado: O Método de lógica Fuzzy pelo Algoritmo Gath-Geva na Tarefa de Clusterização da Shell Orion Data Mining Engine.

Também em 2009 o Bacharel Ricardo Lineburger Mondardo, desenvolveu o trabalho intitulado: O Algoritmo C4.5 na Tarefa de Classificação na Shell Orion de Data Mining Engine.

Inicialmente a *Shell* Orion possuía conexão com os Sistemas Gerenciador de Banco de Dados (SGBD) PostgressSQL e Firebird. Em 2008 foi proposta uma nova interface que possibilitou a conexão com qualquer banco de dados, necessitando apenas do *driver* de conexão do banco (CASSETTARI JUNIOR, 2008).

Ao configurar o *driver* e executar a conexão corretamente, o usuário pode selecionar a tabela para executar a tarefa de *data mining*. Considerando que cada tarefa possui características próprias, deve-se selecionar a que mais se identifica com o objetivo que se deseja atingir.

Até o momento a *Shell* Orion possui implementadas as respectivas tarefas, métodos e algoritmos conforme apresentados na Tabela 1 (CASSETTARI JUNIOR, 2008).

Tabela 1. Algoritmos implementados na *Shell* Orion *Data Mining Engine*

Ano	Tarefa	Algoritmo	Método	Atributos	Referência
2005	Associação	*Apriori*	Regras de associação	Numéricos	(CASAGRANDE, 2005)
2005	Classificação	ID3	Árvores de Decisão	Nominais	(PELEGRIN,2005)
2007	Classificação	CART	Árvores de Decisão	Nominais e Numéricos	(RAIMUNDO, 2007)
2007	Clusterização	*K-Means*	Particionamento	Numéricos	(MARTINS, 2007)
2007	Clusterização	*Kohomen*	Redes Neurais	Numéricos	(BORTOLOTTO, 2007)
2008	Clusterização	*Gustafson-Kessel*	Lógica Fuzzy	Numéricos	(CASSETTARI JUNIOR, 2008)
2009	Clusterização	*Gath-Geva*	Lógica Fuzzy	Numéricos	(PEREGO, 2009)
2009	Classificação	*C4.5*	Árvores de Decisão	Nominais e Numéricos	(MONDARDO, 2009)

Fonte: Adaptado de CASSETARI JUNIOR, J. M.. (2008)

Esta ferramenta está sendo desenvolvida por meio da tecnologia Java, que foi escolhida principalmente por ser multiplataforma, possuir ferramentas de desenvolvimento gratuitas e ter a possibilidade de reutilização de código, visto que a *Shell* está em constante evolução (PELEGRIN, 2005).

2.3.1 Interface Da *Shell* Orion *Data Mining Engine*

A interface facilita a adição dos algoritmos desenvolvidos de forma organizada e clara, facilitando a navegação, como se pode visualizar na Figura 3.

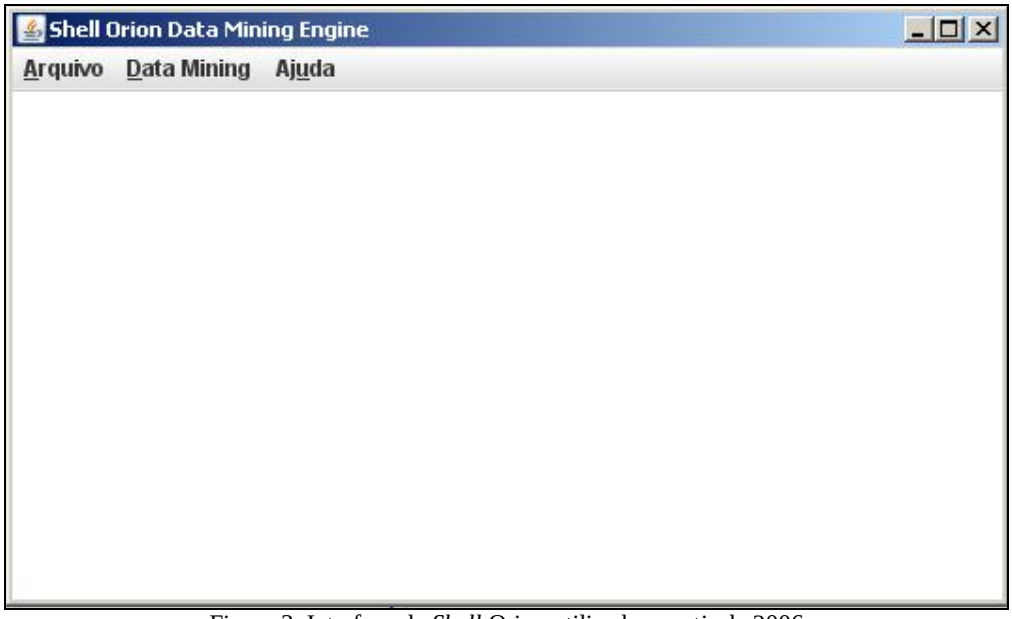


Figura 3. Interface da *Shell* Orion utilizada a partir de 2006

Ao iniciar a *Shell* o usuário necessita configurar o banco de dados, para isso é necessário efetuar uma conexão com o banco, hoje a *Shell* possui implementado conexões com os bancos de dados HSQLDB, PostGreSQL, MySQL e Firebird.

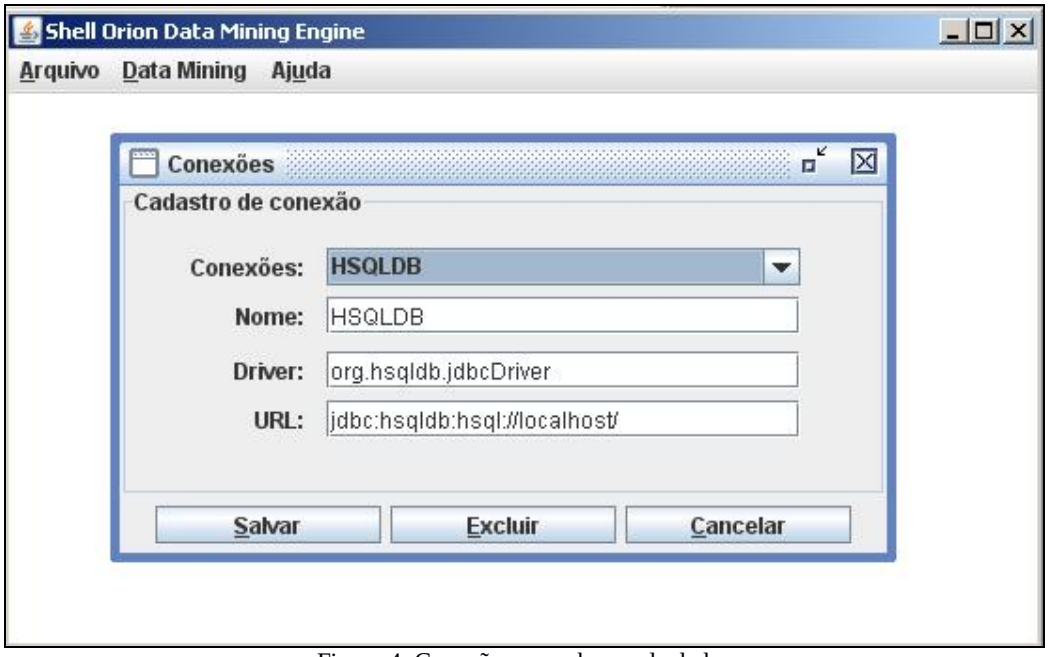


Figura 4. Conexão com o banco de dados

Após fazer a conexão com o banco o usuário pode então usar qualquer módulo da *Shell*, escolhendo entre as tarefas de associação, classificação e clusterização.

2.3.2 Algoritmos de *Data Mining* Implementados na *Shell* Orion *Data Mining Engine*

A seguir estão descritos os algoritmos implementados na *Shell* Orion *Data Mining Engine*, contemplando as tarefas de associação, classificação e clusterização.

a) ***Apriori***: o algoritmo *Apriori* emprega regras de associação para extrair o conhecimento. É bastante utilizado para descobrir *itemsets* freqüentes em regras de associação. O algoritmo *Apriori* é iterativo, onde forma *itemsets* de tamanho K para gerar o próximo conjunto de tamanho K+1, denominado L_1, e a partir dele encontrar o conjunto K+2, denominado L_2, e assim sucessivamente até esgotarem as chances de gerar conjuntos. O algoritmo *apriori* contempla a tarefa de associação na *Shell* Orion, pode-se ver um exemplo do algoritmo em execução na Figura 5. Esse algoritmo utiliza apenas

atributos numéricos (CASAGRANDE, 2005; HAN; KAMBER, 2001, tradução nossa);

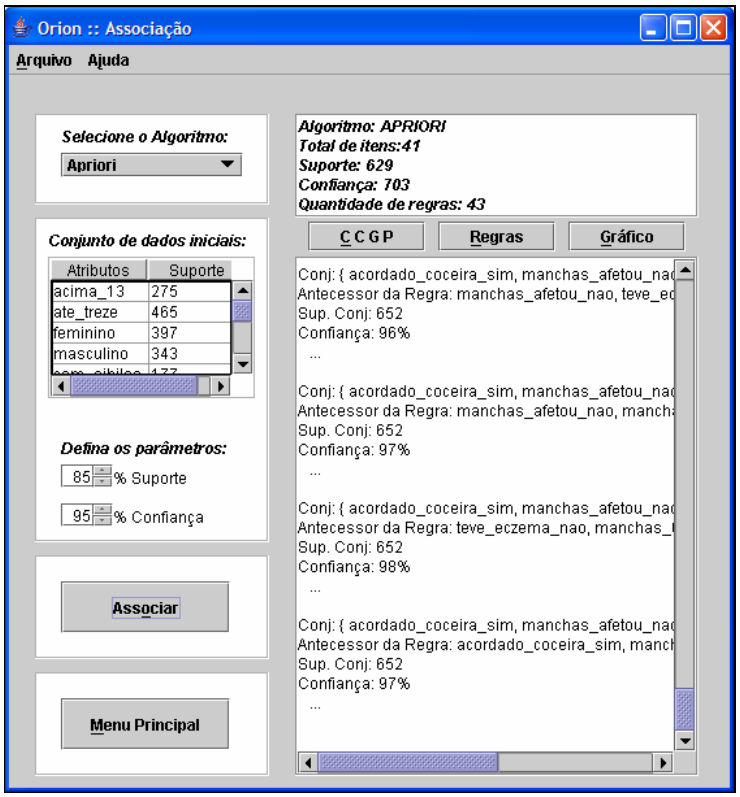


Figura 5. Resultado gerado pelo *Apriori* na *Shell* Orion
Fonte: CASAGRANDE, D. P. (2005)

b) **ID3**: é utilizado para a indução de árvores de decisão *top-down* (de cima para baixo), utiliza uma abordagem recursiva para encontrar a árvore mais simples. O algoritmo escolhe o atributo mais informativo para o nó raiz, a medida que a árvore vai sendo gerada, seus nós filhos vão se tornando mais homogêneos, isso também faz com que a árvore seja menos profunda. O algoritmo ID3 implementa a tarefa de classificação na *Shell* Orion, um exemplo dele em execução pode ser encontrado na Figura 6. O ID3 utiliza atributos nominais (CORDEIRO, 2003; HAN; KAMBER, 2001, tradução nossa; PELEGRIN, 2005);

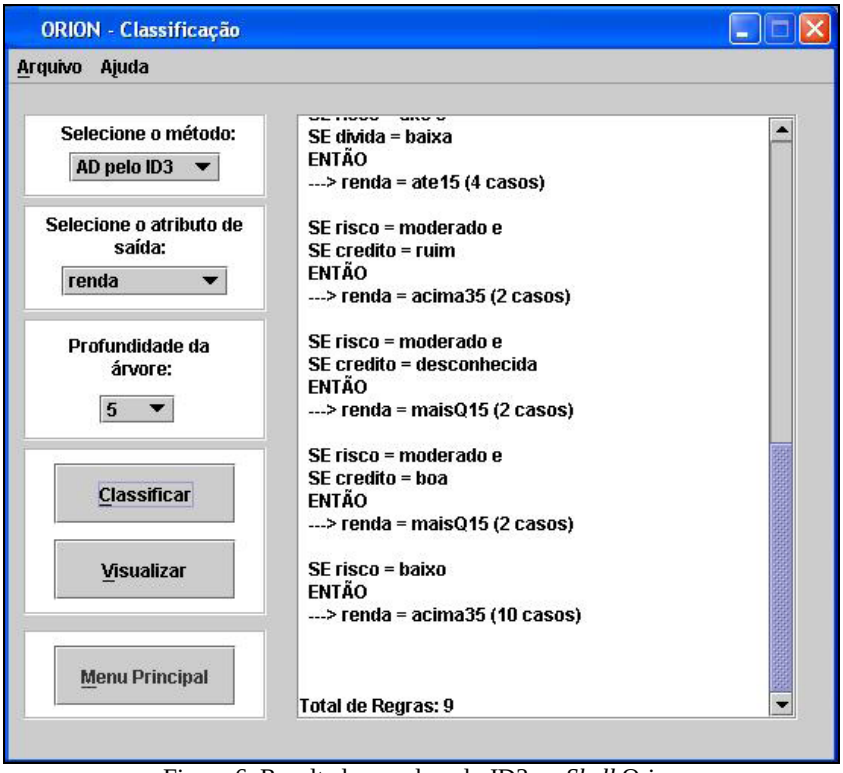


Figura 6. Resultado gerado pelo ID3 na *Shell* Orion
Fonte: PELEGRIN, D. C. (2005)

c) **CART**: o Classification and Regression Trees (CART) tem a capacidade de pesquisar relações entre os dados, mesmo quando essas não são evidentes, produzindo como resultado árvores de decisão binárias simples e legíveis. Podendo-se percorrer da raiz até as folhas respondendo questões como sim/não. É um algoritmo de partição binária recursiva, pois os nós são divididos em dois subconjuntos, aplicando a cada subconjunto recursivamente o algoritmo, isso ocorre até que não seja necessário efetuar nenhuma outra partição. O algoritmo de CART possibilita trabalhar com atributos nominais e numéricos, na Figura 7 encontra-se ilustrado o módulo de classificação pelo CART implementado na *Shell* Orion *Data Mining Engine* (FONSECA, 1994; RAIMUNDO, 2007);

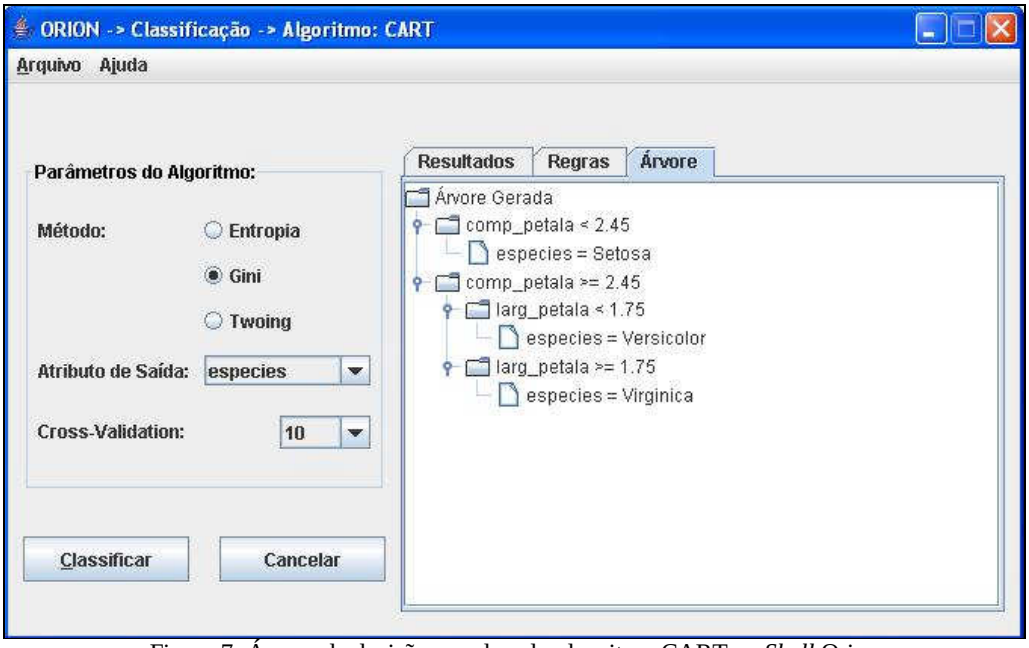


Figura 7. Árvore de decisão gerada pelo algoritmo CART na *Shell* Orion
Fonte: RAIMUNDO, L. R. (2007)

d) ***K-Means***: o algoritmo *k-means* implementa a tarefa de clusterização, podendo somente trabalhar com atributos numéricos. O algoritmo pode ser iniciado com a definição aleatória de K pontos de dados como sendo os centróides do *cluster*. Divide os objetos da base entre os *clusters* de acordo com sua similaridade e de modo que os objetos tenham menos valor de distância entre ele e o centróide. Então um novo centróide é calculado para cada *cluster* com base na média de pontos dos *clusters*, esse processo é repetido até que os centróides não se modifiquem mais ou até atingir o número de iterações especificadas pelo usuário. A Figura 8 mostra o algoritmo em execução na *Shell* Orion (KANTARDZIC, 2003, tradução nossa) (MARTINS, 2007);

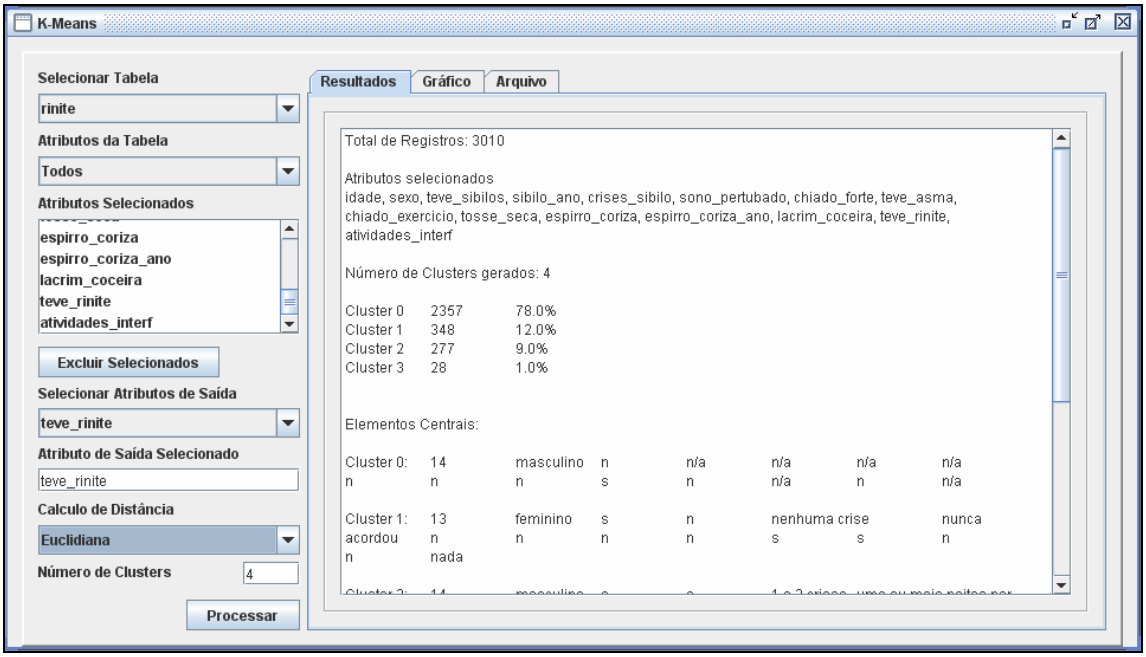


Figura 8. Resultado gerado pelo K-Means na *Shell* Orion
Fonte: MARTINS, D. P. (2007)

e) ***Kohonen***: é um mapa auto-organizável de treinamento não supervisionado, baseado geralmente em uma competição entre os elementos processadores. Seu funcionamento consiste em dividir o conjunto de entrada em grupos com características semelhantes, considerando que os neurônios de saída da rede competem entre si, e apenas um neurônio é o vencedor. Para isso, dispõe os neurônios em uma única camada onde todos recebem os valores do padrão de entrada e somente o neurônio vencedor terá os pesos atualizados, assim ele se aproxima do vetor de entrada. Esse algoritmo trabalha apenas com atributos numéricos. A Figura 9 apresenta o módulo de *Kohonen* em execução na *Shell* Orion (BORTOLOTTO, 2007; GOLDSCHMIDT; PASSOS, 2005);

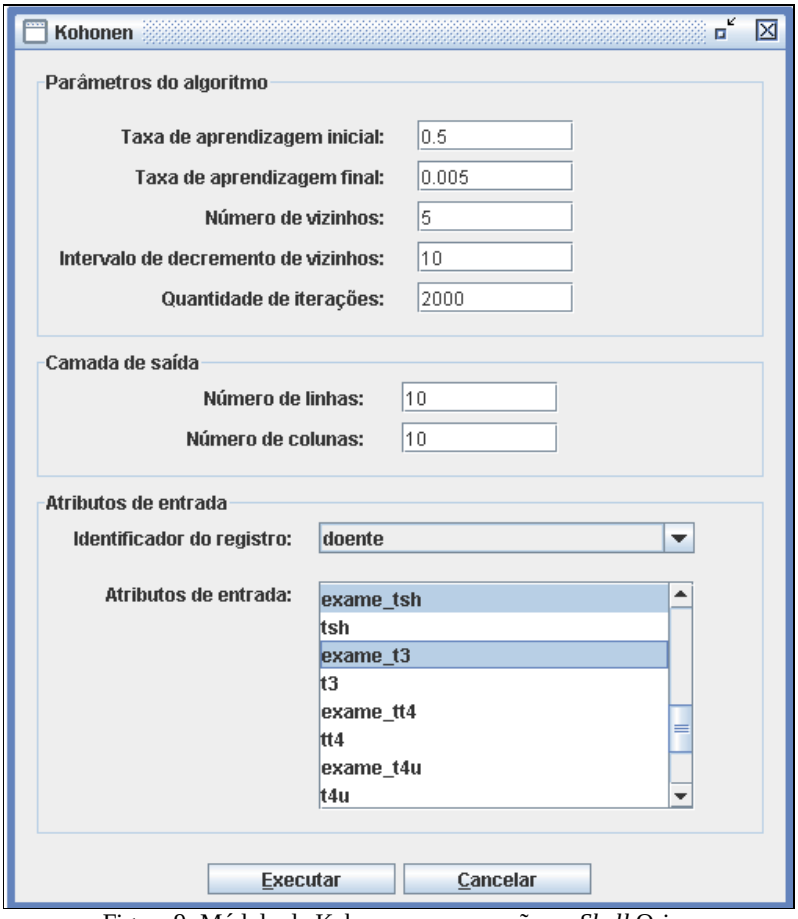


Figura 9. Módulo de Kohonen em execução na *Shell* Orion
Fonte: BORTOLOTTO, L. S. (2007)

f) ***Gustafson-Kessel***: o algoritmo G*ustafson-Kessel* utiliza uma matriz de covariância *fuzzy*, que contém as medidas de distância entre os grupos de elementos, para calcular as relações entre os dados da base de dados e atualizar a distância. Esta matriz permite a adaptação do algoritmo a diferentes formas de *clusters*, assim cada um pode ter suas próprias características de dimensões. O processo de clusterização utiliza apenas operações matemáticas, por esse motivo o algoritmo *Gustafson-Kessel* somente utiliza atributos numérico. A Figura 10 demonstra o algoritmo em execução na *Shell* Orion (CASSETTARI JUNIOR, 2008; GUSTAFSON; KESSEL, 1979).

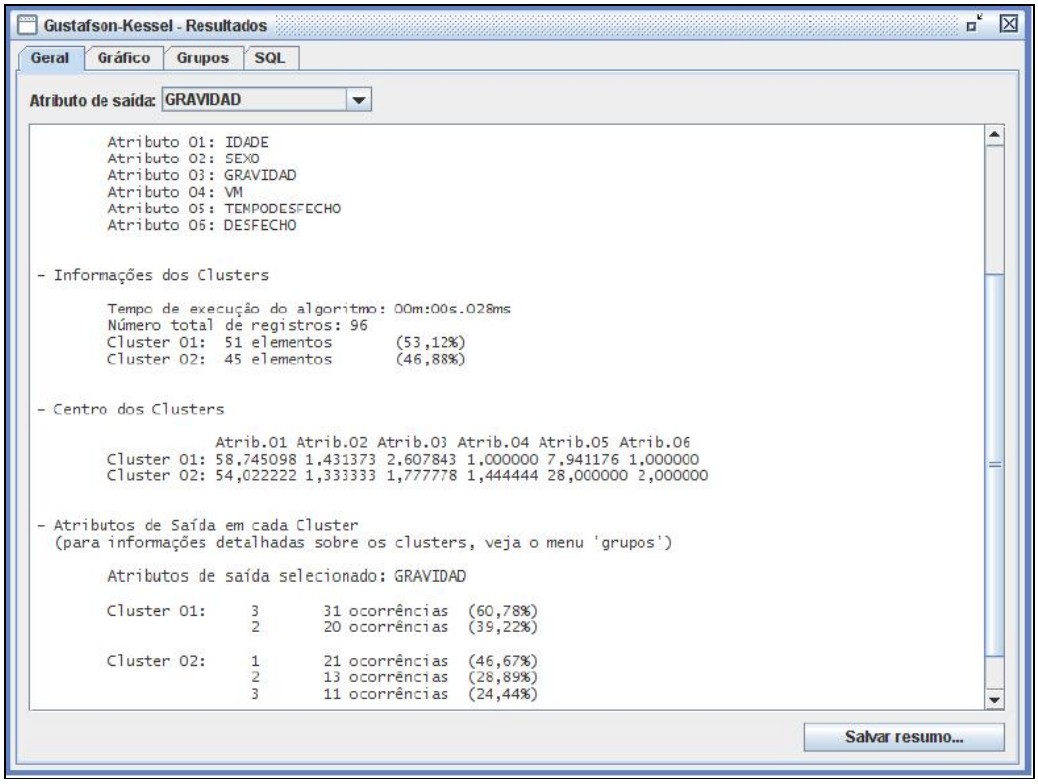


Figura 10. Resultado gerado pelo *Gustafson-Kessel* na *Shell* Orion
Fonte: CASSETTARI JUNIOR, J. M. (2008)

g) ***Gath-Geva***: o algoritmo *Gath-Geva* é uma modificação dos algoritmos *Fuzzy*, *K-Means* e *Gustafson-Kessel*. Esse algoritmo é considerado uma extensão do algoritmo *Gustafson-Kessel*, pois ele tem a possibilidade de calcular *clusters* de tamanhos e densidades diferentes, sendo possível fazer isso usando medidas diferentes nas distâncias dos *clusters* ou por uma adequada definição. Esse algoritmo apresenta uma distância exponencial e utiliza uma matriz de covariância para obter *clusters* diferentes entre si. A Figura 11 mostra os resultados obtidos quando executados o algoritmo na *Shell* Orion (PEREGO, 2009).

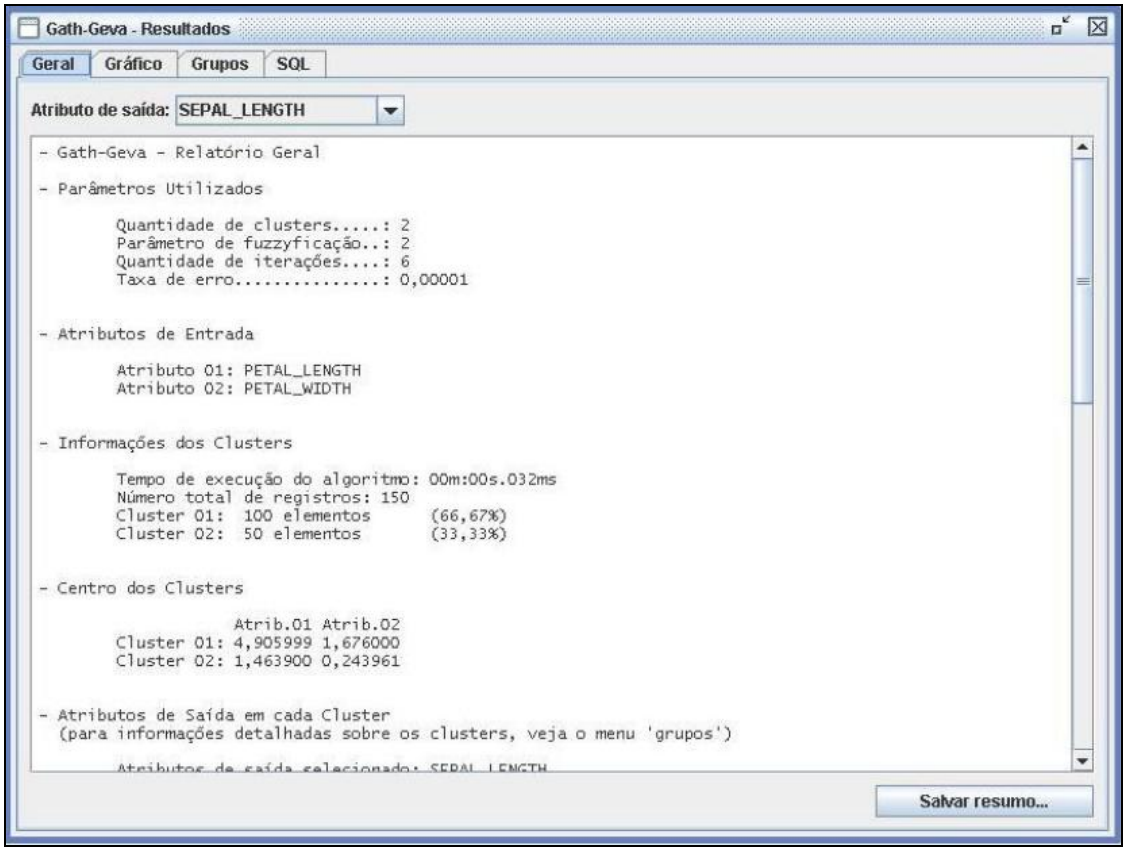


Figura 11. Resultado gerado pelo Gath-Geva na *Shell* Orion
Fonte: PEREGO, D. (2009)

h) **C4.5:** o algoritmo C4.5 utiliza indução de árvores de decisão para realizar a classificação. O C4.5 é uma evolução do algoritmo ID3, sendo a principal diferença entre os dois algoritmos a capacidade do C4.5 de simplificação da árvore, excluindo as regras sem valores significativos para a classificação. A Figura 12 mostra os resultados obtidos quando executados o algoritmo na *Shell* Orion (MONDARDO, 2009).

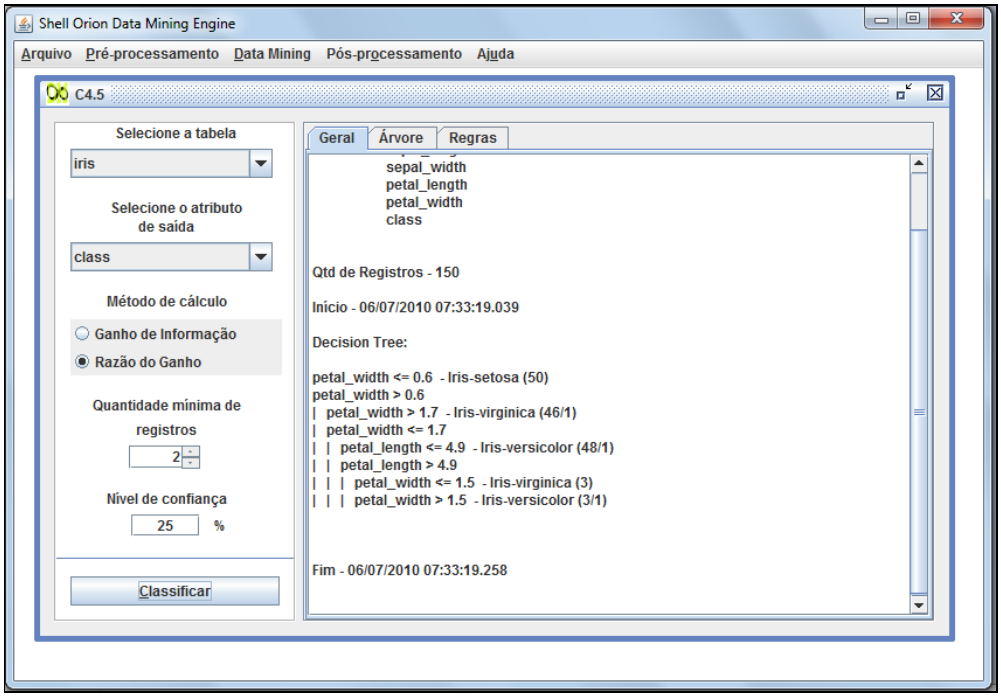


Figura 12. Módulo C4.5 em execução na *Shell* Orion
Fonte: MONDARDO, R. (2009)

Torna-se importante ressaltar que os algoritmos aqui citados não esgotam os métodos de DM que podem ser aplicados em cada tarefa de *data mining* (GOLDSCHMIDT; PASSOS, 2005). A Tabela 2 ilustra outros dos possíveis métodos, tarefas e algoritmos.

Tabela 2. Métodos de DM que podem ser aplicados em cada tarefa de DCBD

Tarefas de DCBD	**Métodos de Data Mining**
Descoberta de Associações	Basic, Apriori, DHP, Partition, DIC, ASCX-2P
Descoberta de Sequências	GSP, MSDD, SPADE
Classificação	Redes Neurais (Ex. Back-Propagation, RBF), C4.5, Rough Sets, Algoritmos Genéticos (Ex. Rule Evolver), CART, K-NN, Classificadores Bayseanos
Regressão	Redes Neurais (Ex. back-Propagation), Lógica Nebulosa
Sumarização	C4.5, Algoritmos Genéticos (Ex. Rule Evolver)
Clusterização	K-Means, K-Modes, K-Prototypes, Fuzzy K-Means, Algoritmos Genéticos, Redes Neurais (Kohonen)
Previsão de Séries Temporais	Redes Neurais (Ex. back-Propagation), Lógica Nebulosa (Ex. Wang_Mendel)

Fonte: Adaptado de GOLDSCHMIDT, R.; PASSOS, E. (2005)

Com a descoberta de conhecimento a partir desses algoritmos surge a necessidade e avaliar os resultados obtidos com a execução de cada tarefa, método e algoritmo implementado na *Shell* Orion *Data Mining Engine*, esta avaliação é feita por meio de métodos estatísticos, sendo também necessária a revisão qualitativa do especialista da área da base de dados aplicada.

3 AVALIAÇÃO DE DESEMPENHO EM *DATA MINING*

É importante adotar uma metodologia de avaliação dos algoritmos a fim de se obter uma estimativa de desempenho, bem como suas vantagens e limitações. Para isso o uso de metodologias baseadas em amostragem é bastante difundido (REZENDE, 2005).

Na tarefa de estimar uma medida de precisão da base de dados, comumente são utilizados dois conjuntos disjuntos, onde um conjunto é o de treinamento e o outro o de teste, isso garante a validade estatística dos parâmetros, conforme Figura 3.

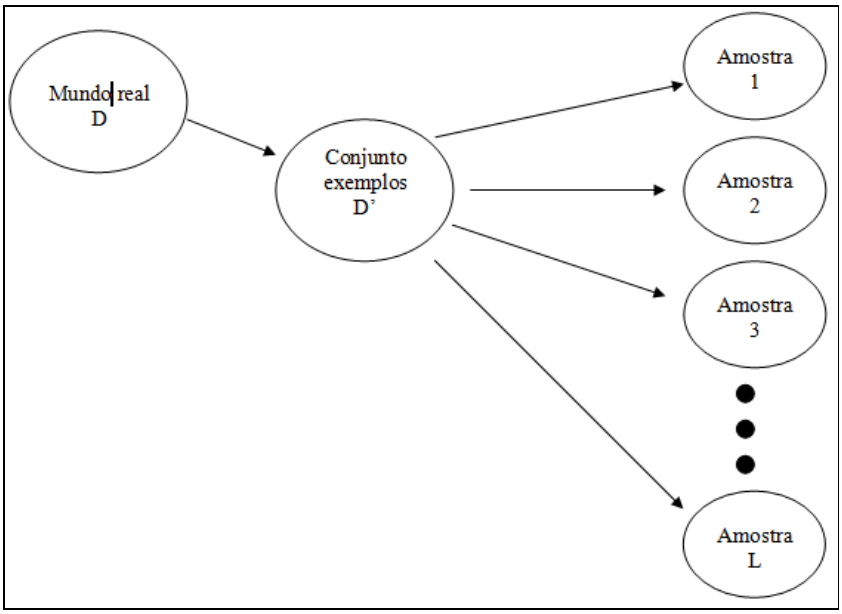


Figura 13. Idéia de amostragem
Fonte: Adaptado de REZENDE, S. (2005)

Onde o mundo real representa exemplos de um dado domínio desconhecido, extraindo exemplos do mundo real obtém-se um conjunto de exemplos D', então treina-se o indutor com amostras extraídas a partir do D`. Testando-se seu desempenho em exemplos extraídos de D`, normalmente com exemplos fora da amostra utilizada para o treinamento. Na estimação de uma medida é importante que as amostras sejam aleatórias, assim obtêm-se medidas verdadeiras. Os métodos existentes para a tarefa de estimar uma medida são

descritos, resumidamente, na Tabela 3. Onde *n* é o numero de exemplos , *r* o número de partições, *p* um número tal que 0<p<1, t um número tal que 0<t<n e L o número de hipóteses induzidas (REZENDE, 2005).

Tabela 3. Parâmetros típicos de estimadores.

	Holdout	***Aleatória***	***Leave-One-Out***	***r-Fold CV***	***r-Fold-Strat CV***	***Boostrap***
Treinamento	*pn*	*t*	*n-1*	*n(r-1)/r*	*n(r-1)/r*	*n*
Teste	*(1-p) n*	*n-t*	*1*	*n/r*	*n/r*	*n-t*
Iterações	*1*	*L « n*	*n*	*r*	*r*	*200*
Reposição	*não*	*não*	*não*	*não*	*não*	*sim*
Prevalência de Classe	*não*	*não*	*não*	*não*	*não*	*sim/não*

Fonte: Adaptado de REZENDE, S. O (2005)

Existe diversos métodos para estimar uma medida verdadeira, descritos a seguir:

a) **holdout:** O método hodout consistem em dividir a base de dados em dois subconjuntos, um para treinamento e outro para teste. Usualmente a base é dividida em 2/3 para treinamento e 1/3 para teste (KANTARDZIC, 2003, tradução nossa), conforme ilustra a Figura 14. A divisão dos dados em treinamento e teste deve ser de forma aleatória, devendo ter a preocupação de se manter a proporção entre as classes nos dois conjuntos, caracterizando assim o hold out estratificado. O conjunto para teste deve ser separado do conjunto de treinamento e a validação dos dados será aplicada somente no conjunto de teste (KANTARDZIC, 2003, tradução nossa). Na aplicação do método hold out é necessário que a base de dados seja numerosa, um subconjunto de teste com no mínimo 1000 registros seria o adequado. Se esta divisão não for possível deve-se aplicar um outro método para validação, como o fator de confiança.

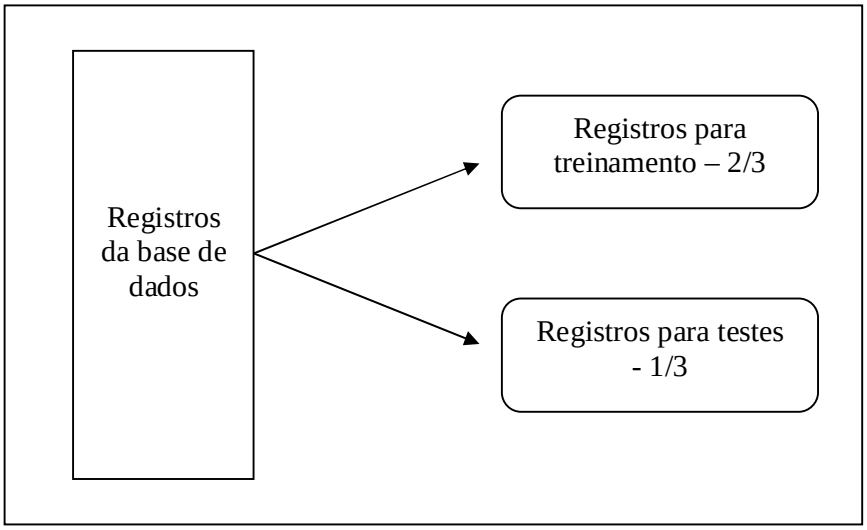


Figura 14. Divisão dos dados

b) **aleatória:** L hipóteses, L « *n*, são induzidas a partir dos conjuntos de treinamento, onde o erro final é a média dos erros de todas as hipóteses calculados em conjuntos de testes independentes e aleatórios (REZENDE, 2005);

c) ***cross-validation:*** o *cross-validation* é um meio termo entre o método *holdout* e o *leave-one-out*. Este método consiste em dividir a base em *r* partições exclusivas da tamanho aproximadamente igual e utilizar (*r* - 1) para treinamento e testar no subconjunto remanescente, repetindo este procedimento para cada subdivisão da base. Sendo o *cross-validation* a média dos erros de cada iteração (BARANAUSKAS, 2001).

d) ***leave-one-out*:** é um caso especial de *cross-validation*, onde as partições usadas nos testes são compostas de exemplos, sua precisão é estimada através da média das precisões de cada teste. É usada frequentemente quando se tem amostras pequenas por ser computacionalmente dispendioso (AMORIM, 2004);

e) ***r-fold cross-validation*:** os dados iniciais são aleatoriamente divididos em

r subconjuntos mutuamente exclusivos (*folds*), $S_1\ S_2 \ldots S_r$, com tamanhos aproximadamente iguais a *n*/*r*, realizando treinamento e testes *r* vezes. Os exemplos *r*-1 partições são usados para treinamento, sendo a hipótese induzida testada na partição remanescente. A estimativa de precisão é dada pelo número total de classificações corretas das iterações *r*, dividido pelo número total de amostras iniciais. É o método mais utilizado principalmente quando a base de dados possui um tamanho limitado (HAN; KANBER, 2001, tradução nossa);

f) ***stratified cross-validation*:** é um método similar ao *cross-validation*, mas a proporção de exemplos em cada uma das classes ao gerar os *folds* é considerada durante a amostragem (AMORIM, 2004; REZENDE, 2005);

g) ***boostrap*:** repete o processo de classificação diversas vezes utilizando um conjunto de treinamento, os exemplos que não estiverem no conjunto de treinamento pertencem ao conjunto de teste. A precisão é estimada pela média das precisões de cada iteração (AMORIM, 2004; REZENDE, 2005);

Na validação de um modelo e determinação de sua significância também deve ser levado em consideração as taxas de erro, porém as taxas de erro encontradas durante os testes são válidas somente para os dados em que modelo foi construído, a precisão pode mudar se esses dados se diferenciarem muito do modelo original (TWO CROWS CORPORATION, 1999, tradução nossa).

Uma maneira de determinar a precisão dos algoritmos de *data mining* é por meio do erro. Existem diferentes tipos de erro que são comumente relatados por uma matriz de confusão (Tabela 4). Uma matriz confusão contém o número de classificações corretas versus o número de relações preditas da classe em um conjunto de exemplos conforme ilustra a

Tabela 4, sendo muito utilizada em casos de regras de classificação para determinar a precisão e compreender os resultados obtidos (TWO CROWS CORPORATION, 1999, tradução nossa).

Tabela 4. Matriz Confusão

Classe	Predita C1	Predita C2	...	Predita Ck
Verdadeira C1	Nun(C1, C1)	Nun(C1, C2)	...	Nun(C1, Ck)
Verdadeira C2	Nun(C2, C1)	Nun(C2, C2)	...	Nun(C2, Ck)
...	...	...	...	...
Verdadeira Ck	Nun(Ck, C1)	Nun(Ck, C2)	...	Nun(Ck, Ck)

Fonte: Adaptado de REZENDE, S. (2005).

Em aplicações do mundo real é preferível projetar um algoritmo que minimize os custos de classificações incorretas ao invés de se diminuir a taxa de erro normal (REZENDE, 2005). Portanto um modelo com precisão global menor pode ser preferível a um modelo com maior precisão mas que possui um erro maior devido aos tipos de erros cometidos e as taxas de erros associadas a esses erros.Sendo necessário levar em consideração diferentes tipos de erro existentes e atribui-los custos, por isso é preciso saber mais sobre os tipos de erros e os custos associados a eles. (TWO CROWS CORPORATION, 1999, tradução nossa).

Além da matriz confusão para o desenvolvimento da avaliação do desempenho pode-se dividir em duas categorias básicas, os métodos subjetivos e objetivos.

a) **método subjetivo:** o usuário precisa estabelecer o conhecimento ou as crenças que o sistema utilizará como entrada no conjunto de padrões descoberto pelo algoritmo, buscando satisfazer o usuário. Isso implica que o usuário tenha um conhecimento prévio do domínio da aplicação (CARVALHO, D; NETO, W; BUENO, M, 2003).

b) **método objetivo:** não é necessário que se tenha um conhecimento previamente estabelecido. Isso facilita que o usuário entenda melhor o conjunto de regras descobertas a partir de uma base que contenha, por exemplo, dados de paciente com diabetes (CARVALHO, D; NETO, W; BUENO, M, 2003).

A descoberta de conhecimento em base de dados tem sido amplamente difundida no objetivo de encontrar padrões em bases de dados bastante robustas, porém esses padrões muitas vezes não contêm o conhecimento desejado, ou são muito óbvios. Por isso é importante se utilizar medidas que validem a importância dos padrões descobertos, para que tenham relevância e que agreguem conhecimento a área analisada (HAMILTON; HILDERMAN,1999, tradução nossa).

Nas próximas sessões encontram-se descritos os métodos para avaliação do desempenho utilizados no desenvolvimento deste trabalho, esses métodos são muito utilizados para a validação do desempenho da tarefa de classificação.

3.1 MÉTODOS DE AVALIAÇÃO

As regras descobertas por algoritmos de *data mining* devem ser avaliadas e validadas independente da técnica utilizada para isso. Um critério bastante adotado é a taxa de acerto, sendo uma forma bastante simples de se validar as regras, pois compara os dados classificados corretamente com os dados aplicáveis (ROMÃO, 2002).

$$\text{Taxa de acerto} = \frac{\text{Número de classificações corretas}}{\text{Número de dados aplicáveis}}$$

Este método pode não ser muito eficaz por não levar em consideração a distribuição de freqüência das classes que tende a ser desbalanceada, e ignorar também o custo de classificações erradas, pois diferentes tipos de erros podem ter diferentes custos (MICHIE et al., 1994; ROMÃO, 2002).

No complemento a este método de validação existem inúmeros outros métodos eficazes e mais completos, que podem ser aplicados com o objetivo de avaliar a classificação na descoberta do conhecimento.

3.1.1 Fator de Confiança

O fator de confiança é uma forma simples de se avaliar a precisão das regras, podendo ser calculado a partir da razão X/Y, com X sendo o número de registros que satisfazem o conseqüente e o antecedente da regra e Y o total de registros que satisfazem o antecedente da regra (ROMÃO, 2002).

Na utilização deste método, pode acontecer o fenômeno chamado de *overfitting*, que ocorre quando a regra descoberta ajusta-se demais as peculiaridades dos dados, gerando regras muito especificas, o *overfitting* acontece principalmente quando uma regra inclui poucos registros (BARANAUSKAS, 2001).

Um exemplo disso seria uma regra incluindo somente um registro corretamente, então teríamos X/Y=100% gerando um resultado indesejado já que a regra é muito especifica. Com o intuito de resolver este problema Freitas (1999), propôs em seu trabalho a utilização do fator -1/2, no numerando do cálculo do fator de confiança, ou seja:

Fator de confiança = (X – 1/2)/Y

A inserção do fator -1/2, penaliza as regras com cobertura de poucos registros não penalizando de forma significativa as regras com cobertura maior de registros, desta forma pode-se distinguir regras muito especializadas de regras generalizadas (ROMÃO, 2002).

3.1.2 Validação Cruzada

A validação cruzada permite a utilização de todos os dados na fase de treinamento e teste. Sendo bastante empregada quando se tem poucos dados. Este método consiste em dividir toda a base de dados em K subconjuntos iguais, preocupando-se em manter a proporção das classes (BARANAUSKAS, 2001).

A Figura 15 apresenta o algoritmo para a validação cruzada (ROMÃO, 2002):

```
Inicio
Para i = 1...k
    Executar o algoritmo de DM usando K-1
    subconjuntos de    dados;
    Computar a taxa de acerto, das regras obtidas,
    sobre os dados do subconjunto i não utilizado no
    passo anterior;
Fim Para;
Taxa de acerto = média das taxas de acerto nas K
iterações anteriores;
Fim.
```

Figura 15. Algoritmo para validação cruzada

No cálculo da média e desvio padrão[7] na validação cruzada pode-se empregar as seguintes equações (BARANAUSKAS, 2001):

$$\bar{a} = \frac{1}{k}\sum_{i=1}^{k} x_i$$

$$dp(\bar{a}) = \sqrt{\frac{1}{k}\left[\frac{1}{k-1}\sum_{i=1}^{k}(x_i - \bar{a})^2\right]}$$

onde:

a) $\bar{a}$ = média;

b) dp = desvio padrão;

c) x_i = taxa de acerto da partição i;

[7] Em Estatística, a média é o valor que aponta para onde mais se concentram os dados de uma distribuição o desvio padrão é a medida mais comum da dispersão (SPIEGEL, 1993).

d) K = número de partições.

Estas equações permitem estimar o acerto dos dados gerados a partir da tarefa de classificação, muito utilizado para a validação e avaliação do conhecimento descoberto (ROMÃO, 2002).

3.1.3 Matriz Confusão

A matriz confusão oferece uma medida efetiva da classificação em casos que o custo de validar uma regra errada é alto, como a classificação de falsos negativos e falsos positivos. Ela mostra em uma matriz de duas dimensões o número de classificações corretas e o número de classificações erradas, em relação as classes verdadeiras e as classes preditas conforme Figura 16 (BARANAUSKAS, 2001).

		Classe predita	
		Sim	Não
Classe real	Sim	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Figura 16. Matriz Confusão de ordem dois

O algoritmo de *data mining* deve ser executado a partir da base de dados e divididos em:

a) classificações “sim” realizadas corretamente (VP);

b) classificações “sim” realizadas erroneamente (FN);

c) classificações “não” realizadas corretamente (VN);

d) classificações “não” realizadas erroneamente (FP).

O número de acertos de cada classe encontra-se na diagonal principal e o número de erros são os elementos FP e FN, um classificador ideal tem esses elementos igual a zero, já que não comete erros (BARANAUSKAS, 2001).

A partir dos dados da matriz confusão é possível determinar medidas como sensibilidade e abrangência, acurácia, especificidade e precisão, obtidos a partir das expressões:

a) acurácia = $\frac{VP+VN}{n}.100\%$

b) erro = $\frac{FP+FN}{n}.100\%$

c) sensibilidade, abrangência = $\frac{VP}{VP+FP}.100\%$

d) especificidade = $\frac{VN}{VN+FP}.100\%$

e) precisão = $\frac{VP}{VP+FP}.100\%$

Onde:

a) $n = VP + VN + FP + FN$

As variáveis VP, VN, FP e FN, são contadores internos da matriz confusão, assim o contador VP totaliza as classificações “sim” classificadas corretamente e o FN por exemplo, totaliza as classificações “sim”classificadas erroneamente.

A sensibilidade é a capacidade da classificação gerar resultados verdadeiro positivos, sendo também conhecida como taxa de verdadeiros positivos. Em contrapartida a especificidade corresponde a taxa de verdadeiros negativos. Essas medidas se tornam importantes na avaliação do desempenho, pois demonstram a quantidade de resultados positivos serem classificados como positivo dentro da aplicação do domínio, e também a

quantidade de classificações negativas serem realmente classificadas como negativas.

3.1.4 Cálculo do Erro Médio

A taxa de erro de um classificador *h* é uma medida bastante usada, normalmente definida por err*(h),* equação (9) a qual faz comparação da classe predita pelo classificador com a classe real (REZENDE, 2005).

$$err(h) = \frac{1}{n}\sum_{n=1}^{n} \left\| y_i \neq h(x_i) \right\| \qquad (9)$$

Onde:

a) n é o número de partições do conjunto de dados;

b) y_i é a classe real;

c) $h(x_i)$ é o retorno do classificador;

d) $\left\| y_i \neq h(x_i) \right\|$ retorna 1 se verdadeira e 0 se falsa.

O erro médio é obtido pela equação (10).

$$mean(A) = \frac{1}{n}\sum_{i=1}^{n} err(h) \qquad (10)$$

Onde:

a) n é o número de partições do conjunto de dados;

b) $err(h)$ é o resultado da equação (9).

O resultado do erro médio é então utilizado para calcular o desvio padrão do classificador, dado pelas equação (11) e (12).

$$\text{var}(A) = \frac{1}{n}\left[\frac{1}{n-1}\sum_{i=1}^{n} (err(h_i) - mean(A))^2 \right] \qquad (11)$$

Onde:

a) var(A) é a variância das partições;

b) n é o número de partições do conjunto de dados;

c) $err(h_i)$ é o erro obtido em cada partição;

d) mean(A) é o erro médio.

$$sd(A) = \sqrt{\text{var}(A)} \qquad (12)$$

$sd(A)$ é denominado o desvio padrão do classificador.

Uma medida importante para a avaliação do desempenho é a validação do conhecimento que consiste na compreensão e validação por parte do especialista da área da base de dados que foi submetida ao *data mining*. Isso contribui para que o sistema seja avaliado quanto a compreensibilidade do conhecimento gerado.

3.1.5 Validação do Conhecimento Adquirido

Um classificador pode gerar muitas regras, todas corretas, mas é necessário levar em consideração o interesse do usuário nessas regras, muitas regras podem não contribuir com conhecimento útil na finalidade esperada pelo usuário. Por isso é necessário que as regras geradas por algoritmos de *data mining* além de corretas e compreensíveis sejam de interesse estratégico na resolução de problemas (ROMÃO, 2002).

Um fator a ser considerado é a compreensibilidade das regras geradas pelo classificador. Nesse ponto a entrevista com o especialista da área se torna de suma importância, pois ele pode avaliar as regras quanto ao interesse e ao conhecimento adquirido a partir de sistemas de descoberta de conhecimento em bases de dados.

4 TRABALHOS CORRELATOS

A descoberta de conhecimento em base de dados vem sendo bastante estudada ao longo dos anos por possibilitar a identificação de relações e padrões em bases de dados. As tarefas e métodos de *data mining* têm possibilitado isso, porém o conhecimento gerado pelos algoritmos precisam ser confiáveis e atender as características desejáveis. Assim, a avaliação de desempenho torna-se importante na verificação das relações e dos padrões que vêm sendo descobertos, quando aplicam-se bases de dados específicas.

4.1 DESCOBERTA DE CONHECIMENTO RELEVANTE EM BANCO DE DADOS SOBRE CIÊNCIA E TECNOLOGIA

Tese apresentada ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal de Santa Catarina, como parte dos requisitos para a obtenção do título de Doutor em Engenharia de Produção, em fevereiro de 2002 por Wesley Romão.

Nesta tese o objetivo é resolver uma generalização da tarefa de classificação, conhecida como modelagem de dependência, para extrair conhecimento na forma de regras de previsão válidas, utilizando para isso métricas de avaliação de desempenho (ROMÃO, 2002).

Propõe-se um novo algoritmo genético capaz de descobrir regras de previsão difusas. O objetivo é descobrir conhecimento relevante, surpreendente e novo, oculto em banco de dados adaptando-se uma técnica que emprega impressões gerais fornecidas pelo usuário, um assunto pouco explorado. Utilizando como referencial de eficiência e qualidade do protótipo proposto os resultados obtidos na avaliação do algoritmo J4.8 (ROMÃO, 2002).

No desenvolvimento do trabalho utilizou-se uma base de dados com registros da agência de fomento a Ciência e Tecnologia dos estados do Sul do Brasil. Para avaliar e validar as descobertas dos algoritmos utilizados, foram empregadas técnicas objetivas como a

validação cruzada, fator de confiança e a matriz confusão, e técnicas subjetivas que correspondem a aplicação de entrevista com o especialista da área (ROMÃO, 2002).

4.2 EXTRAÇÃO AUTOMÁTICA DE CONHECIMENTO POR MÚLTIPLOS INDUTORES

Tese apresentada ao Instituto de Ciência e Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências da Computação e Matemática Computacional, por José Augusto Baranauskas em 2001.

Os objetivos desta tese são: a avaliação da seleção de atributos e o seu impacto em um classificador simbólico; a avaliação da composição de atributos guiada pelo conhecimento fornecido por um especialista; uma metodologia para prover a combinação de classificadores simbólicos em um classificador final também simbólico; e foi proposto e implementado o sistema xruler (BARANAUSKAS, 2001).

A tese descreve conceitos e definições sobre aprendizado supervisionado e as linguagens de representação freqüentemente utilizadas, encontra-se também a descrição de técnicas e métodos para validar e avaliar o desempenho de algoritmos de data mining utilizados no desenvolvimento do trabalho (BARANAUSKAS, 2001).

Para a validação do algoritmo descrito em seu trabalho Baranauskas propõe métodos baseados em amostragem como o *Hold out*, amostragem aleatória, *Cross-Validation*, *Stratified Cross-Validation*, *Leave-one-out* e *Booststrap* (BARANAUSKAS, 2001).

Na validação do desempenho apresenta o cálculo da média e do desvio padrão comumente utilizadas em aprendizado de máquina, com a utilização de amostragem, e a avaliação do conhecimento adquirido com o especialista (BARANAUSKAS, 2001).

As técnicas de avaliação de desempenho foram empregadas com a finalidade de comparar os algoritmos C4.5 e CN2 implementados no sistema xruler, chegando-se ao

resultado de o sistema proporcionou uma redução da taxa de erro de 0,5% para o C4.5 e de quase 10% para o CN2 (BARANAUSKAS, 2001).

4.3 AVALIAÇÃO DO CONHECIMENTO DESCOBERTO A PARTIR DE ALGORITMOS DE *DATA MINING*

O artigo de iniciação científica foi desenvolvido na Universidade Tuiuti do Paraná em parceria com o Instituto Paranaense de Desenvolvimento Econômico e Social, no ano de 2003, pelos acadêmicos Deborah R. Carvalho, Wilson Alves Neto e Marcos Bueno.

Os autores desenvolvem uma revisão de literatura em relação a avaliação de desempenho em algoritmos de *data mining*, e realizam um estudo de caso aplicando as métricas propostas na avaliação (CARVALHO; BUENO; NETO, 2003).

No processo de avaliação do conhecimento descoberto em bases de dados por meio dos algoritmos de *data mining*, descreve algumas técnicas que podem ser aplicadas no pós-processamento, abordando duas categorias básicas, os métodos subjetivos e objetivos(CARVALHO; BUENO; NETO, 2003).

O método subjetivo busca satisfazer o usuário, isso implica que ele apresente um conhecimento prévio do domínio de aplicação. Durante a realização desta avaliação o especialista da área precisa ser consultado, como por exemplo, em forma de entrevista, com o objetivo de determinar o interesse das regras descobertas (CARVALHO; BUENO; NETO, 2003).

Já no método objetivo os autores utilizam métricas como a taxa de acerto, o suporte, sensibilidade e especificidade, e a medida de interesse na regra. Na realização dos testes foi utilizado o algoritmo C4.5 para indução de árvores de decisão, que demonstrou uma taxa de acerto igual a 94,25% (CARVALHO; BUENO; NETO, 2003).

4.4 DESCOBERTA DE CONHECIMENTO E MEDIDAS DE INTERESSE: UM EXAME

O relatório técnico foi desenvolvido por Robert J. Hamilton, e Howard J. Hilderman do Departamento de Ciência da Computação da Universidade de Regina, Canadá em 1999.

O trabalho contém um estudo da descoberta de desconhecimento em bases de dados. Uma área de *data mining* que necessita ser bastante estudada e que causa muitos desafios é mensurar os padrões encontrados com relação a usabilidade e validade das regras encontradas (HAMILTON, HILDERMAN, 1999, tradução nossa).

Algumas técnicas para classificar as medidas efetivas dos padrões encontrados nessas bases de dados são abordadas no trabalho, que propõem uma classificação dessas regras quanto a significância e utilidade (HAMILTON; HILDERMAN, 1999, tradução nossa).

Segundo os autores existem diferentes métricas para avaliação de desempenho em aplicações de *data mining*, dependendo da técnica utilizada. A classificação pode ser avaliada em relação ao (HAMILTON; HILDERMAN, 1999, tradução nossa):

a) interesse da função da regra;
b) a média relativa das regras usadas para determinar as melhores regras;
c) a seleção de regras potencialmente interessantes de acordo com a simplicidade e significância estatística;
d) o interesse nas regras geradas na classificação de acordo com a necessidade e suficiência da regra;
e) as impressões gerais utilizadas para avaliar a importância das regras geradas;
f) a distância das medidas das regras utilizadas para determinar as regras com maior cobertura de dados.

Ainda no trabalho os autores descrevem métricas para avaliação e validação na tarefa de associação. São elas: *Agrawal and Srikant´s Itemset Measures; Klemettien Rule Templates; Gray and Orlowska´s Interestingness; Dong and Li´s Interestingness; Liu Reliable Exceptions;* e *Zhong Peculiarity*. (HAMILTON, HILDERMAN, 1999, tradução nossa).

5 AVALIAÇÃO DE DESEMPENHO DO MÓDULO DE CLASSIFICAÇÃO DA ORION *DATA MINING ENGINE*

A *Shell* Orion é um projeto acadêmico com o objetivo de desenvolver uma ferramenta gratuita para a descoberta de conhecimento em bases de dados e fornecer à comunidade acadêmica informações sobre *data Mining*, demonstração matemática e implementação de alguns algoritmos.

O módulo de classificação da *Shell* Orion conta com os algoritmos CART, C4.5 e ID3. A avaliação dos algoritmos ID3 e C4.5 pertencentes a este módulo é mais uma contribuição para o desenvolvimento do projeto, trazendo além das métricas empregadas na avaliação do desempenho, a entrevista com o especialista da área da base de dados utilizada, na validação do conhecimento, para isso foi aplicado um questionário contendo questões relacionadas a funcionalidade da aplicação, usabilidade e as regras geradas.

Na avaliação do conhecimento e do desempenho, foi necessário utilizar uma base de dados, nesta pesquisa optou-se por uma base da área médica, com informações de pacientes portadores de sepse.

5.1 BASE DE DADOS

A base de dados utilizada nesta pesquisa foi disponibilizada pelo Dr. Felipe Dal Pizzol, do curso de Medicina da UNESC, contendo registros de pacientes com sepse da Unidade de Terapia Intensiva (UTI) do Hospital das Clínicas de Porto Alegre, Rio Grande do Sul.

A sepse é uma inflamação causada por germes patogênicos, podendo atacar somente um órgão do doente, porém provoca uma resposta infecciosa que acaba atingindo

todo o organismo humano, por isso a sepse é comumente chamada de infecção generalizada (CECIL; GOLDMAN; AUSIELLO, 2005).

A base de dados contém no total 96 registros, cada registro possui 45 atributos com informações dos pacientes e de exames realizados durante a internação.

A base foi pré-processada e normalizada, as alterações necessárias para a classificações dos algoritmos foi realizada nessa etapa, como a transformação do atributo de saída em ultimo atributo da tabela no caso do C4.5. A Tabela 5 descreve os atributos da base de dados.

O objetivo da classificação com os dados da base da sepse é saber o desfecho do paciente que pode ser óbito ou cura, a base conta com 45 registros com o desfecho cura e 51 registros com o desfecho óbito, os algoritmos de classificação calculam o atributo com o maior ganho para eleger como atributo classificador.

Tabela 5. Descrição Base de Dados

Atributo	Descrição	Valor
caso	Número do caso registrado pelo hospital.	Número inteiro
idade	Idade do paciente.	Número inteiro
sexo	Sexo do paciente.	1 para masculino e 2 para feminino
gravidad	Gravidade em que o paciente chegou ao hospital.	1 para leve, 2 para grave e 3 para estado de choque
ira	Se foi registrado insuficiência renal aguda (perda das funções dos rins) durante a internação.	1 para sim e 2 para não
hepatico	Se o paciente possuía hepatite (inflamação no fígado).	1 para sim e 2 para não
vm	Se o paciente utilizou ventilação mecânica (aparelho que auxilia a respiração) durante a internação.	1 para sim e 2 para não
fio2, fio21, fio22 e fio23	Se utilizou ventilação mecânica, qual a fração inspirada de oxigênio pelo paciente (em porcentagem). Os quatro atributos representam respectivamente os valores dos quatro primeiros dias de internação.	Número decimal
neurolog	Se o caso é neurológico, ou seja, se afetou o sistema nervoso central	1 para sim e 2 para não
vasopres	Se o paciente utilizou vasopressor (substância com o objetivo de aumentar a pressão arterial)	1 para sim e 2 para não
nora, nora1, nora2 e nora3	Se utilizou vasopressor, qual a quantidade de noradrenalina utilizada (substância que mantém a pressão arterial estabilizada). Os quatro atributos representam respectivamente os valores dos quatro primeiros dias de internação.	Número decimal
apacheII	Grau de gravidade em que o paciente chegou na UTI. Este valor determina a previsão de sobrevivência do doente.	Número inteiro
mods, mods2, mods3 e mods4	Grau de disfunção de múltiplos órgãos do paciente. Os quatro atributos representam respectivamente os valores dos quatro primeiros dias de internação.	Número inteiro
tempodesfecho	Tempo de desfecho do paciente (em dias).	Número inteiro
desfecho	Desfecho do paciente.	Número inteiro
tba1, tba2, tba3 e tba4	Marcador de oxidação de lipídios (alterações dos lipídios pelo oxigênio).	Número decimal
carbonyl, carb2, carb3 e carb4	Marcador de oxidação de proteínas (alterações das proteínas pelo oxigênio).	Número decimal
sodcategorica	Categoria do superóxido-dismutase (enzima que realiza a dismutação do radical superóxido em peróxido de hidrogênio e oxigênio).	1 para sod maior que 5.5 e 2 para sod menor que 5.5
sod, sod2, sod3 e sod4	Marcador do superóxido-dismutase. Os quatro atributos representam respectivamente os valores dos quatro primeiros dias de internação.	Número decimal
cat1, cat2, cat3 e cat4	Quantidade de catalase (enzima que transforma o peróxido de hidrogênio em água e oxigênio) produzida pelo paciente. Os quatro atributos representam respectivamente os valores dos quatro primeiros dias de internação.	Número decimal
xo1, xo2, xo3 e xo4	Quantidade de xantina oxidase (enzima que catalisa hipoxantina com oxigênio, produzindo ácido úrico e o radical superóxido) produzida pelo paciente. Os quatro atributos representam respectivamente os valores dos quatro primeiros dias de internação.	Número decimal

Fonte: Adaptado de CASSETTARI JUNIOR, J. M. (2008)

No desenvolvimento dessa pesquisa adotou-se uma metodologia com o objetivo de compreender os temas abordados, executar os algoritmos de classificação da *Shell* Orion e avaliar o seu desempenho.

5.2 METODOLOGIA

As seguintes etapas foram adotadas para a validação do conhecimento: levantamento bibliográfico dos temas abordados; definição da base de dados a ser utilizada; normalização da base de acordo com as necessidades dos algoritmos; execução dos algoritmos de classificação a partir da base; realização dos cálculos de avaliação; entrevista com o especialista.

O levantamento bibliográfico abordou os temas inerentes a este trabalho, tais como: descoberta de conhecimento em base de dados, *data mining*, as tarefas e métodos de *data mining*, a *Shell* Orion *Data Mining Engine*, estudo dos algoritmos da tarefa de classificação da *Shell*, avaliação do desempenho e métricas.

5.2.1 Avaliação de Desempenho do Algoritmo C4.5 na Tarefa de Classificação da *Shell* Orion *Data Mining Engine*

O algoritmo C4.5 utiliza o método de indução de árvores de decisão para realizar a classificação. Este algoritmo é uma evolução do ID3, sua principal diferença é possuir a capacidade de simplificação da árvore gerada, ou seja, efetua o cálculo da poda, excluindo as regras sem valores significativos para a classificação.

Como etapa de pré-processamento foi efetuado a adequação da base de dados, para a classificação utilizando o algoritmo C4.5 é necessário que o atributo de saída seja o último atributo da base, que nesse caso é o *desfecho*.

Após a etapa de pré-processamento, foi possível realizar a classificação e com base nos resultados obtidos calcular a matriz confusão, o fator de confiança, realizar a validação cruzada e o cálculo do erro médio.

5.2.1.1 Cálculo da Matriz Confusão do Algoritmo C4.5

Na realização da avaliação foram utilizados todos os registros da base de dados. O algoritmo C4.5 elegeu como atributo de melhor ganho o *tempodesfecho* que contém a quantidade de dias que o paciente permaneceu internado, conforme ilustra a Figura 17:

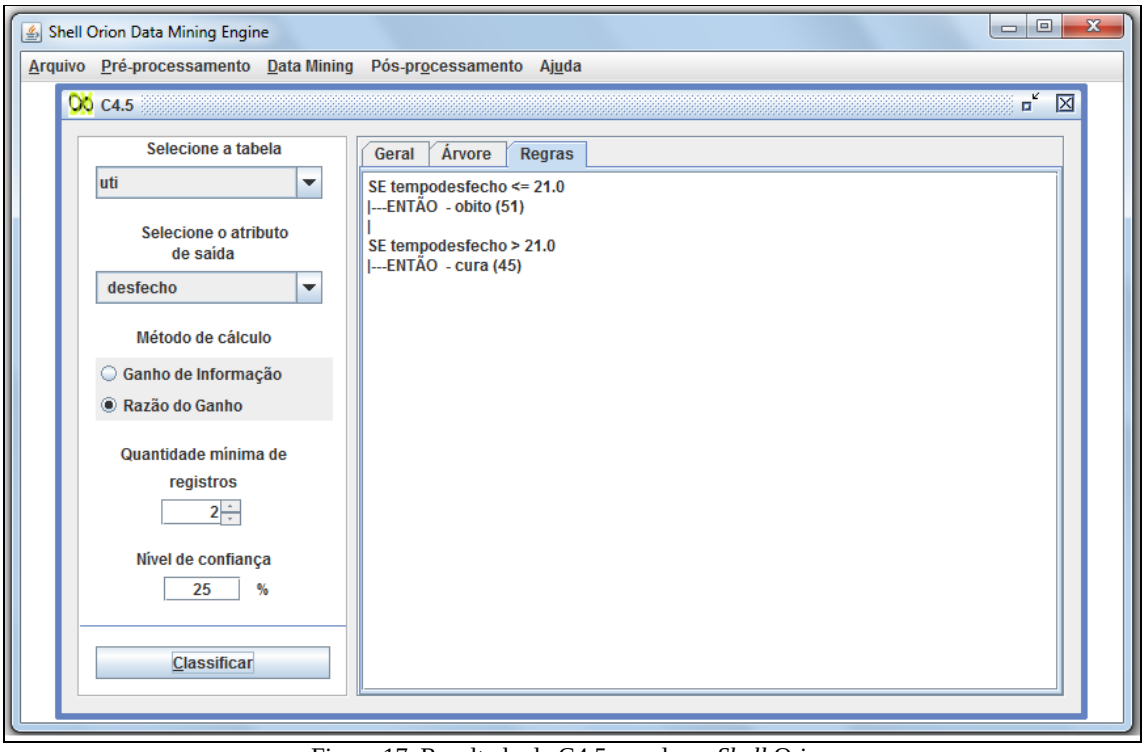


Figura 17. Resultado do C4.5 gerado na *Shell* Orion

A matriz confusão é uma matriz de ordem dois por dois que contém as classificações corretas *versus* o número de classificações preditas para cada classe. A partir dos resultados do algoritmo de *data mining* C4.5 pode-se gerar a matriz confusão ilustrada na Figura 18:

		Classe prevista	
		Óbito	Cura
Classe real	Óbito	51	0
	Cura	0	45

Figura 18. Matriz Confusão do Algoritmo C4.5

Possibilitando calcular: acurácia, erro, sensibilidade e abrangência, especificidade e precisão, do algoritmo C4.5.

Acurácia: $\frac{VP+VN}{n}.100\% = \frac{51+45}{96}.100\% = 100\%$

Erro: $\frac{FP+FN}{n}.100\% = \frac{0+0}{96}.100\% = 0\%$

Cálculo da taxa dos verdadeiros positivos (sensibilidade e abrangência):

$\frac{VP}{VP+FP}.100\% = \frac{51}{51+0}.100\% = 100\%$

Cálculo da taxa de verdadeiros negativos (especificidade):

$\frac{VN}{VN+FP}.100\% = \frac{45}{45+0}.100\% = 100\%$

Precisão: $\frac{VP}{VP+FP}.100\% = \frac{51}{51+0}.100\% = 100\%$

Onde:

a) VP: classificações verdadeiras positivo;

b) VN: classificações verdadeiras negativo;

c) FP: classificações falsas positivo;

d) FN: classificações falsas negativo.

É de fácil visualização na matriz confusão, a ausência de classificações erradas, o que comprova a qualidade do conhecimento descoberto, que pode ser comparado ao de um classificador ideal, uma vez que não comete erros.

As taxas de sensibilidade e especificidade demonstram que o classificador obteve os resultados esperados, a alta taxa de sensibilidade garante que não se tenha perda de casos candidatos no auxilio a tomada de decisão, já a alta taxa de especificidade garante que não sejam levados em consideração casos com resultados negativos.

5.2.1.2 Cálculo da Taxa de Acerto

Considerando as regras geradas na Figura 16 com a classificação pelo algoritmo C4.5 na *Shell* Orion, realizou-se o cálculo da taxa de acerto para cada saída da regra:

$$\text{Taxa de acerto} = \frac{\text{Número de classificações corretas}}{\text{Número de dados aplicáveis}}$$

a) regra com desfecho óbito.

$$\text{Taxa de acerto: } \frac{51}{51} = 1 = 100\%$$

b) regra com desfecho cura.

$$\text{Taxa de acerto: } \frac{45}{45} = 1 = 100\%$$

A taxa de acerto é igual a 100% para as duas regras, uma vez que o algoritmo não comete erros.

5.2.1.3 Cálculo do Fator de Confiança

Realização do cálculo do fator de confiança para cada regra gerada a partir da classificação no algoritmo C4.5.

Fator de confiança = (X – 1/2)/Y

a) Fator de confiança para a regra: SE tempodesfecho <= 21.0 ENTÃO óbito, contendo 51 casos.

$$\left(51-\frac{1}{2}\right)/51 = 99{,}01\%$$

b) Fator de confiança para a regra: SE tempodesfecho > 21.0 ENTÃO cura, contendo 45 casos.

$$\left(45-\frac{1}{2}\right)/45 = 98{,}88\%$$

O fator (- ½), penaliza as regras com menor abrangência de dados, não penalizando significativamente regras com abrangência maior. Este fator ajuda a identificar regras que se ajustam demasiadamente a peculiaridade da base de dados.

5.2.1.4 Validação Cruzada para o Algoritmo C4.5

No cálculo da validação cruzada dividiu-se a base de dados em cinco subconjuntos, sendo que destes subconjuntos, quatro eram compostos por dezenove registros e um subconjunto possuía os vinte dados restantes, que totalizaram os 96 registros. Cada subconjunto compondo dez casos de desfecho óbito e nove casos de desfecho cura, com exceção do ultimo com onze casos com desfecho óbito e nove com desfecho cura, o Apêndice A contém os subconjuntos gerados. Cada subconjunto foi submetido ao algoritmo C4.5 e calculado para cada iteração a taxa de acerto das regra gerada.

$$\text{Taxa de acerto} = \frac{\text{Número de classificações corretas}}{\text{Número de dados aplicáveis}}$$

Iteração 1:

a) SE tempodesfecho <= 15.0 ENTÃO óbito, contendo 10 casos:

Taxa de acerto: $\frac{10}{10} = 1 = 100\%$

b) SE tempodesfecho > 15.0 ENTÃO cura, contendo 9 casos:

Taxa de acerto: $\frac{9}{9} = 1 = 100\%$

O resultado do cálculo para a taxa de acerto de cada iteração foi igual a 100%, no Apêndice B encontram-se os cálculos para cada iteração.

Com os resultados das taxas de acerto é possível calcular a média (ā) e do desvio padrão (dp) para cada regra. Regras com desfecho óbito:

$$\bar{a} = \frac{1}{k}\sum_{i=1}^{k} x_i \; = \; \frac{1}{5}.5 = 1$$

$$dp(\bar{a}) = \sqrt{\frac{1}{k}\left[\frac{1}{k-1}\sum_{i=1}^{k}(x_i - \bar{a})^2\right]} = \sqrt{\frac{1}{5}.\left[\frac{1}{4}.5\right]} = 0{,}5$$

Onde:

a) ā = média;

b) dp = desvio padrão;

c) x_i = taxa de acerto da partição i;

d) K = número de partições.

Cálculo da média (ā) e do desvio padrão (dp) para as regras com desfecho cura:

$$\bar{a} = \frac{1}{k}\sum_{i=1}^{k} x_i \; = \; \frac{1}{5}.5 = 1$$

$$dp(\bar{a}) = \sqrt{\frac{1}{k}\left[\frac{1}{k-1}\sum_{i=1}^{k}(x_i - \bar{a})^2\right]} = \sqrt{\frac{1}{5}.\left[\frac{1}{4}.5\right]} = 0{,}5$$

O cálculo da média e do desvio padrão na validação cruzada permitem estimar o acerto dos dados gerados a partir da tarefa de classificação, muito utilizado para a validação e avaliação do conhecimento descoberto.

5.2.1.5 Cálculo do Erro Médio do Algoritmo C4.5

Para calcular o erro médio do algoritmo C4.5 utilizou-se a mesma base particionada, Apêndice A, do cálculo da validação cruzada.

Cálculo do erro por partição, a partir da expressão abaixo:

$$err(h) = \frac{1}{n}\sum_{n=1}^{n} \left\| y_i \neq h(x_i) \right\|$$

Onde:

a) n é o número de partições do conjunto de dados;

b) y_i é a classe real;

c) $h(x_i)$ é o retorno do classificador;

d) $\left\| y_i \neq h(x_i) \right\|$ retorna 1 se verdadeira e 0 se falsa.

Partição 1:

$$err(h) = \frac{1}{19}.(0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0) = 0$$

Partição 2:

$$err(h) = \frac{1}{19}.(0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0) = 0$$

Partição 3:

$$err(h) = \frac{1}{19}.(0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0) = 0$$

Partição 4:

$$err(h) = \frac{1}{19}.(0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0) = 0$$

Partição 5:

$$err(h) = \frac{1}{20}.(0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0) = 0$$

Cálculo da média dos erros das partições:

$$mean(A) = \frac{1}{n}\sum_{i=1}^{n} err(h)$$

Onde:

a) n é o número de partições do conjunto de dados;

b) $err(h)$ é o resultado do cálculo do erro.

$$mean(A) = \frac{1}{5}.(0+0+0+0+0) = 0$$

Cálculo do desvio padrão e da variância do classificador:

$$\text{var}(A) = \frac{1}{n}\left[\frac{1}{n-1}\sum_{i=1}^{n}(err(h_i) - mean(A))^2\right]$$

Onde:

a) var(A) é a variância das partições;

b) n é o número de partições do conjunto de dados;

c) $err(h_i)$ é o erro obtido em cada partição;

d) mean(A) é o erro médio.

$$\text{var}(A) = \frac{1}{5}\left[\frac{1}{5}.(0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2\right] = 0$$

Cálculo do desvio padrão do classificador:

$$dp(A) = \sqrt{\text{var}(A)}$$

$$dp(A) = \sqrt{0} = 0$$

O erro médio do classificado é 0, ou seja o classificador não apresenta erros. Os valores obtidos com os cálculos do erro médio no algoritmo C4.5 demonstram que a ferramenta é confiável.

5.2.2 Avaliação de Desempenho do Algoritmo ID3 na Tarefa de Classificação da *Shell* Orion *Data Mining Engine*

O algoritmo ID3 foi executado na *Shell* Orion com a base da sepse e profundidade da árvore igual a 1, gerando 16 regras, mudando a profundidade da árvore o algoritmo gerou as mesmas regras.

Na etapa de pré-processamento foi necessário excluir o atributo *caso* que continha o número do caso registrado pelo hospital, um número inteiro de 1 a 96, este tratamento foi realizado para o algoritmo não eleger o atributo *caso* como atributo classificador, gerando assim uma regra para cada caso. Após o pré-processamento o algoritmo elegeu o campo *tempodesfecho* como atributo classificador, conforme ilustra a Figura 19.

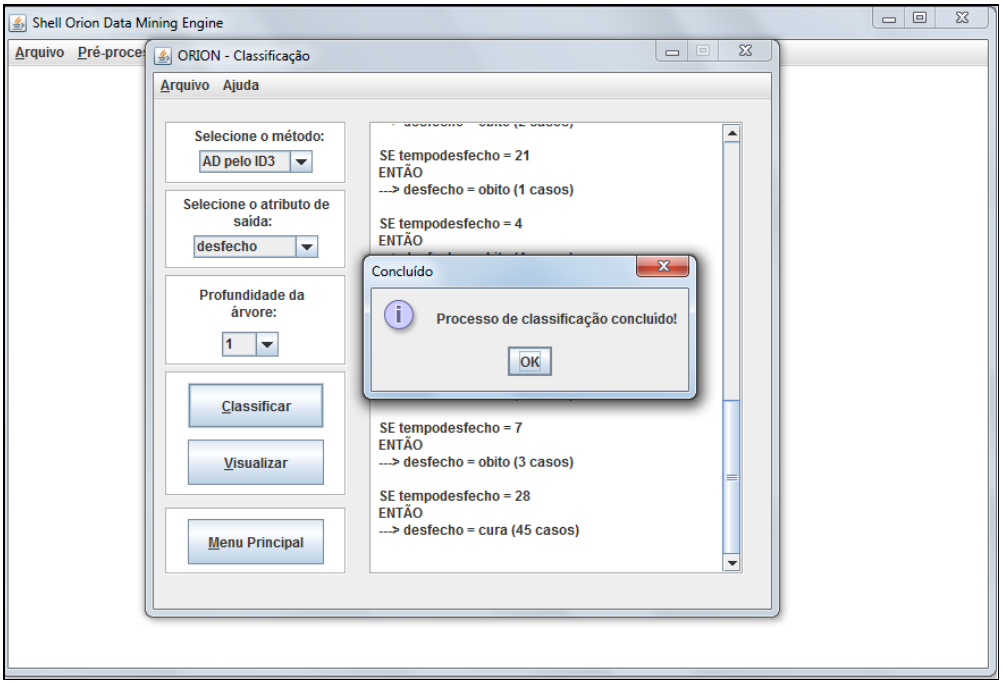


Figura 19. Classificação pelo algoritmo ID3 na Shell Orion

A partir das regras obtidas com a classificação por meio do ID3, foi possível calcular a matriz confusão, o fator de confiança, realizar a validação cruzada e o cálculo do erro médio.

5.2.2.1 Cálculo da Matriz Confusão do Algoritmo ID3

Na avaliação do desempenho do algoritmo ID3 a partir da matriz confusão, foram utilizados todos os registros da base de dados. O algoritmo ID3 elegeu como atributo de melhor ganho o *tempodesfecho* que contém a quantidade de dias que o paciente permaneceu internado, gerando assim 16 regras: 15 regras para saída óbito, totalizando 51 casos; e 1 regra para saída cura, totalizando 45 casos, Figura 20:

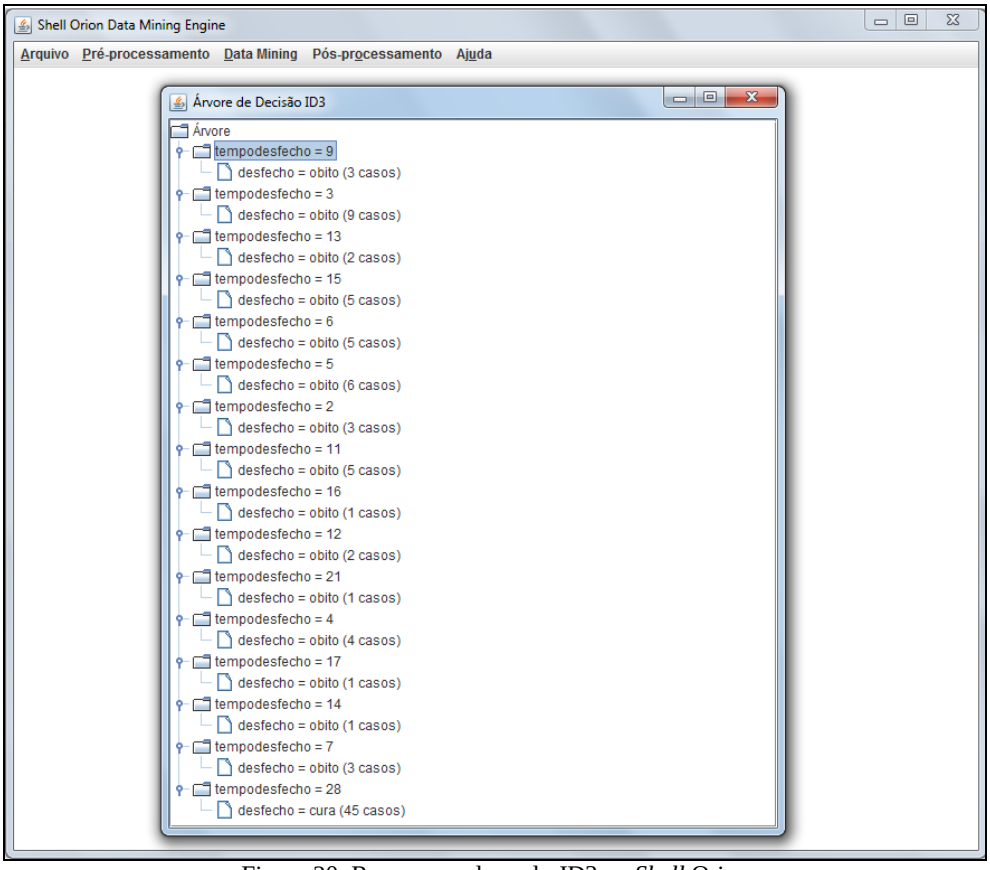


Figura 20. Regras geradas pelo ID3 na *Shell* Orion

A partir dos resultados gerados pelo algoritmo ID3 na tarefa de classificação da *Shell* Orion, foi possível gerar a matriz confusão para o algoritmo, Figura 21. Uma matriz confusão contém as classificações corretas *versus* as classificações preditas e possibilita avaliar o desempenho de um classificador.

		Classe Prevista	
		Óbito	Cura
Classe Real	Óbito	51	0
	Cura	0	45

Figura 21. Matriz Confusão do algoritmo ID3

Acurácia: $\frac{VP+VN}{n}.100\% = \frac{51+45}{96}.100\% = 100\%$

Erro: $\frac{FP+FN}{n}.100\% = \frac{0+0}{96}.100\% = 0\%$

Cálculo da taxa dos verdadeiros positivos (sensibilidade e abrangência):

$\frac{VP}{VP+FP}.100\% = \frac{51}{51+0}.100\% = 100\%$

Cálculo da taxa de verdadeiros negativos (especificidade):

$\frac{VN}{VN+FP}.100\% = \frac{45}{45+0}.100\% = 100\%$

Precisão: $\frac{VP}{VP+FP}.100\% = \frac{51}{51+0}.100\% = 100\%$

Onde:

a) VP: classificações verdadeiras positivo;

b) VN: classificações verdadeiras negativo;

c) FP: classificações falsas positivo;

d) FN: classificações falsas negativo.

É de fácil visualização na matriz confusão, a ausência de classificações erradas, o que comprova a qualidade do conhecimento descoberto pelo algoritmo ID3, que pode ser comparado ao de um classificador ideal, uma vez que não comete erros.

As taxas de sensibilidade e especificidade obtiveram um percentual de 100% o que demonstra que o classificador obteve os resultados esperados. A sensibilidade garante que não se tenha perda de casos candidatos no auxilio a tomada de decisão, já a especificidade garante que não sejam levados em consideração casos com resultados negativos.

5.2.2.2 Cálculo da Taxa de Acerto

A partir dos dados contidos na Figura 20 foi realizado o cálculo da taxa de acerto para cada saída da regra:

$$\text{Taxa de acerto} = \frac{\text{Número de classificações corretas}}{\text{Número de dados aplicáveis}}$$

a) Regra com desfecho óbito.

Taxa de acerto: $\frac{51}{51} = 1 = 100\%$

b) Regra com desfecho cura.

Taxa de acerto: $\frac{45}{45} = 1 = 100\%$

5.2.2.3 Cálculo do Fator de Confiança

Realização do cálculo do fator de confiança para cada regra, a partir das regras geradas com a classificação, conforme Figura 20.

Fator de confiança = (X – 1/2)/Y

Fator de confiança para a regra: SE tempodesfecho = 9.0 ENTÃO óbito, contendo 3 casos.

$$\left(3-\frac{1}{2}\right)/3 = 83{,}33\%$$

O cálculo do fator de confiança das demais regras geradas pelo ID3 encontram-se no Apêndice C. Nas regras geradas pelo algoritmo ID3 tem-se o *overfitting*, que acontece quando a regra se ajusta demais as peculiaridades dos dados. Isso ocorre principalmente por esse algoritmo não realizar o cálculo da poda, o fator de confiança ajuda a identificar as regras com maior abrangência dos dados da base.

5.2.2.4 Validação Cruzada para o Algoritmo ID3

No cálculo da validação cruzada a base de dados foi divida em cinco subconjuntos, quatro subconjuntos contendo 19 registros e um subconjunto contento 20 registros, totalizando os 96 registros. Cada subconjunto contém dez casos de desfecho óbito e nove casos de desfecho cura, com exceção do ultimo com onze casos com desfecho óbito e nove com desfecho cura. Cada subconjunto foi submetido ao algoritmo ID3 e calculado para cada iteração a taxa de acerto das regras geradas, o Apêndice A contém os subconjuntos gerados.

Taxa de acerto = Número de classificações corretas
Número de dados aplicáveis

Iteração 1:

a) SE idade = 75.0 ENTÃO óbito, contendo 2 casos

SE idade = 57.0 ENTÃO óbito, contendo 1 casos

SE idade = 71.0 ENTÃO óbito, contendo 2 casos

SE idade = 41.0 ENTÃO óbito, contendo 2 casos

SE idade = 69.0 ENTÃO óbito, contendo 1 casos

SE idade = 50.0 ENTÃO óbito, contendo 1 casos

SE idade = 58.0 ENTÃO óbito, contendo 1 casos:

Taxa de acerto: $\frac{10}{10} = 1 = 100\%$

b) SE idade = 48.0 ENTÃO cura, contendo 1 casos

SE idade = 30.0 ENTÃO cura, contendo 1 casos

SE idade = 31.0 ENTÃO cura, contendo 1 casos

SE idade = 78.0 ENTÃO cura, contendo 1 casos

SE idade = 84.0 ENTÃO cura, contendo 1 casos

SE idade = 59.0 ENTÃO cura, contendo 1 casos

SE idade = 33.0 ENTÃO cura, contendo 1 casos

SE idade = 52.0 ENTÃO cura, contendo 1 casos

SE idade = 65.0 ENTÃO cura, contendo 1 casos:

Todos os cálculos para a taxa de acerto de cada iteração foram iguais a 100%, no Apêndice D encontram-se os cálculos de cada iteração. Com os resultados das taxas de acerto é possível calcular a média (ā) e do desvio padrão (dp) para as regras.

Cálculo da média (ā) e do desvio padrão (dp) para as regras com desfecho óbito:

$$\bar{a} = \frac{1}{k}\sum_{i=1}^{k} x_i = \frac{1}{5}.5 = 1$$

$$dp(\bar{a}) = \sqrt{\frac{1}{k}\left[\frac{1}{k-1}\sum_{i=1}^{k}(x_i - \bar{a})^2\right]} = \sqrt{\frac{1}{5}.\left[\frac{1}{4}.5\right]} = 0{,}5$$

Onde:

a) ā = média;

b) dp = desvio padrão;

c) x_i = taxa de acerto da partição i;

d) K = número de partições.

Cálculo da média (ā) e do desvio padrão (dp) para as regras com desfecho cura:

$$\bar{a} = \frac{1}{k}\sum_{i=1}^{k} x_i = \frac{1}{5}.5 = 1$$

$$dp(\bar{a}) = \sqrt{\frac{1}{k}\left[\frac{1}{k-1}\sum_{i=1}^{k}(x_i - \bar{a})^2\right]} = \sqrt{\frac{1}{5}.\left[\frac{1}{4}.5\right]} = 0{,}5$$

Os cálculos da média e do desvio padrão na validação cruzada permitem estimar o acerto dos dados gerados a partir da tarefa de classificação, sendo uma métrica muito utilizada para a validação e avaliação do conhecimento descoberto.

5.2.1.5 Cálculo do Erro Médio do Algoritmo ID3

No cálculo do erro médio do algoritmo ID3 foi utilizado a mesma base de dados particionada no cálculo da validação cruzada, que pode ser encontrada no Apêndice A.

Cálculo do erro por partição:

$$err(h) = \frac{1}{n}\sum_{n=1}^{n} \left\| y_i \neq h(x_i) \right\|$$

Onde:

a) n é o número de partições do conjunto de dados;

b) y_i é a classe real;

c) $h(x_i)$ é o retorno do classificador;

d) $\left\| y_i \neq h(x_i) \right\|$ retorna 1 se verdadeira e 0 se falsa.

Partição 1 =

$$err(h) = \frac{1}{19}.(0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0) = 0$$

Partição 2 =

$$err(h) = \frac{1}{19}.(0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0) = 0$$

Partição 3 =

$$err(h) = \frac{1}{19}.(0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0) = 0$$

Partição 4 =

$$err(h) = \frac{1}{19}.(0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0) = 0$$

Partição 5 =

$$err(h) = \frac{1}{20}.(0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0) = 0$$

A partir dos resultados do cálculo do erro é possível obter a média dos erros das partições:

$$mean(A) = \frac{1}{n}\sum_{i=1}^{n} err(h)$$

Onde:

a) n é o número de partições do conjunto de dados;

b) $err(h)$ é o resultado da equação anterior.

$$mean(A) = \frac{1}{5}.(0+0+0+0+0) = 0$$

Cálculo do desvio padrão e da variância do classificador:

$$\text{var}(A) = \frac{1}{n}\left[\frac{1}{n-1}\sum_{i=1}^{n}(err(h_i) - mean(A))^2\right]$$

Onde:

a) var(A) é a variância das partições;

b) n é o número de partições do conjunto de dados;

c) $err(h_i)$ é o erro obtido em cada partição;

d) mean(A) é o erro médio.

$$\text{var}(A) = \frac{1}{5}\left[\frac{1}{5}.(0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2\right] = 0$$

$$sd(A) = \sqrt{\text{var}(A)} =$$

$$dp(A) = \sqrt{0} = 0$$

O erro médio do classificado é 0, ou seja o classificador não apresenta erros. Os valores obtidos com os cálculos do erro médio no algoritmo ID3 demonstram que a ferramenta é confiável.

5.2.3 Método de Validação do Conhecimento dos Algoritmos C4.5 E ID3

A validação do conhecimento foi efetuada com base na entrevista com o especialista, aplicada na forma de questionário, o especialista da área que concedeu a entrevista foi a MSc. Cristiane Damiani Tomasi, enfermeira e mestre em Ciência da Saúde, bolsista de Doutorado orientada pelo Prof. Dr. Felipe Dal Pizzol, da Universidade do Extremo Sul Catarinense.

A metodologia aplicada na entrevista seguiu os seguintes passos:

a) apresentação da base de dados;

b) breve explicação sobre o funcionamento dos algoritmos C4.5 e ID3;

c) apresentação da *Shell* Orion *Data Mining Engine*;

d) aplicação dos algoritmos;

e) aplicação dos questionários.

Os questionários aplicados continham questões sobre a funcionalidade e usabilidade da aplicação, e apresentava as regras geradas pelo algoritmo que podiam ser classificadas conforme o interesse em: Alto; Médio, e baixo interesse. O questionário submetido para a validação do conhecimento a partir do C4.5 encontra-se no Apêndice E, e o aplicado para o ID3 no Apêndice F.

5.3 RESULTADOS OBTIDOS

Os resultados obtidos na avaliação de desempenho da tarefa de classificação da *Shell* Orion *Data Mining Engine*, foram a partir da análise do funcionamento do módulo e de seus algoritmos.

Inicialmente, verificou-se as capacidades dos algoritmos em relação a descoberta de grupos na base de dados de pacientes com sepse, com o objetivo de verificar e validar os resultados obtidos. A segunda etapa foi a realização dos cálculos para a validação do desempenho obtidos com os algoritmos C4.5 e ID3. Como terceira etapa foi desenvolvida e aplicada a entrevista com o especialista da área.

5.3.1 Validação e análise do Algoritmo C.45

A finalização da execução do algoritmo C4.5 na *Shell* Orion *Data Mining Engine* através da base de dados da sepse, utilizando o todos os registros para validação e para os testes, o algoritmo gerou apenas duas regras, uma contendo 51 casos de pacientes com o desfecho óbito e a segunda com 45 casos de pacientes com desfecho cura.

O algoritmo não apresentou nenhum erro, ou classificação incorreta, o que pode ser comprovado com os cálculos da matriz confusão e do erro médio. A partir dos dados da

matriz confusão também é possível comprovar que o algoritmo teve uma taxa de acurácia, sensibilidade, abrangência, especificidade e precisão igual a 100%, resultados que demonstram que o algoritmo obteve sucesso e resultados satisfatórios.

A taxa de confiança para as regras foi de 99,01% para o desfecho óbito e 98,88% para o desfecho cura, o que comprova que os resultados do algoritmo são confiáveis e que as regras geradas não se especializaram com detalhes da base, evitando assim o *orverfitting*.

Os resultados da validação cruzada e do cálculo do erro médio também foram muito satisfatórios, já que o algoritmo não apresentou erros. Assim, pode-se afirmar que a tarefa de classificação pelo algoritmo C4.5 da *Shell* Orion *Data Mining Engine* é uma ferramenta confiável na busca do conhecimento em bases de dados.

Conforme entrevista com o especialista que encontra-se no Apêndice E e F, o algoritmo C4.5 gera regras de alto interesse, e de fácil interpretação. A entrevistada também classificou o conhecimento gerado pela Shell Orion em ótimo quanto a eficiência na aquisição dos resultados, ótimo quanto ao conhecimento gerado corresponder as expectativas, ótimo também na compreensão do conhecimento gerado e quanto ao auxilio na confirmação de conhecimento pré-existente. Classificando como bons a contribuição com conhecimento novo e a contribuição de conhecimento útil a sua atividade.

Quanto a usabilidade a entrevistada classificou a *Shell* Orion como Bom em todos os requisitos: facilidade em aprender a manipular a *Shell*; facilidade de utilização e o ambiente da aplicação.

Um ponto muito importante na validação com o especialista da área é a isenção de respostas ruins ou de regras com baixo interesse, o que torna-se importante para contribuição de confiabilidade na aplicação.

5.3.2 Validação e Análise do Algoritmo ID3

Na finalização da execução do algoritmo ID3 na *Shell* Orion *Data Mining Engine* por meio da base de dados da sepse, utilizando todos os registros para validação e para os testes, o algoritmo gerou 16 regras, 15 regras contendo 51 casos de pacientes com o desfecho óbito e 1 regra com 45 casos de pacientes com desfecho cura.

O algoritmo não apresentou erros ou classificações erradas, a partir da matriz confusão gerada, pode-se afirmar que a acurácia, a sensibilidade, abrangência, especificidade e a precisão tiveram resultados iguais a 100%, já o erro foi igual a 0%. Esse resultados são excelentes e demonstram que a tarefa de classificação da *Shell* Orion *Data Mining Engine* obteve êxito no objetivo de descobrir conhecimento em bases de dados.

O cálculo do fator de confiança demonstra que os resultados do algoritmo são confiáveis. O algoritmo ID3 gerou mais regras se comparado ao C4.5, isso acontece porque o ID3 não possui o cálculo da poda da árvore. A taxa de confiança mostra quais regras são mais confiáveis, por possuírem mais casos iguais, o algoritmo ID3 apresentou *overfitting* em casos de regras com desfecho óbito, este fato não caracteriza um erro do algoritmo pois é uma peculiaridade do próprio ID3.

Os resultados da validação cruzada e do cálculo do erro médio comprovaram que o ID3 na *Shell* Orion é confiável, não apresentando taxas de erros. Desta forma, pode-se afirmar que a descoberta de conhecimento através da tarefa de classificação da *Shell* Orion *Data Mining Engine* apresenta resultados corretos e confiáveis.

A partir da entrevista com a especialista da área foi realizada a validação do conhecimento do algoritmo ID3. O questionário utilizado na entrevista pode ser encontrado na integra no Apêndice E. O ID3 gerou dezesseis regras, quinze com desfecho óbito e uma para desfecho cura, que poderiam ser classificadas pela especialista de acordo com o interesse em: alto; médio ou baixo interesse.

A especialista classificou cinco regras como de alto interesse, cinco como médio interesse e seis regras como de baixo interesse. As regras de alto interesse foram as que continham mais casos relacionados, assim como as de médio interesse foram aquelas com uma quantidade de casos que variou de três a cinco casos, e as regras que continham apenas um caso foram classificadas como de baixo interesse.

A usabilidade da ferramenta obteve resposta “ótima” para a fácil compreensão do conhecimento. E resposta “bom” quanto a eficiência na obtenção dos resultados, quanto ao conhecimento gerado corresponder as expectativas, ao acréscimo de conhecimento útil, confirmação de conhecimentos pré existentes e a contribuição de conhecimentos novos. Em contrapartida a usabilidade foi classificada como “ótimo” em todos os requisitos, facilidade em aprender a manipular a *Shell*, facilidade de utilização e o ambiente da aplicação. A validação quanto à usabilidade e a funcionalidade não obtiveram resposta para classificação como “ruim”.

5.3.3 Quadro Comparativo dos Resultados Obtidos nos Algoritmos C4.5 e ID3

A partir dos resultado obtidos no cálculo da matriz confusão, taxa de acerto, fator de confiança , validação cruzada e erro médio, a Figura 22 ilustra os resultados contidos nas tabelas comparativas, Tabela 6, Tabela 7, Tabela 8, e Tabela 9.

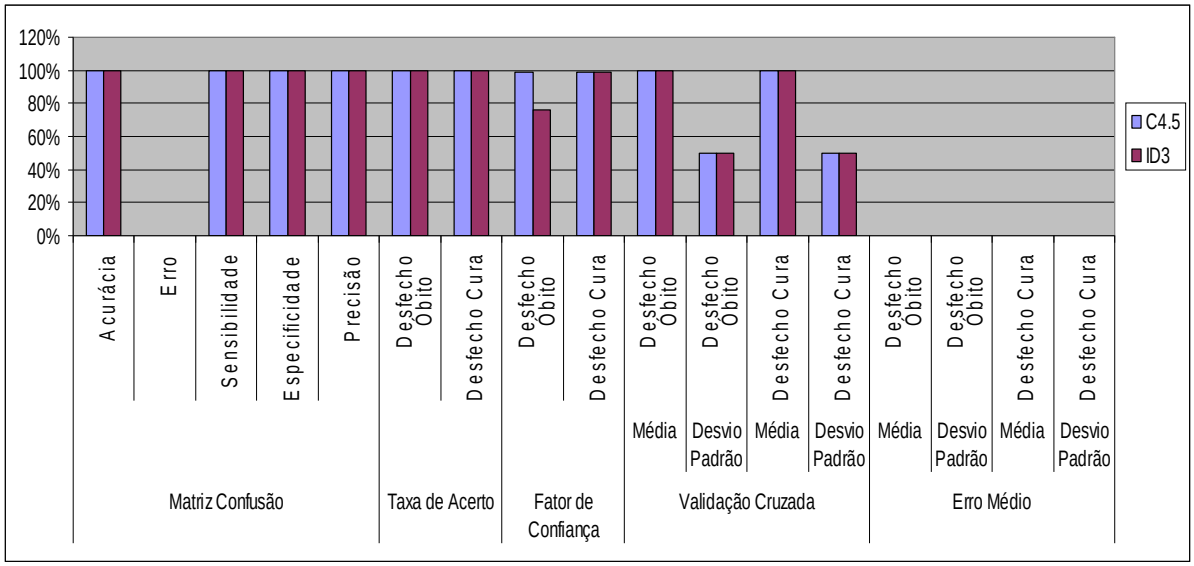


Figura 22. Gráfico comparativo da avaliação dos algoritmos C4.5 e ID3

Tabela 6. Tabela comparativa dos resultados do cálculo da matriz confusão

	Matriz Confusão				
	Acurácia	Erro	Sensibilidade	Especificidade	Precisão
C4.5	100%	0%	100%	100%	100%
ID3	100%	0%	100%	100%	100%

Tabela 7. Tabela comparativa dos resultados do cálculo da taxa de acerto e do fator de confiança

	Taxa de Acerto		Fator de Confiança	
	Desfecho Óbito	Desfecho Cura	Desfecho Óbito	Desfecho Cura
C4.5	100%	100%	99,01%	98,88%
ID3	100%	100%	76,24%	98,88%

Tabela 8. Tabela comparativa dos resultados do cálculo da validação cruzada

	Validação Cruzada			
	Média Desfecho Óbito	Desvio Padrão Desfecho Óbito	Média Desfecho Cura	Desvio Padrão Desfecho Cura
C4.5	1	0,5	1	0,5
ID3	1	0,5	1	0,5

Tabela 9. Tabela comparativa dos resultados do cálculo erro médio

	Erro Médio			
	Média Desfecho Óbito	Desvio Padrão Desfecho Óbito	Média Desfecho Cura	Desvio Padrão Desfecho Cura
C4.5	0	0	0	0
ID3	0	0	0	0

A partir da entrevista com a especialista, realizada por meio de questionário, foi possível representar através de uma tabela comparativa (Tabela 10) os resultados obtidos com as respostas do questionário, e ilustrados na Figura 23.

Tabela 10. Tabela comparativa dos resultados obtidos através da entrevista

	Interesse na Regra			Funcionalidade			Usabilidade		
	Alto	Médio	Baixo	Ótimo	Bom	Ruim	Ótimo	Bom	Ruim
C4.5	100%	-	-	66,66%	33,33%	-	-	100%	-
ID3	31,25%	31,25%	37,50%	16,67%	83,33%	-	100%	-	-

Tabela 11. Tabela comparativa da contagem dos resultados obtidos através da entrevista

	Interesse na Regra			Funcionalidade			Usabilidade		
	Alto	Médio	Baixo	Ótimo	Bom	Ruim	Ótimo	Bom	Ruim
C4.5	2	-	-	4	2	-	-	3	-
ID3	5	5	6	1	5	-	3	-	-

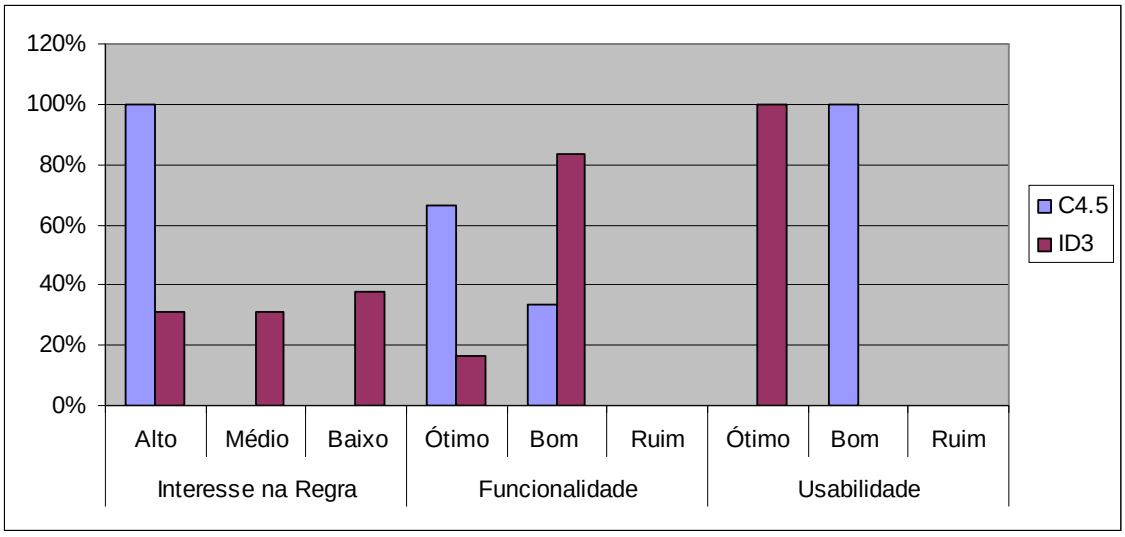


Figura 23. Gráfico comparativo do questionário a respeito dos algoritmos C4.5 e ID3

Comparando os gráficos construídos com os dados dos cálculos realizados e com os dados extraídos do questionário aplicado a especialista da área, pode-se verificar que na avaliação do desempenho realizado a partir dos cálculos os dois algoritmos obtiveram resultados parecidos, se diferenciando no fator de confiança, isso acontece porque o algoritmo ID3 gerou mais regras que o C4.5, por este mesmo motivo também a diferença no gráfico gerado com base nos resultados da entrevista.

CONCLUSÃO

A procura de ferramentas de *Data Mining* no auxílio a tomadas de decisões, está se tornando crescente, isso acontece principalmente por empresas e organizações terem investido mais no armazenamento de informações digitalmente. O aumento da capacidade computacional proporcionou a essas entidades perceberem que essas bases de dados podem conter conhecimento desconhecido.

A tarefa de classificação na *Shell* Orion *Data Mining Engine* conta com os algoritmos C4.5 e ID3, para o *data mining*. Esta ferramenta demonstrou ótimos resultados para os algoritmos ID3 e C4.5 na base de dados da sepse. Esta pesquisa veio fortalecer a confiança na ferramenta desenvolvida por alguns acadêmicos do curso de Ciência da Computação e mantida pelo Grupo de Pesquisa em Inteligência Computacional, da UNESC.

Os módulos utilizados obtiveram êxito quanto a sua exatidão e operacionalidade na realização dos testes. Os algoritmos ID3 e C4.5 na tarefa de classificação do *Data Mining* na base de dados da área médica com informações de pacientes com sepse, obtiveram resultados muito satisfatórios e puderam assim contribuir para a evolução da ferramenta quanto a sua confiabilidade.

O objetivo deste trabalho foi alcançado com a realização de estudos sobre *data* m*ining* e os métodos de avaliação de desempenho, com foco nos métodos específicos para a tarefa de classificação. Na realização dos testes foi possível perceber como é construído um classificador, além da compreensão de como funcionam ferramentas que envolvem descoberta de conhecimento em bases de dados.

Na avaliação do desempenho realizada, pode-se constatar que os algoritmos C4.5 e ID3 não apresentaram erros nas classificações, as taxas de sensibilidade e abrangência reforçaram a confiabilidade nos algoritmos, assim como a entrevista realizada com a especialista da área da base de dados aplicada, obteve bons resultados.

Considerando o tema abordado nesta pesquisa e a necessidade da continuidade do desenvolvimento da *Shell* Orion *Data Mining Engine*. Ficam como sugestões de trabalhos futuros a avaliação do desempenho do algoritmo CART da tarefa de classificação, avaliação de outras tarefas e algoritmos da *Shell* Orion, bem como a aplicação de outras métricas para a avaliação do desempenho em aplicações de *data mining*.

REFERÊNCIAS

AMORIM, Bruno P. **Desenvolvimento de uma Plataforma Híbrida para Descoberta de Conhecimento em Bases de Dados.** 2004. 188 f. Dissertação de Mestrado (Faculdade de Ciência da Computação) - Universidade Federal de Pernambuco, Pernambuco, Recife, 2004.

BORTOLOTTO, Leandro S. **O Método de Redes Neurais pelo Algoritmo de Kohonen para Clusterização na *Shell* Orion *Data Mining Engine*.** 2007. 91 f. Trabalho de Conclusão de Curso (Faculdade de Ciência da Computação) - Universidade do Extremo Sul Catarinense, Criciúma, Santa Catarina, 2007.

BARANAUSKAS, José Augusto. **Extração Automática de Conhecimento por Múltiplos Indutores.** USP – Instituto de Ciências Matemáticas e de Computação, São Carlos – SP. Tese de Doutorado, 2001.

CARVALHO, Deborah R.; NETO, Wilson Alves; BUENO, Marcos. **Avaliação do conhecimento descoberto a partir de algoritmos de *Data Mining*.** Universidade Tuiuti do Paraná, Instituto Paranaense de Desenvolvimento Econômico e Social, Paraná, 2003.

CARVALHO, Luiz Alfredo Vidal de. ***Data* mining: a mineração de dados no marketing, medicina, economia, engenharia e administração.** Rio de Janeiro: Ciência Moderna, 2005.

CASAGRANDE, Diego P. **O Módulo da Tarefa de Associação pelo Algoritmo *Apriori* no Desenvolvimento da *Shell* de *Data Mining Orion*.** 2005. 79 f.Trabalho de Conclusão de Curso (Especialização) - Faculdade de Ciência da Computação, Universidade do Extremo Sul Catarinense, Criciúma, Santa Catarina, 2005.

CASSETTARI JUNIOR, José Márcio. **O método de Lógica *Fuzzy* pelo Algoritmo *Gustafson-Kessel* Na Tarefa de Clusterização Da *Shell* Orion *Data Mining Engine*.** 2008.147 f. Trabalho de Conclusão de Curso (Faculdade de Ciência da Computação) - Universidade do Extremo Sul Catarinense, Criciúma, Santa Catarina, 2008.

CECIL, Russell L.; GOLDMAN, Lee; AUSIELLO, Dennis A. **Cecil: tratado de medicina interna.** Rio de Janero: Elsevier, 2005.

CORDEIRO, João Paulo da C. **Extração de elementos relevantes em texto/páginas da World Wide Web.** 2003. 174 f. Dissertação (Mestrado em Inteligência Artificial e Computação) – Faculdade de Ciências da Universidade do Porto, Porto. Disponível em: <http://www.di.ubi.pt/~jpaulo/tese.pdf> Acesso em: 21 maio 2009.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. **From data mining to knowledge discovery in databases**. 1996. AI Magazine. Menlo Park, 1996. Disponível em: <http://kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf> Acesso em 30 abr. 09.

FREITAS, A.A. A **genetic algorithm for generalized rule induction.** In: R. Roy et al. Advances in Soft Computing - Engineering Design and Manufacturing, 340-353. (Proc. WSC3, *3rd On-Line World Conference on Soft Computing*, hosted on the Internet, July 1998.) Springer-Verlag, 1999.

FONSECA, José Manuel Matos Ribeiro da. **Indução de Árvores de Decisão**. Tese de Mestrado - Departamento de Informática – Universidade Novas de Lisboa, Lisboa 1994.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel**.** ***Data Mining*****: um guia prático**. Rio de Janeiro: Elsevier, 2005.

GUSTAFSON, Donald E.; KESSEL, Willian C. ***Fuzzy Clustering with fuzzy covariance matrix***. Proceedings of the IEE Control and Decision Conference, San Diego, p. 761-766, jan 1979.

HAMILTON, Robert J.; HILDERMAN, Howard J. **knowledge Discovery and Interestingness Measures: A Survey**. Technical report, Department of Computer Science, University of Regina, Regina, Saskatchewan, Canada, 1999.

HAN, Jiawei; KAMBER, Micheline. ***Data mining*****: concepts and techniques.** San Francisco: Morgan Kaufmann, 2001.

HAND, David; HEIKKI, Maminila; SMITH, Padhraic. **Principles of *data mining*.** The MIT Press, 2001.

IBGE, **ENCE – Escola Nacional de Ciências Estatisticas.**Disponível em: < http://www.ence.ibge.gov.br/estatistica/default.asp> Acesso em: 15 jun. 2010.

KANTARDZIC, Mehmed. **Data Mining: Concepts, Models, Methods, and Algorithms**. John Wiley & Sons, 2003.

MARTINS, Denis P. **A Tarefa de Clusterização pelo Método de Particionamento *K-means* na *Shell* Orion *Data Mining Engine*.** 2007. 77 f. Trabalho de Conclusão de Curso (Faculdade de Ciência da Computação) - Universidade do Extremo Sul Catarinense, Criciúma, Santa Catarina, 2007.

MICHIE, D. et al. (Eds.) ***Machine Learning, Neural and Statistical Classification***, 1994.

MONDARDO, Ricardo L. **O Algoritmo C4.5 na Tarefa de Classificação na *Shell* Orion de *Data Mining Engine*.** 2009. 104 f. Trabalho de Conclusão de Curso (Faculdade de Ciência da Computação) - Universidade do Extremo Sul Catarinense, Criciúma, Santa Catarina, 2009.

NASSAR, S. M. (2007), **Tratamento de Incerteza: Sistemas Especialistas Probabilísticos.** Departamento de Informática e de Estatística da Universidade Federal de Santa Catarina. Disponível em: <http://www.inf.ufsc.br/~silvia/disciplinas/sep/material_didatico/MaterialDidatico.pdf> Acesso em: 15 jun. 2010.

NAVEGA, Sergio. **Princípios Essenciais do Data Mining.** Publicado Nos Anais do Infoimagem 2002, Cenadem, Novembro 2002.

PELEGRIN, Diana C. **A Tarefa de Classificação e o Algoritmo ID3 para Indução de Árvores de Decisão na *Shell* de *Data Mining* Orion.** 2005. 92 f. Trabalho de Conclusão de Curso (Faculdade de Ciência da Computação) - Universidade do Extremo Sul Catarinense, Criciúma, Santa Catarina, 2005.

PEREGO, Daniel. **O Método de Lógica *Fuzzy* pelo Algoritmo *Gath-Geva* na Tarefa de Clusterização da *Shell* Orion *Data Mining Engine*.** 2009. 124 f. Trabalho de Conclusão de Curso (Faculdade de Ciência da Computação) - Universidade do Extremo Sul Catarinense, Criciúma, Santa Catarina, 2009.

RAIMUNDO, Lidiane R. **O Algoritmo CART pelo Critério de Ginina Tarefa Classificação da *Shell* de *Data Mining* Orion.** 2007. 106 f. Trabalho de Conclusão de Curso (Faculdade de Ciência da Computação) - Universidade do Extremo Sul Catarinense, Criciúma, Santa Catarina, 2007.

REZENDE, Solange Oliveira. **Sistemas Inteligentes: Fundamentos e Aplicação**. São Paulo: Editora Manole, 2005.

RICH, Elaine; KNIGHT, Kevin: **Inteligência Artificial**. Makron Books. 2ª. Edição. São Paulo, 1994.

ROMÃO, Wesley. **Descoberta de conhecimento relevante em banco de dados sobre ciência e tecnologia.** 2002. 252 f. Tese apresentada ao Programa de Pós-Graduação em Engenharia de Produção como parte dos requisitos para a obtenção do título de Doutor em Engenharia de Produção. Universidade Federal de Santa Catarina, Florianópolis, Santa Catarina, 2002.

RUSSELL, S. J.; NORVIG, P. **Inteligência artificial**. Tradução da segunda edição. Rio de Janeiro: Elsevier, 2004.

SILVA, Marcelino Pereira dos Santos. **SKDQL uma linguagem declarativa de especificação de consultas e processos para descoberta de conhecimento em bancos de dados e sua implementação**. 2002. 113f. Dissertação (Mestre em Ciência da Computação) – Centro de Informática da Universidade Federal de Pernambuco, Recife, BR – PE. Disponível em: <http://www.dpi.inpe.br/~mpss/docs/DissertacaoMarcelino.pdf> Acesso em: 29 abr. 2009.

SOMMERVILLE, Ian. **Engenharia de software.** São Paulo: Addison-Wesley, 2003.

SPIEGEL, MURRAY R. **Estatística**, editora MAKRON, 1993

TWO CROWS CORPORATION, **Introduction to data mining and knowledge discovery**. 3ª. Edição. USA: Two Crows, 1999.

APÊNDICE A – SUBCONJUNTOS DA BASE DE DADOS

Subconjunto 1:

caso	idade	sexo	gravidad	vm	tempo_desfecho	desfecho
1.0	75.0	1.0	3.0	1.0	7.0	obito
2.0	57.0	1.0	3.0	1.0	4.0	obito
3.0	71.0	1.0	3.0	1.0	3.0	obito
4.0	41.0	1.0	2.0	1.0	9.0	obito
5.0	69.0	1.0	3.0	1.0	3.0	obito
6.0	71.0	2.0	3.0	1.0	13.0	obito
7.0	48.0	1.0	1.0	2.0	28.0	cura
8.0	30.0	1.0	1.0	1.0	28.0	cura
9.0	50.0	2.0	2.0	1.0	15.0	obito
10.0	41.0	1.0	3.0	1.0	6.0	obito
11.0	75.0	2.0	2.0	1.0	5.0	obito
12.0	31.0	2.0	2.0	2.0	28.0	cura
13.0	78.0	1.0	2.0	1.0	28.0	cura
14.0	58.0	1.0	3.0	1.0	2.0	obito
15.0	84.0	2.0	1.0	2.0	28.0	cura
16.0	59.0	1.0	2.0	1.0	28.0	cura
17.0	33.0	2.0	3.0	1.0	28.0	cura
18.0	52.0	1.0	3.0	1.0	28.0	cura
21.0	65.0	1.0	3.0	1.0	28.0	cura

Subconjunto 2:

caso	idade	sexo	gravidad	vm	tempo_desfecho	desfecho
19.0	61.0	2.0	3.0	1.0	11.0	obito
20.0	62.0	2.0	3.0	1.0	5.0	obito
22.0	62.0	2.0	3.0	1.0	16.0	obito
23.0	41.0	1.0	2.0	1.0	28.0	cura
24.0	76.0	2.0	2.0	1.0	28.0	cura
25.0	63.0	2.0	2.0	1.0	28.0	cura
26.0	36.0	1.0	3.0	1.0	28.0	cura
27.0	57.0	2.0	3.0	1.0	3.0	obito
28.0	43.0	1.0	3.0	1.0	6.0	obito
29.0	46.0	2.0	3.0	1.0	9.0	obito
30.0	67.0	1.0	3.0	1.0	15.0	obito
31.0	71.0	1.0	2.0	1.0	11.0	obito
32.0	77.0	1.0	3.0	1.0	28.0	cura
33.0	77.0	2.0	2.0	1.0	12.0	obito
34.0	23.0	2.0	3.0	1.0	6.0	obito
35.0	57.0	2.0	1.0	2.0	28.0	cura
37.0	56.0	1.0	1.0	2.0	28.0	cura
38.0	51.0	2.0	1.0	2.0	28.0	cura
39.0	44.0	2.0	1.0	2.0	28.0	cura

Subconjunto 3:

caso	idade	sexo	gravidad	vm	tempo_desfecho	desfecho
36.0	79.0	1.0	2.0	1.0	21.0	obito
40.0	45.0	1.0	1.0	2.0	28.0	cura
41.0	56.0	1.0	1.0	2.0	28.0	cura
42.0	47.0	2.0	1.0	2.0	28.0	cura
43.0	57.0	1.0	1.0	2.0	28.0	cura
44.0	56.0	1.0	1.0	2.0	28.0	cura
45.0	48.0	1.0	1.0	2.0	28.0	cura
46.0	55.0	2.0	1.0	2.0	28.0	cura
47.0	41.0	2.0	1.0	2.0	28.0	cura
48.0	55.0	1.0	1.0	2.0	28.0	cura
52.0	47.0	2.0	3.0	1.0	5.0	obito
53.0	54.0	1.0	2.0	1.0	12.0	obito
54.0	58.0	2.0	3.0	1.0	4.0	obito
55.0	67.0	1.0	3.0	1.0	3.0	obito
56.0	74.0	1.0	2.0	1.0	15.0	obito
58.0	77.0	2.0	2.0	1.0	11.0	obito
59.0	29.0	2.0	3.0	1.0	5.0	obito
60.0	65.0	1.0	2.0	1.0	3.0	obito
61.0	55.0	1.0	2.0	1.0	17.0	obito

Subconjunto 4:

caso	idade	sexo	gravidad	vm	tempo_desfecho	desfecho
49.0	65.0	1.0	1.0	2.0	28.0	cura
50.0	61.0	2.0	1.0	2.0	28.0	cura
51.0	33.0	1.0	3.0	1.0	28.0	cura
57.0	75.0	1.0	3.0	1.0	28.0	cura
62.0	44.0	1.0	3.0	1.0	2.0	obito
63.0	59.0	1.0	2.0	1.0	14.0	obito
64.0	56.0	1.0	3.0	1.0	3.0	obito
65.0	78.0	2.0	3.0	1.0	6.0	obito
66.0	43.0	1.0	1.0	2.0	28.0	cura
67.0	33.0	1.0	1.0	1.0	28.0	cura
68.0	53.0	2.0	2.0	1.0	15.0	obito
69.0	32.0	1.0	3.0	1.0	3.0	obito
70.0	77.0	2.0	2.0	1.0	11.0	obito
71.0	43.0	2.0	3.0	1.0	28.0	cura
72.0	71.0	1.0	2.0	1.0	28.0	cura
73.0	55.0	1.0	3.0	1.0	4.0	obito
74.0	89.0	2.0	1.0	2.0	28.0	cura
77.0	51.0	2.0	3.0	1.0	3.0	obito
78.0	49.0	1.0	2.0	1.0	5.0	obito

Subconjunto 5:

caso	idade	sexo	gravidad	vm	tempo_desfecho	desfecho
75.0	55.0	1.0	3.0	1.0	28.0	cura
76.0	43.0	1.0	3.0	1.0	28.0	cura
79.0	62.0	2.0	3.0	1.0	3.0	obito
80.0	64.0	1.0	3.0	1.0	4.0	obito
81.0	66.0	1.0	2.0	1.0	13.0	obito
82.0	73.0	1.0	3.0	1.0	28.0	cura
83.0	56.0	2.0	2.0	1.0	15.0	obito
84.0	37.0	2.0	3.0	1.0	7.0	obito
85.0	55.0	1.0	2.0	1.0	9.0	obito
86.0	76.0	1.0	2.0	1.0	11.0	obito
87.0	59.0	1.0	3.0	1.0	7.0	obito
88.0	68.0	1.0	2.0	1.0	6.0	obito
89.0	49.0	1.0	3.0	1.0	5.0	obito
90.0	68.0	2.0	3.0	1.0	2.0	obito
91.0	58.0	1.0	2.0	1.0	28.0	cura
92.0	43.0	1.0	2.0	1.0	28.0	cura
93.0	51.0	1.0	2.0	1.0	28.0	cura
94.0	49.0	1.0	2.0	1.0	28.0	cura
95.0	52.0	1.0	2.0	1.0	28.0	cura
96.0	53.0	2.0	2.0	1.0	28.0	cura

APÊNDICE B – CÁLCULO DA TAXA DE ACERTO PARA CADA ITERAÇÃO NA VALIDAÇÃO CRUZADA DO ALGORITMO C4.5

Taxa de acerto = Número de classificações corretas / Número de dados aplicáveis

Iteração 1:

a) SE tempodesfecho <= 15.0 ENTÃO óbito, contendo 10 casos:

Taxa de acerto: $\frac{10}{10} = 1 = 100\%$

b) SE tempodesfecho > 15.0 ENTÃO cura, contendo 9 casos:

Iteração 2:

a) SE tempodesfecho <= 16.0 ENTÃO óbito, contendo 10 casos:

Taxa de acerto: $\frac{10}{10} = 1 = 100\%$

b) SE tempodesfecho > 16.0 ENTÃO cura, contendo 9 casos:

Taxa de acerto: $\frac{9}{9} = 1 = 100\%$

Iteração 3:

a) SE vm <= 1.0 ENTÃO óbito, contendo 10 casos:

Taxa de acerto: $\frac{10}{10} = 1 = 100\%$

b) SE vm > 1.0 ENTÃO cura, contendo 9 casos:

Taxa de acerto: $\frac{9}{9} = 1 = 100\%$

Iteração 4:

a) SE tempodesfecho <= 15.0 ENTÃO óbito, contendo 10 casos:

Taxa de acerto: $\frac{10}{10} = 1 = 100\%$

b) SE tempodesfecho > 15.0 ENTÃO cura, contendo 9 casos:

Taxa de acerto: $\frac{9}{9} = 1 = 100\%$

Iteração 5:

a) SE tempodesfecho <= 15.0 ENTÃO óbito, contendo 11 casos:

Taxa de acerto: $\frac{11}{11} = 1 = 100\%$

b) SE tempodesfecho > 15.0 ENTÃO cura, contendo 9 casos:

Taxa de acerto: $\frac{9}{9} = 1 = 100\%$

APÊNDICE C – CÁLCULO DO FATOR DE CONFIANÇA DAS REGRAS GERADAS PELO ID3 NA *SHELL* ORION

SE tempodesfecho = 3.0 ENTÃO óbito, contendo 9 casos.

$$\left(9-\frac{1}{2}\right)/9 = 94{,}44\%$$

SE tempodesfecho = 13.0 ENTÃO óbito, contendo 2 casos.

$$\left(2-\frac{1}{2}\right)/2 = 75\%$$

SE tempodesfecho = 15.0 ENTÃO óbito, contendo 5 casos.

$$\left(5-\frac{1}{2}\right)/5 = 90\%$$

SE tempodesfecho = 6.0 ENTÃO óbito, contendo 5 casos.

$$\left(5-\frac{1}{2}\right)/5 = 90\%$$

SE tempodesfecho = 5.0 ENTÃO óbito, contendo 6 casos.

$$\left(6-\frac{1}{2}\right)/6 = 91{,}66\%$$

SE tempodesfecho = 2.0 ENTÃO óbito, contendo 3 casos.

$$\left(3-\frac{1}{2}\right)/3 = 83{,}33\%$$

SE tempodesfecho = 11.0 ENTÃO óbito, contendo 5 casos.

$$\left(5-\frac{1}{2}\right)/5 = 90\%$$

SE tempodesfecho = 16.0 ENTÃO óbito, contendo 1 casos.

$$\left(1-\frac{1}{2}\right)/1 = 50\%$$

SE tempodesfecho = 12.0 ENTÃO óbito, contendo 2 casos.

$$\left(2-\frac{1}{2}\right)/2 = 75\%$$

SE tempodesfecho = 21.0 ENTÃO óbito, contendo 1 casos.

$$\left(1-\frac{1}{2}\right)/1 = 50\%$$

SE tempodesfecho = 4.0 ENTÃO óbito, contendo 4 casos.

$$\left(4-\frac{1}{2}\right)/4 = 87{,}5\%$$

SE tempodesfecho = 17.0 ENTÃO óbito, contendo 1 casos.

$$\left(1-\frac{1}{2}\right)/1 = 50\%$$

SE tempodesfecho = 14.0 ENTÃO óbito, contendo 1 casos.

$$\left(1-\frac{1}{2}\right)/1 = 50\%$$

SE tempodesfecho = 7.0 ENTÃO óbito, contendo 3 casos.

$$\left(3-\frac{1}{2}\right)/3 = 83{,}33\%$$

SE tempodesfecho = 28.0 ENTÃO cura, contendo 45 casos.

$$\left(45-\frac{1}{2}\right)/45 = 98{,}88\%$$

APÊNDICE D – CÁLCULO DA TAXA DE ACERTO PARA CADA ITERAÇÃO NA VALIDAÇÃO CRUZADA DO ALGORITMO ID3

Iteração 2:

a) SE tempodesfecho = 11 ENTÃO óbito, contendo 2 casos

SE tempodesfecho = 5 ENTÃO óbito, contendo 1 casos

SE tempodesfecho = 16 ENTÃO óbito, contendo 1 casos

SE tempodesfecho = 3 ENTÃO óbito, contendo 1 casos

SE tempodesfecho = 6 ENTÃO óbito, contendo 2 casos

SE tempodesfecho = 9 ENTÃO óbito, contendo 1 casos

SE tempodesfecho = 15 ENTÃO óbito, contendo 1 casos

SE tempodesfecho = 12 ENTÃO óbito, contendo 1 casos:

$$\text{Taxa de acerto: } \frac{10}{10} = 1 = 100\%$$

b) SE tempodesfecho = 28 ENTÃO cura, contendo 9 casos:

$$\text{Taxa de acerto: } \frac{9}{9} = 1 = 100\%$$

Iteração 3:

a) SE gravidad = 2 ENTÃO óbito, contendo 6 casos

SE gravidad = 3 ENTÃO óbito, contendo 4 casos:

$$\text{Taxa de acerto: } \frac{10}{10} = 1 = 100\%$$

b) SE gravidad = 1 ENTÃO cura, contendo 9 casos:

$$\text{Taxa de acerto: } \frac{9}{9} = 1 = 100\%$$

Iteração 4:

a) SE idade = 44.0 ENTÃO óbito, contendo 1 casos

SE idade = 59.0 ENTÃO óbito, contendo 1 casos

SE idade = 56.0 ENTÃO óbito, contendo 1 casos

SE idade = 78.0 ENTÃO óbito, contendo 1 casos

SE idade = 53.0 ENTÃO óbito, contendo 1 casos

SE idade = 32.0 ENTÃO óbito, contendo 1 casos

SE idade = 77.0 ENTÃO óbito, contendo 1 casos

SE idade = 55.0 ENTÃO óbito, contendo 1 casos

SE idade = 51.0 ENTÃO óbito, contendo 1 casos

SE idade = 49.0 ENTÃO óbito, contendo 1 casos:

$$\text{Taxa de acerto: } \frac{10}{10} = 1 = 100\%$$

b) SE idade = 65.0 ENTÃO cura, contendo 1 casos

SE idade = 461.0 ENTÃO cura, contendo 1 casos

SE idade = 33.0 ENTÃO cura, contendo 2 casos

SE idade = 75.0 ENTÃO cura, contendo 1 casos

SE idade = 43.0 ENTÃO cura, contendo 2 casos

SE idade = 71.0 ENTÃO cura, contendo 1 casos

SE idade = 89.0 ENTÃO cura, contendo 1 casos:

$$\text{Taxa de acerto: } \frac{9}{9} = 1 = 100\%$$

Iteração 5:

a) SE tempodesfecho = 13 ENTÃO óbito, contendo 1 casos

SE tempodesfecho = 15 ENTÃO óbito, contendo 1 casos

SE tempodesfecho = 7 ENTÃO óbito, contendo 2 casos

SE tempodesfecho = 9 ENTÃO óbito, contendo 1 casos

SE tempodesfecho = 11 ENTÃO óbito, contendo 1 casos

SE tempodesfecho = 6 ENTÃO óbito, contendo 1 casos

SE tempodesfecho = 5 ENTÃO óbito, contendo 1 casos

SE tempodesfecho = 2 ENTÃO óbito, contendo 1 casos

SE tempodesfecho = 3 ENTÃO óbito, contendo 1 casos

SE tempodesfecho = 4 ENTÃO óbito, contendo 1 casos:

Taxa de acerto: $\frac{11}{11} = 1 = 100\%$

b) SE tempodesfecho = 28 ENTÃO cura, contendo 1 casos:

Taxa de acerto: $\frac{9}{9} = 1 = 100\%$

APÊNDICE E – ENTREVISTA COM O ESPECIALISTA PARA VALIDAÇÃO DO CONHECIMENTO RESULTANTE DO ALGORITMO C4.5

Validação feita por:
Nome: Cristiane Damiani Pomari
Titulo: Enfermeira – Mestre em Ciências da Saúde

Entrevista com o especialista da área para a validação quanto a funcionalidade e usabilidade.
Módulo Classificação da Shell Orion Data Mining Engine
Algoritmo C4.5

Funcionalidade da aplicação;

1. A Shell Orion faz o que necessita de forma eficiente?

(X)Ótimo; ( )Bom; ( )Ruim.

2. O conhecimento gerado corresponde as expectativas?

(X)Ótimo; ( )Bom; ( )Ruim.

3. As respostas obtidas acrescentaram conhecimento útil a sua atividade?

( )Ótimo; (X)Bom; ( )Ruim.

4. O conhecimento obtido é de fácil compreensão?

(X)Ótimo; ( )Bom; ( )Ruim.

5. As respostas da Shell ajudaram a confirmar o conhecimento pré-existente?

(X)Ótimo; ( )Bom; ( )Ruim.

6. A Shell contribuiu com novos conhecimentos?

( )Ótimo; (X)Bom; ( )Ruim.

Usabilidade;

7. Fácil de aprender como se usa a Shell?

( )Ótimo; (X)Bom; ( )Ruim.

8. Fácil de Usar?

( )Ótimo; (X)Bom; ( )Ruim.

9. O ambiente da aplicação é agradável?

( )Ótimo; (X)Bom; ( )Ruim.

Validação e avaliação das regras descobertas.

O conhecimento obtido a partir de cada regra pode ser classificado de acordo com o interesse em: baixo interesse; médio interesse; ou grande interesse.

Previsão da Regra: Desfecho como óbito ou cura.
Algoritmo: C4.5

1. SE tempodesfecho <= 21.0 dias
   |---ENTÃO - obito (51)

( ) Baixo Interesse; ( ) Interesse Médio; (X) Alto Interesse.

2. SE tempodesfecho > 21.0 dias
   |---ENTÃO - cura (45)

( ) Baixo Interesse; ( ) Interesse Médio; (X) Alto Interesse.

Observações:
Ao contrário do algoritimo ID3, o C4.5 dicotomiza o tempo do desfecho, o que facilita a interpretação dos dados. Porém, também seria interessante a associação de um terceiro atributo.
O ponto de corte do C4.5 facilita a interpretação dos dados.

APÊNDICE F – ENTREVISTA COM O ESPECIALISTA PARA VALIDAÇÃO DO CONHECIMENTO RESULTANTE DO ALGORITMO ID3

Validação feita por:

Nome: Cristiani Damiani Tomasi

Titulo: Enfermeira - Mestre em Ciências da Saúde

Bolsista de Doutorado - FAPESC Orientador: Prof. Dr. Felipa Dal-Pizzol

Linha de Pesquisa: Fisiopatologia Experimental Área de Estudo: Delirium em pacientes internados em Unidade de Terapia Intensiva

Entrevista com o especialista da área para a validação quanto a funcionalidade e usabilidade.

Módulo Classificação da Shell Orion Data Mining Engine

Algoritmo ID3.

Funcionalidade da aplicação;

1. A Shell Orion faz o que necessita de forma eficiente?

( )Ótimo; (X)Bom; ( )Ruim.

2. O conhecimento gerado corresponde as expectativas?

( )Ótimo; (X)Bom; ( )Ruim.

3. As respostas obtidas acrescentaram conhecimento útil a sua atividade?

( )Ótimo; (X)Bom; ( )Ruim.

4. O conhecimento obtido é de fácil compreensão?

(X)Ótimo; ( )Bom; ( )Ruim.

5. As respostas da Shell ajudaram a confirmar o conhecimento pré-existente?

( )Ótimo; (X)Bom; ( )Ruim.

6. A Shell contribuiu com novos conhecimentos?

( )Ótimo; (X)Bom; ( )Ruim.

Usabilidade;

7. Fácil de aprender como se usa a Shell?

(X)Ótimo; ( )Bom; ( )Ruim.

8. Fácil de Usar?

(X)Ótimo; ( )Bom; ( )Ruim.

9. O ambiente da aplicação é agradável?

(X)Ótimo; ( )Bom; ( )Ruim.

Validação e avaliação das regras descobertas.

O conhecimento obtido a partir de cada regra pode ser classificado de acordo com o interesse em: baixo interesse; médio interesse; ou grande interesse.

Previsão da Regra: Desfecho como óbito (1) ou cura (2).
Algoritmo: ID3

1. SE tempodesfecho = 7 dias
   ENTÃO
   ---> desfecho = 1 (3 casos)

( ) Baixo Interesse; (X) Interesse Médio; ( ) Alto Interesse.

2. SE tempodesfecho = 4 dias
   ENTÃO
   ---> desfecho = 1 (4 casos)

( ) Baixo Interesse; (X) Interesse Médio; ( ) Alto Interesse.

3. SE tempodesfecho = 3 dias
   ENTÃO
   ---> desfecho = 1 (9 casos)

( ) Baixo Interesse; ( ) Interesse Médio; (X) Alto Interesse.

4. SE tempodesfecho = 9 dias
   ENTÃO
   ---> desfecho = 1 (3 casos)

( ) Baixo Interesse; (X) Interesse Médio; ( ) Alto Interesse.

5. SE tempodesfecho = 13 dias
   ENTÃO
   ---> desfecho = 1 (2 casos)

(X) Baixo Interesse; ( ) Interesse Médio; ( ) Alto Interesse.

6. SE tempodesfecho = 28 dias
   ENTÃO
   ---> desfecho = 2 (45 casos)

( ) Baixo Interesse; (X) Interesse Médio; (X) Alto Interesse.

7. SE tempodesfecho = 15 dias
ENTÃO
--> desfecho = 1 (5 casos)

( ) Baixo Interesse; (X) Interesse Médio; ( ) Alto Interesse.

8. SE tempodesfecho = 6 dias
ENTÃO
--> desfecho = 1 (5 casos)

( ) Baixo Interesse; ( ) Interesse Médio; (X) Alto Interesse.

9. SE tempodesfecho = 5 dias
ENTÃO
--> desfecho = 1 (6 casos)

( ) Baixo Interesse; ( ) Interesse Médio; (X) Alto Interesse.

10. SE tempodesfecho = 2 dias
ENTÃO
---> desfecho = 1 (3 casos)

( ) Baixo Interesse; ( ) Interesse Médio; (X) Alto Interesse.

11. SE tempodesfecho = 11 dias
ENTÃO
--> desfecho = 1 (5 casos)

( ) Baixo Interesse; (X) Interesse Médio; ( ) Alto Interesse.

12. SE tempodesfecho = 16 dias
ENTÃO
---> desfecho = 1 (1 casos)

(X) Baixo Interesse; ( ) Interesse Médio; ( ) Alto Interesse.

13. SE tempodesfecho = 12 dias
ENTÃO
---> desfecho = 1 (2 casos)

(X) Baixo Interesse; ( ) Interesse Médio; ( ) Alto Interesse.

14. SE tempodesfecho = 21 dias
ENTÃO
---> desfecho = 1 (1 casos)

(X) Baixo Interesse; ( ) Interesse Médio; ( ) Alto Interesse.

15. SE tempodesfecho = 17 dias
ENTÃO
---> desfecho = 1 (1 casos)

(X) Baixo Interesse; ( ) Interesse Médio; ( ) Alto Interesse.

16. SE tempodesfecho = 14 dias
ENTÃO
--> desfecho = 1 (1 casos)

(X) Baixo Interesse; ( ) Interesse Médio; ( ) Alto Interesse.

Observações:

As relações feitas a partir dos dados, são muito interessantes, no entanto, seria de maior valor a associação de três variáveis, como por exemplo: tempo desdesfecho; desfecho e gravidade da sepse. Visto que para trabalhadores da área da saúde, já se sabe que há associação da gravidade da sepse com o desfecho. Não associando as 3 variáveis algumas informações ficam vagas. Ex: será que o tempo de desfecho é maior porque a sepse foi mais grave? A ~~maior~~ mortalidade nos primeiros dias é maior porque a sepse é mais grave?
Mesmo com estas "lacunas" a ferramenta é interessante, visto que nos oferece várias combinações de atributos para a ~~anal~~ análise.

APÊNDICE G – ARTIGO

Avaliação de Desempenho dos Algoritmos para Classificação em *Data Mining*, ID3 E C4.5, na *Shell* Orion

Abstract. *The process of extracting knowledge from databases has in the main stage, the data mining process. This research was conducted to evaluate performance during the classification's tasks by ID3 and C4.5 algorithms. This implementations are inside the software called Shell Orion Data Mining Engine. During the developing the performance evaluation statistical methods were employed such as: cross validation; calculation of confidence factor; confusion matrix and calculation of average error. Also was conducted the validation of the knowledge, presenting the results in forms of rules to the domain's specialist of application who validated the functionality and usability. The Shell Orion algorithms' evaluation obtained satisfactory results, mainly for tax exemption of misclassification. It confirms the usefulness and relevance of the implemented algorithms.*

Keywords: *Data Mining; Shell Orion Data Mining Engine, Classification, Performance Evaluation, Knowledge Validation..*

Resumo. *O processo de extração de conhecimento de bases de dados tem como principal etapa o data mining. Nesta pesquisa foi desenvolvida a avaliação de desempenho da tarefa de classificação pelos algoritmos ID3 e C4.5, na Shell Orion Data Mining Engine. Na avaliação de desempenho foram empregados métodos estatísticos como: validação cruzada; fator de confiança; matriz confusão e erro médio. Na avaliação do conhecimento, foram apresentados os resultados obtidos em formas de regras ao especialista da área, que avaliou o interesse na regra, funcionalidade e usabilidade. Os algoritmos da Orion obtiveram isenção de taxas de erro de classificação, confirmando a utilidade e relevância dos algoritmos avaliados.*

Palavras-Chave: *Data Mining; Shell Orion Data Mining Engine, Classificação, Avaliação de Desempenho, Validação do Conhecimento*.

1. Introdução

O desenvolvimento da tecnologia da informação possibilitou uma maior capacidade de armazenamento de dados, mediante isso, teve-se a necessidade de transformar estes dados em conhecimento úteis. Assim, novas metodologias e ferramentas computacionais fizeram-se primordiais para auxiliar a análise. Surgiu então, o conceito de Descoberta de Conhecimento em Bases de Dados, que visa a identificação de padrões e relações. A etapa de *data mining* é considerada a principal, pois ela é efetivamente a responsável pela identificação de relações, envolvendo métodos e algoritmos que evidenciam padrões e vinculações entre as variáveis registradas na base de dados. (GOLDSHMDT; PASSOS, 2005).

Na aplicação de *data mining* empregam-se ferramentas computacionais, denominadas *Shell*. Essas ferramentas em sua maioria são comerciais, considerando isso, o Grupo de Pesquisa em Inteligência Computacional Aplicada, do curso de Ciência da Computação da Universidade do Extremo Sul Catarinense (UNESC), tem em desenvolvimento desde o ano de 2005 o projeto da *Shell* Orion *Data Mining Engine*. Cada algoritmo tem sido desenvolvido por um acadêmico da UNESC em seu Trabalho de Conclusão de Curso.

Esta ferramenta tem sido utilizada para análise de diferentes tipos de bases de dados, como por exemplo, a oriunda do Curso de Medicina da UNESC, contendo registros referentes a sepse em pacientes da Unidade de Terapia Intensiva do Hospital de Clínicas de Porto

Alegre. A sepse é uma doença desenvolvida a partir de uma infecção sistêmica grave, conhecida como infecção generalizada (CECIL; GOLDMAN; AUSIELLO, 2005).

A Shell tem gerado resultados satisfatórios, porém torna-se interessante avaliar as relações e os padrões que vêm sendo descobertos. Salienta-se que esta avaliação é importante para verificar se o conhecimento descoberto está correto e atende as expectativas do especialista do domínio de aplicação, para que se possa confiar nas relações obtidas (SOMMERVILLE, 2003).

Esta pesquisa apresenta uma metodologia para avaliar o desempenho e validar o conhecimento descoberto pela aplicação do método de classificação e dos algoritmos de *data mining*, ID3 e C4.5 implementado na *Shell* Orion, aplicando-se para análise uma base de dados.

2. Avaliação do Desempenho do Módulo de Classificação da *Shell* Orion *Data Mining Engine*

Na classificação foi utilizada a base de dados da sepse. Disponibilizada pelo Dr. Felipe Dal Pizzol, do curso de Medicina da UNESC, contendo registros de pacientes com sepse da Unidade de Terapia Intensiva (UTI) do Hospital das Clínicas de Porto Alegre, Rio Grande do Sul. A base de dados contém no total 96 registros, cada registro possui 45 atributos com informações dos pacientes e de exames realizados durante a internação.

O objetivo da classificação com os dados da base da sepse é saber o desfecho do paciente que pode ser óbito ou cura, a base conta com 45 registros com o desfecho cura e 51 registros com o desfecho óbito, os algoritmos de classificação calculam o atributo com o maior ganho para eleger como atributo classificador.

As seguintes etapas foram adotadas para a validação do conhecimento: levantamento bibliográfico dos temas abordados; normalização da base de acordo com as necessidades dos algoritmos; execução dos algoritmos de classificação a partir da base; realização dos cálculos de avaliação; entrevista com o especialista.

A avaliação com o especialista foi realizada por meio de um questionário com alternativas relacionadas a:

a) funcionalidade da aplicação: ótimo, bom ou ruim;

b) usabilidade: ótimo, bom ou ruim;

c) interesse nas regras geradas: baixo, médio e alto interesse.

Após a etapa de pré-processamento, foi possível realizar a classificação e com base nos resultados obtidos calcular a matriz confusão, o fator de confiança, realizar a validação cruzada e o cálculo do erro médio.

3. Cálculo da Matriz Confusão dos algoritmos C4.5 e ID3

Na realização da avaliação foram utilizados todos os registros da base de dados. Os algoritmos C4.5 e ID3 elegeram como atributo de melhor ganho o *tempodesfecho* que contém a quantidade de dias que o paciente permaneceu internado. O algoritmo C4.5 gerou 2 regras: 1 para o saída óbito, totalizando 51 casos, e 1 para o saída cura, totalizando 45 casos. Já o ID3 gerou 16 regras: 15 regras para saída óbito, totalizando 51 casos; e 1 regra para saída cura, totalizando 45 casos. A matriz confusão dos dois algoritmos é apresentada na Figura 1.

		Classe Prevista	
		Óbito	Cura
Classe Real	Óbito	51	0
	Cura	0	45

Figura 1. Matriz Confusão dos algoritmos C4.5 e ID3

Acurácia: $\frac{VP+VN}{n}.100\% = \frac{51+45}{96}.100\% = 100\%$

Erro: $\frac{FP+FN}{n}.100\% = \frac{0+0}{96}.100\% = 0\%$

Cálculo da taxa dos verdadeiros positivos (sensibilidade e abrangência):

$\frac{VP}{VP+FP}.100\% = \frac{51}{51+0}.100\% = 100\%$

Cálculo da taxa de verdadeiros negativos (especificidade):

$\frac{VN}{VN+FP}.100\% = \frac{45}{45+0}.100\% = 100\%$

Precisão: $\frac{VP}{VP+FP}.100\% = \frac{51}{51+0}.100\% = 100\%$

Onde:

e) VP: classificações verdadeiras positivo;

f) VN: classificações verdadeiras negativo;

g) FP: classificações falsas positivo;

h) FN: classificações falsas negativo.

É de fácil visualização na matriz confusão, a ausência de classificações erradas, o que comprova a qualidade do conhecimento descoberto pelos algoritmos C4.5 e ID3, que pode ser comparado ao de um classificador ideal, uma vez que não comete erros.

As taxas de sensibilidade e especificidade obtiveram um percentual de 100% o que demonstra que o classificador obteve os resultados esperados. A sensibilidade garante que não se tenha perda de casos candidatos no auxilio a tomada de decisão, já a especificidade garante que não sejam levados em consideração casos com resultados negativos.

4. Cálculo da taxa de Acerto dos Algoritmos C4.5 e ID3

Considerando as regras geradas com a classificação pelos algoritmos C4.5 e ID3 na *Shell* Orion, realizou-se o cálculo da taxa de acerto para cada saída da regra, que podem ser visualizados na Tabela 1:

$$\text{Taxa de acerto} = \frac{\text{Número de classificações corretas}}{\text{Número de dados aplicáveis}}$$

Tabela 1. Taxas de acerto C4.5 e ID3

	Taxa de Acerto	
	Desfecho Óbito	Desfecho Cura
C4.5	100%	100%
ID3	100%	100%

A taxa de acerto é igual a 100% para as duas regras, uma vez que o algoritmo não comete erros.

5. Cálculo do Fator de Confiança dos Algoritmos C4.5 e ID3

Realização do cálculo do fator de confiança para cada regra gerada a partir da classificação nos algoritmo C4.5 e ID3, visualizados na Tabela 2.

Fator de confiança = (X – 1/2)/Y

Onde X é o número de registros que satisfazem o conseqüente e o antecedente da regra e Y o total de registros que satisfazem o antecedente da regra (ROMÃO, 2002).

Tabela 2. Fator de confiança C4.5 e ID3

	Fator de Confiança	
	Desfecho Óbito	Desfecho Cura
C4.5	99,01%	98,88%
ID3	76,24%	98,88%

O fator (- ½) penaliza as regras com menor abrangência de dados, não penalizando significativamente regras com abrangência maior. Este fator ajuda a identificar regras que se ajustam demasiadamente a peculiaridade da base de dados.

6. Cálculo da Validação Cruzada dos Algoritmos C4.5 e ID3

No cálculo da validação cruzada dividiu-se a base de dados em cinco subconjuntos, sendo que destes subconjuntos, quatro eram compostos por dezenove registros e um subconjunto possuía os vinte dados restantes, que totalizaram os 96 registros. Cada subconjunto compondo dez casos de desfecho óbito e nove casos de desfecho cura, com exceção do ultimo com onze casos com desfecho óbito e nove com desfecho cura, o Apêndice A contém os subconjuntos gerados. Cada subconjunto foi submetido ao algoritmo C4.5 e ao ID3 e calculado para cada iteração a taxa de acerto das regra gerada, pora cada algoritmo.

Com os resultados das taxas de acerto é possível calcular a média (ā) e do desvio padrão (dp) para cada regra, Tabela 3. Regras com desfecho óbito:

$$\bar{a} = \frac{1}{k}\sum_{i=1}^{k} x_i \;=\; \frac{1}{5}.5 = 1$$

$$dp(\bar{a}) = \sqrt{\frac{1}{k}\left[\frac{1}{k-1}\sum_{i=1}^{k}(x_i - \bar{a})^2\right]} = \sqrt{\frac{1}{5}.\left[\frac{1}{4}.5\right]} = 0{,}5$$

Onde:

e) ā = média;

f) dp = desvio padrão;

g) x_i = taxa de acerto da partição i;

h) K = número de partições.

Cálculo da média (ā) e do desvio padrão (dp) para as regras com desfecho cura:

$$\bar{a} = \frac{1}{k}\sum_{i=1}^{k} x_i = \frac{1}{5}.5 = 1$$

$$dp(\bar{a}) = \sqrt{\frac{1}{k}\left[\frac{1}{k-1}\sum_{i=1}^{k}(x_i - \bar{a})^2\right]} = \sqrt{\frac{1}{5}.\left[\frac{1}{4}.5\right]} = 0{,}5$$

Tabela 3. Validação cruzada dos algoritmos C4.5 e ID3

	Validação Cruzada			
	Média Desfecho Óbito	Desvio Padrão Desfecho Óbito	Média Desfecho Cura	Desvio Padrão Desfecho Cura
C4.5	1	0,5	1	0,5
ID3	1	0,5	1	0,5

O cálculo da média e do desvio padrão na validação cruzada permitem estimar o acerto dos dados gerados a partir da tarefa de classificação, muito utilizado para a validação e avaliação do conhecimento descoberto.

7. Cálculo do Erro Médio dos Algoritmos C4.5 e ID3

Para calcular o erro médio dos algoritmos C4.5 e ID3 utilizou-se a mesma base particionada, do cálculo da validação cruzada, procedimento realizado para cada agoritmo, Tabela 4.

Cálculo do erro por partição, a partir da expressão abaixo:

$$err(h) = \frac{1}{n}\sum_{n=1}^{n} \left\| y_i \neq h(x_i) \right\|$$

Onde:

e) n é o número de partições do conjunto de dados;

f) y_i é a classe real;

g) $h(x_i)$ é o retorno do classificador;

h) $\left\| y_i \neq h(x_i) \right\|$ retorna 1 se verdadeira e 0 se falsa.

Partição 1:

$$err(h) = \frac{1}{19}.(0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0) = 0$$

Partição 2:

$$err(h) = \frac{1}{19}.(0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0) = 0$$

Partição 3:

$$err(h) = \frac{1}{19}.(0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0) = 0$$

Partição 4:

$$err(h) = \frac{1}{19}.(0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0) = 0$$

Partição 5:

$$err(h) = \frac{1}{20}.(0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0) = 0$$

Cálculo da média dos erros das partições:

$$mean(A) = \frac{1}{n}\sum_{i=1}^{n} err(h)$$

Onde:

c) n é o número de partições do conjunto de dados;

d) $err(h)$ é o resultado do cálculo do erro.

$$mean(A) = \frac{1}{5}.(0+0+0+0+0) = 0$$

Cálculo do desvio padrão e da variância do classificador:

$$\text{var}(A) = \frac{1}{n}\left[\frac{1}{n-1}\sum_{i=1}^{n}(err(h_i) - mean(A))^2\right]$$

Onde:

e) var(A) é a variância das partições;

f) n é o número de partições do conjunto de dados;

g) $err(h_i)$ é o erro obtido em cada partição;

h) mean(A) é o erro médio.

$$\text{var}(A) = \frac{1}{5}\left[\frac{1}{5}.(0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2\right] = 0$$

Cálculo do desvio padrão do classificador:

$$dp(A) = \sqrt{\text{var}(A)} = \sqrt{0} = 0$$

Tabela 4. Cálculo do Erro Médio dos algoritmos C4.5 e ID3

	Erro Médio			
	Média Desfecho Óbito	Desvio Padrão Desfecho Óbito	Média Desfecho Cura	Desvio Padrão Desfecho Cura
C4.5	0	0	0	0
ID3	0	0	0	0

O erro médio do classificado é 0, ou seja o classificador não apresenta erros. Os valores obtidos com os cálculos do erro médio nos algoritmos C4.5 e ID3 demonstram que a ferramenta é confiável.

8. Método de Validação do Conhecimento dos Algoritmos C4.5 E ID3

A validação do conhecimento foi efetuada com base na entrevista com o especialista, aplicada na forma de questionário, o especialista da área que concedeu a entrevista foi a MSc. Cristiane Damiani Tomasi, enfermeira e mestre em Ciência da Saúde, bolsista de Doutorado orientada pelo Prof. Dr. Felipe Dal Pizzol, da Universidade do Extremo Sul Catarinense.

A metodologia aplicada na entrevista seguiu os seguintes passos:

f) apresentação da base de dados;

g) breve explicação sobre o funcionamento dos algoritmos C4.5 e ID3;

h) apresentação da *Shell* Orion *Data Mining Engine*;

i) aplicação dos algoritmos;

j) aplicação dos questionários.

Os questionários aplicados continham questões sobre a funcionalidade e usabilidade da aplicação, e apresentava as regras geradas pelo algoritmo que podiam ser classificadas conforme o interesse em: Alto; Médio, e baixo interesse. O questionário submetido para a validação do conhecimento a partir do C4.5 encontra-se no Apêndice E, e o aplicado para o ID3 no Apêndice F.

9. Resultados Obtidos

Os resultados obtidos na avaliação de desempenho da tarefa de classificação da *Shell* Orion *Data Mining Engine*, foram a partir da análise do funcionamento do módulo e de seus algoritmos.

Inicialmente, verificou-se as capacidades dos algoritmos em relação a descoberta de grupos na base de dados de pacientes com sepse, com o objetivo de avaliar os resultados obtidos. A segunda etapa foi a realização dos cálculos para a validação do desempenho obtidos com os algoritmos C4.5 e ID3. Como terceira etapa foi desenvolvida e aplicada a entrevista com o especialista da área.

9.1 Validação e análise do Algoritmo C.45 e ID3

A finalização da execução do algoritmo C4.5 na *Shell* Orion *Data Mining Engine* através da base de dados da sepse, utilizando o todos os registros para validação e para os testes, o algoritmo gerou apenas duas regras, uma contendo 51 casos de pacientes com o desfecho óbito e a segunda com 45 casos de pacientes com desfecho cura, enquanto o algoritmo ID3 gerou 16 regras, 15 regras contendo 51 casos de pacientes com o desfecho óbito e 1 regra com 45 casos de pacientes com desfecho cura.

Os algoritmos não apresentaram erros, ou classificações erradas, o que pode ser comprovado com os cálculos da matriz confusão e do erro médio. A partir dos dados da matriz confusão também é possível comprovar que os algoritmos tiveram uma taxa de acurácia, sensibilidade, abrangência, especificidade e precisão igual a 100%, já o erro foi igual a 0%, resultados que demonstram que os algoritmos obtiveram sucesso e resultados satisfatórios.

A taxa de confiança para as regras geradas pelo algoritmo C4.5 foi de 99,01% para o desfecho óbito e 98,88% para o desfecho cura, e de 76,24% para o desfecho óbito e 98,88% para o desfecho cura no ID3, o que comprova que os resultados do algoritmo são confiáveis e que as regras geradas não se especializaram com detalhes da base, evitando assim o *orverfitting*. O algoritmo ID3 gerou mais regras se comparado ao C4.5, isso acontece porque o ID3 não possui o cálculo da poda da árvore. A taxa de confiança mostra quais regras são mais confiáveis, por possuírem mais casos iguais, o algoritmo ID3 apresentou *overfitting* em casos de regras com desfecho óbito, este fato não caracteriza um erro do algoritmo pois é uma peculiaridade do próprio ID3.

Os resultados da validação cruzada e do cálculo do erro médio também foram muito satisfatórios, já que os algoritmos não apresentaram erros. Desta forma, pode-se afirmar que a tarefa de classificação pelos algoritmos C4.5 e ID3 da *Shell* Orion *Data Mining Engine* é uma ferramenta confiável na busca do conhecimento em bases de dados, apresentando resultados corretos e confiáveis. A Figura 2 apresenta um gráfico comparativo entre os algoritmos C4.5 e ID3.

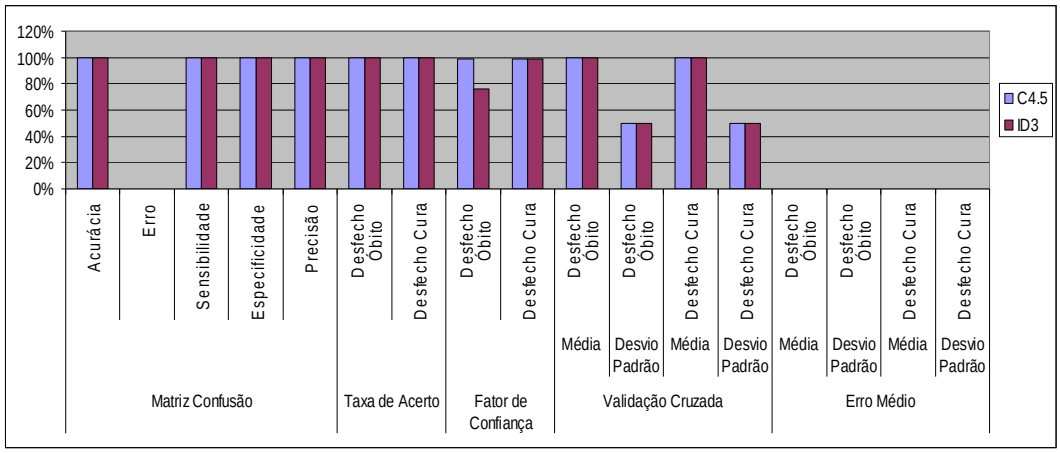


Figura 2. Gráfico comparativo da avaliação dos algoritmos C4.5 e ID3

Conforme entrevista com o especialista da área, o algoritmo C4.5 gera regras de alto interesse, e de fácil interpretação. A entrevistada também classificou o conhecimento gerado pela *Shell* Orion em ótimo quanto a eficiência na aquisição dos resultados, ótimo quanto ao conhecimento gerado corresponder as expectativas, ótimo também na compreensão do conhecimento gerado e quanto ao auxilio na confirmação de conhecimento pré-existente. Classificando como bons a contribuição com conhecimento novo e a contribuição de conhecimento útil a sua atividade.

Quanto a usabilidade no algoritmo C4.5 a entrevistada classificou a *Shell* Orion como Bom em todos os requisitos: facilidade em aprender a manipular a *Shell*; facilidade de utilização e o ambiente da aplicação.

A especialista classificou quanto ao resultado do ID3, cinco regras como de alto interesse, cinco como médio interesse e seis regras como de baixo interesse. As regras de alto interesse foram as que continham mais casos relacionados, assim como as de médio interesse foram aquelas com uma quantidade de casos que variou de três a cinco casos, e as regras que continham apenas um caso foram classificadas como de baixo interesse.

A usabilidade no ID3 obteve resposta “ótima” para a fácil compreensão do conhecimento. E resposta “bom” quanto a eficiência na obtenção dos resultados, quanto ao conhecimento gerado corresponder as expectativas, ao acréscimo de conhecimento útil, confirmação de conhecimentos pré existentes e a contribuição de conhecimentos novos. Em contrapartida a usabilidade foi classificada como “ótimo” em todos os requisitos, facilidade em aprender a manipular a *Shell*, facilidade de utilização e o ambiente da aplicação.

A partir da entrevista com a especialista, realizada por meio de questionário, foi possível representar através de uma tabela comparativa (Tabela 5) os resultados obtidos com as respostas do questionário, e ilustrados na Figura 3.

Tabela 5. Tabela comparativa dos resultados obtidos através da entrevista

	Interesse na Regra						Funcionalidade						Usabilidade					
	Alto		Médio		Baixo		Ótimo		Bom		Ruim		Ótimo		Bom		Ruim	
	%	Contagem	%	Contagem	%	Contagem	%	Contagem	%	Contagem	%	Contagem	%	Contagem	%	Contagem	%	Contagem
C4.5	100%	2	-	-	-	-	66,66%	4	33,33%	2	-	-	-	-	100%	3	-	-
ID3	31,25%	5	31,25%	5	37,50%	6	16,67%	1	83,33%	5	-	-	100%	3	-	-	-	-

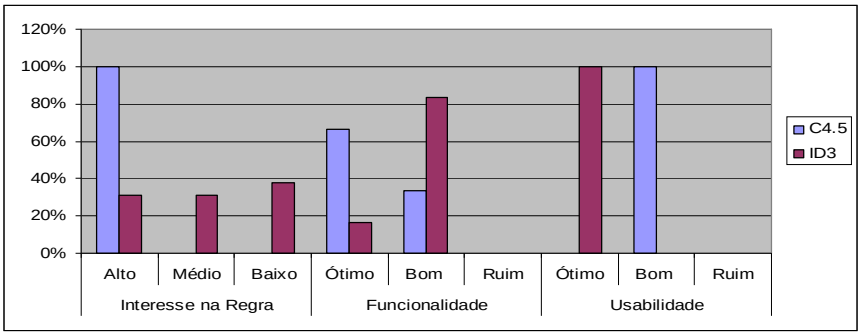


Figura 3. Gráfico comparativo do questionário a respeito dos algoritmos C4.5 e ID3

Um ponto muito importante na validação com o especialista da área é a isenção de respostas ruins ou de regras com baixo interesse, o que torna-se importante para contribuição de confiabilidade na aplicação.

10. Conclusão

A tarefa de classificação na *Shell* Orion *Data Mining Engine* conta com os algoritmos C4.5 e ID3, para o *data mining*. Esta ferramenta demonstrou ótimos resultados para os algoritmos ID3 e C4.5 na base de dados da sepse. Esta pesquisa veio fortalecer a confiança na ferramenta desenvolvida por alguns acadêmicos do curso de Ciência da Computação e mantida pelo Grupo de Pesquisa em Inteligência Computacional, da UNESC.

Os módulos utilizados obtiveram êxito quanto a sua exatidão e operacionalidade na realização dos testes. Os algoritmos ID3 e C4.5 na tarefa de classificação do *Data Mining* na base de dados da área médica com informações de pacientes com sepse, obtiveram resultados muito satisfatórios e puderam assim contribuir para a evolução da ferramenta quanto a sua confiabilidade.

O objetivo deste trabalho foi alcançado com a realização de estudos sobre *data* m*ining* e os métodos de avaliação de desempenho, com foco nos métodos específicos para a tarefa de classificação. Na realização dos testes foi possível perceber como é construído um classificador, além da compreensão de como funcionam ferramentas que envolvem descoberta de conhecimento em bases de dados.

Na avaliação do desempenho realizada, pode-se constatar que os algoritmos C4.5 e ID3 não apresentaram erros nas classificações, as taxas de sensibilidade e abrangência reforçaram a confiabilidade nos algoritmos, assim como a entrevista realizada com a especialista da área da base de dados aplicada, obteve bons resultados.

Considerando o tema abordado nesta pesquisa e a necessidade da continuidade do desenvolvimento da *Shell* Orion *Data Mining Engine*. Ficam como sugestões de trabalhos futuros a avaliação do desempenho do algoritmo CART da tarefa de classificação, avaliação de outras tarefas e algoritmos da *Shell* Orion, bem como a aplicação de outras métricas para a avaliação do desempenho em aplicações de *data mining*.

Referências

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel**.** ***Data Mining*****: um guia prático**. Rio de Janeiro: Elsevier, 2005.

CECIL, Russell L.; GOLDMAN, Lee; AUSIELLO, Dennis A. **Cecil: tratado de medicina interna.** Rio de Janero: Elsevier, 2005.

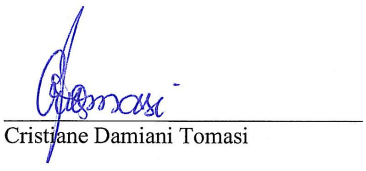
SOMMERVILLE, Ian. **Engenharia de software.** São Paulo: Addison-Wesley, 2003.

ANEXO A – Autorização para Publicação de Pesquisa de Opinião

AUTORIZAÇÃO PARA PUBLICAÇÃO DE PESQUISA DE OPINIÃO

Eu, Cristiane Damiani Tomasi, RG 49215914, residente à Rua Luiz Martinello, 285, na cidade de Criciúma/SC, aluna de Doutorado do Programa de Pós-Graduação em Ciências da Saúde da Universidade do Extremo Sul Catarinense (UNESC), AUTORIZO Maiara Costa Motta, RG 4.122.705, residente na Rua José Sumara de Souza, 10, Criciúma/SC, acadêmica do curso de Ciência da Computação da UNESC, a anexar em seu Trabalho de Conclusão de Curso, intitulado Avaliação de Desempenho dos Algoritmos para Classificação em Data Mining, ID3 e C4.5, na Shell Orion, o questionário para validação do conhecimento por mim respondido.

Criciúma, 17 de junho de 2010.



Cristiane Damiani Tomasi