H.pylori-MIND*Sys*: Um Protótipo de Sistema Baseado em Conhecimento para Auxílio na Predição da Existência da Bactéria *Helicobacter pylori* em Doenças Pépticas

Daniel de Faveri Honorato[1], Huei Diana Lee[1], Renato Bobsin Machado[1], Claudio Saddy Rodrigues Coy[2], João José Fagundes[2], and Feng Chung Wu[1,2]

[1] Laboratório de Bioinformática (LABI)
Universidade Estadual do Oeste do Paraná (UNIOESTE)
Parque Tecnológico Itaipu
Foz do Iguaçu, Paraná, Brasil CEP 85870-650
[2] Serviço de Coloproctologia da Faculdade de Ciências Médicas
Universidade Estadual de Campinas (UNICAMP)
Campinas, São Paulo, Brasil CEP 13084-970
{dfaverih,hueidianalee}@gmail.com

Abstract. Research related to peptic diseases has raised interest especially because of their high incidence in the population. This work presents a system prototype, named *H.pylori*-MIND*Sys*, to predict the existence of the *Helicobacter pylori* bacteria. The Data Mining process was employed to help finding patterns from a real Upper Digestive Endoscopy database. The extracted knowledge was then evaluated and validated by domain specialists and used in the construction of the knowledge base of the developed system.

Key words: data mining, knowledge-based system, bioinformatics

1 Introdução

Com o avanço tecnológico, conceitos e técnicas de Inteligência Artificial (IA) têm sido cada vez mais aplicados na solução de problemas reais por meio de sistemas de IA, tais como os relacionados aos Sistemas Baseados em Conhecimento. Esses sistemas podem ser construídos para qualquer área do conhecimento e utilizam conhecimento representado explicitamente para resolver problemas. Uma das áreas na qual podem ser utilizadas técnicas de IA para auxiliar na tomada de decisão é a área médica. Nessa área, pesquisas relacionadas a doenças pépticas têm despertado grande interesse, principalmente devido ao seu alto índice de incidência na população [7]. Uma das causas relacionadas a essas doenças é a presença da bactéria *Helicobacter pylori* (HP). Essa bactéria danifica células da parede gástrica, tornando-a mais suscetível à ação do ácido clorídrico e da pepsina do suco gástrico e causando lesões conhecidas como gastrites e úlceras. Existem várias maneiras de detectar a bactéria HP, tais como teste de urease,

anatomia patológica (AP), exame de sangue e análise do ar expirado [13]. Os dois primeiros são realizados por meio da biópsia do tecido extraído no exame de Endoscopia Digestiva Alta (EDA). Os dois últimos, embora menos invasivos, apresentam custo maior.

Este trabalho faz parte do projeto de Análise Inteligente de Dados desenvolvido conjuntamente por pesquisadores do Laboratório de Bioinformática (LABI[3]) da Universidade Estadual do Oeste do Paraná, do Serviço de Coloproctologia da Faculdade de Ciências Médicas da Universidade Estadual de Campinas, e do Hospital Municipal de Paulínia (HMP), e tem por objetivo desenvolver um protótipo de Sistema Baseado em Conhecimento (SBC), o qual poderá auxiliar médicos na predição da existência ou não da bactéria HP, em doenças pépticas gastroduodenais. O sistema poderá auxiliar também na diminuição do custo da detecção da existência da HP e prover suporte a médicos durante a realização do exame de EDA, alertando-os sobre a probabilidade de existência dessa bactéria e sobre outras características importantes do paciente que está sendo examinado.

O conhecimento utilizado no SBC foi adquirido por meio do processo de Mineração de Dados, o qual foi extraído de uma base de dados real de Endoscopia Digestiva Alta e resultados de teste para a detecção da bactéria HP. Esse conhecimento extraído de maneira automática foi avaliado e validado por especialistas do domínio e inserido no SBC.

Este trabalho está organizado da seguinte maneira: na Seção 2 são apresentados conceitos relacionados a Sistemas Baseados em Conhecimento. Na Seção 3 são abordados conceitos relacionados ao processo de Mineração de Dados. Na Seção 4 é apresentada a metodologia utilizada e na Seção 5 são apresentados os resultados e discussões. Ao final, na Seção 6 são apresentadas as considerações finais.

2 Sistemas Baseados em Conhecimento

Sistemas Baseados em Conhecimento são programas de computador que usam conhecimento, representado explicitamente, para resolver problemas. Esses sistemas, manipulam conhecimento e informação de maneira inteligente e são usados em problemas que requerem uma grande quantidade de conhecimento especializado [10]. Um SBC possui quatro componentes principais: a Interface com o Usuário, a Base de Conhecimento (BC), o Motor de Inferência e a Memória de Trabalho. A interação entre esses componentes é apresentada na Figura 1.

A Interface é responsável pela interação entre o usuário e o SBC, a qual pode ser realizada por meio de um interpretador de comando ou por outros mecanismos mais amigáveis, por exemplo, o uso de janelas, menus, gráficos, animações e cores.

A Base de Conhecimento contém a descrição do conhecimento necessário para a resolução do problema abordado na aplicação. Desse modo, é necessário que o conhecimento esteja organizado de maneira adequada para que o Motor

[3] www.foz.unioeste.br/labi

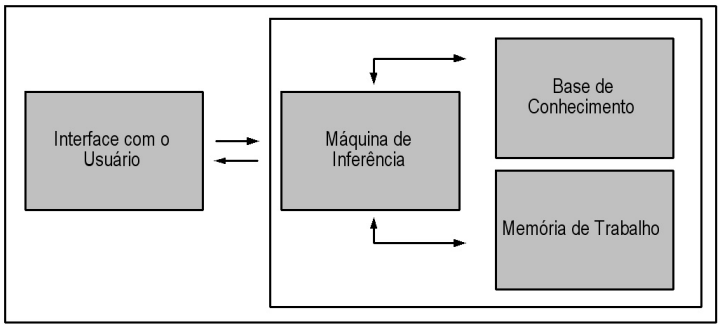

Figura 1. Estrutura de um Sistema Baseado em Conhecimento.

de Inferência consiga tratá-lo convenientemente. Cada representação na Base de Conhecimento é chamada de sentença, a qual é expressa em linguagem para a Representação de Conhecimento (RC), tais como lógica, regras de decisão, redes semânticas, frames, orientação a objetos e orientação a objetos associada a regras [10].

Dentre as linguagens para RC existentes, uma das mais utilizadas é a de regras de decisão. Esse modo de representação é popular, pois as regras apresentam natureza modular, facilidade de explicação e são similares ao processo cognitivo humano [3]. As regras são estruturadas como "SE (condições) ENTÃO (conclusão)", nas quais a parte SE é uma lista de condições a serem satisfeitas e a parte ENTÃO é a conclusão.

Usualmente, o conhecimento utilizado para a construção da BC dos SBCs pode ser adquirido explicitamente por meio de entrevistas com especialistas da área de domínio, pesquisas bibliográficas da área em questão, entre outros [10].

O Motor de Inferência ou Máquina de Inferência é responsável pelo desenvolvimento do raciocínio baseado nas informações fornecidas pelo usuário e no conhecimento representado na BC. A principal característica do Motor de Inferência diz respeito ao modo de raciocínio a ser utilizado. Existem vários tipos de raciocínio que podem ser aplicados em Sistemas Baseados em Conhecimento. Em um sistema de regras de decisão, são aplicáveis, basicamente, dois modos de raciocínio: encadeamento progressivo (*forward chaining*) e encadeamento regressivo (*backward chaining*) [3].

A área de trabalho do SBC, na qual são registradas todas as respostas fornecidas pelo usuário durante as interações realizadas com o sistema, é representada pela Memória de Trabalho. Algumas vantagens da utilização desse componente são: evitar que o usuário responda duas vezes à mesma pergunta, permitir ao usuário ver toda a linha de raciocínio que foi usada para chegar a uma determinada conclusão, e evitar que sejam realizadas repetidas seqüências de raciocínio para obtenção de conclusões intermediárias [10].

3 Mineração de Dados

Com o avanço tecnológico, a quantidade de informações armazenadas digitalmente está cada vez maior. Para que possam ser realizadas análises mais completas, é necessário que essas informações sejam representadas de maneira apropriada, processadas e que um modelo que represente o conhecimento embutido nesses dados seja construído, uma vez que a análise manual é inviável. Um dos processos utilizados para realizar a análise de dados é a Mineração de Dados (MD) cujo objetivo é identificar padrões válidos, novos, potencialmente úteis e compreensíveis embutidos em dados [4].

O processo de MD é interativo e iterativo. Ele é composto, basicamente, por três etapas: pré-processamento, extração de padrões e pós-processamento. O pré-processamento é, freqüentemente, a etapa mais custosa, consumindo em torno de 80% do tempo usado para realizar o processo [8]. Ele tem como objetivo realizar tarefas como preparação, redução e transformação dos dados. Ainda em pré-processamento, é necessário que os dados estejam representados no formato apropriado para a próxima etapa, sendo um dos formatos mais comumente utilizados o atributo-valor. Um dos principais problemas na etapa de pré-processamento está relacionado à Seleção de Atributos (SA) importantes que serão utilizados na extração de conhecimento. A SA permite selecionar um sub-conjunto de atributos do conjunto de dados original por meio da remoção de atributos irrelevantes ou redundantes. A maioria dos algoritmos de MD não trabalha bem com um número grande de atributos, assim, a SA pode melhorar a precisão dos classificadores gerados. Ainda, com um número menor de atributos, modelos mais compreensíveis podem ser construídos [5].

A etapa de extração de padrões tem como característica a configuração, escolha e execução de um ou mais algoritmos de extração de padrões sobre os dados selecionados na etapa de pré-processamento. Essa etapa é realizada de maneira iterativa, sendo necessário realizar diversos ajustes nos parâmetros dos algoritmos de extração de padrões utilizados, com o objetivo de construir modelos do conhecimento extraído dos dados pré-processados.

Após a extração de padrões, inicia-se a etapa de pós-processamento, na qual os modelos construídos são avaliados e validados. Depois de concluído o processo, o conhecimento extraído é disponibilizado ao usuário, o qual pode ser utilizado para auxiliar no processo de tomada de decisões. Neste trabalho, o conhecimento extraído é utilizado na construção da BC do SBC.

4 Metodologia Utilizada para a Construção do Sistema

Conforme mencionado, a construção de um SBC implica, geralmente, na aquisição manual (explícita) de conhecimento por meio de interações com um ou mais especialistas da área. Neste trabalho foi utilizada a aquisição automática de conhecimento, a partir de uma base de dados reais, associada a interações com especialistas do domínio. A base de dados utilizada neste trabalho para a construção da BC foi coletada a partir de laudos, em formato texto, de Endoscopia

Digestiva Alta realizados no período de 01/1999 a 02/2000 no Serviço de Endoscopia Digestiva Alta do Hospital Municipal de Paulínia. Nesse período foram realizados 1271 exames, os quais continham informações relacionadas ao esôfago, estômago, duodeno e anatomia patológica. As informações sobre o esôfago, o estômago e o duodeno dos pacientes são obtidas no ato do exame de endoscopia. As informações relacionadas às características histológicas e a existência ou não da bactéria são obtidas pela análise do tecido extraído do paciente por meio da AP.

Para o desenvolvimento deste trabalho, somente exames de pacientes submetidos à biópsia e posterior teste para verificação da existência de HP poderiam ser considerados para a construção do SBC. Assim, do total de 1271 laudos, foram selecionados 276 casos que atendiam a esse requisito. Esses exames foram mapeados para um conjunto de dados inicial digital no formato atributo-valor descritos por 85 atributos. Nesse formato cada coluna representa um atributo e cada linha representa um exame. Informações de identificação dos pacientes foram removidas. A partir dessa base, foram construídos dois conjuntos de dados, apenas variando o número de atributos. O primeiro conjunto de dados (CD1) foi criado utilizando todos os atributos da base de dados, isto é, os 85 atributos iniciais. O segundo conjunto de dados (CD2) foi construído não considerando os atributos de anatomia patológica, resultando, portanto, em um conjunto de dados com 63 atributos. Nesse último conjunto de dados, uma das razões para a exclusão de todos atributos de anatomia patológica está relacionada à tentativa de predizer a exitência da bactéria *Helicobacter pylori* considerando somente informações de esôfago, estômago e duodeno.

Duas ferramentas principais foram empregadas neste trabalho: uma ferramenta de MD para extração de conhecimento dos conjuntos de dados e uma *Shell*[4] de SBC para a construção do protótipo. O conhecimento extraído, após analisado e validado pelos especialistas da área, foi inserido na Base de Conhecimento do SBC. Dentre as ferramentas de MD existentes, foi selecionada a *See5* [12], a qual implementa a versão posterior dos consagrados algoritmos C4.5 e C4.5-*rules* [9] que permite a indução de árvores de decisão e regras de decisão, respectivamente. Para a construção do SBC, foi utilizada a ferramenta CLIPS [1], a qual provê suporte com Base de Conhecimento e Motor de Inferência. Essa ferramenta foi escolhida por ser gratuita, possuir boa documentação, além de ter como um dos modos de Representação de Conhecimento, as regras de produção, as quais foram utilizadas neste trabalho.

A linguagem *JAVA*[5] foi utilizada para a construção da interface do sistema. Essa linguagem foi selecionada, pois possui um conjunto de *APIs* (*Application Programming Interfaces*) que disponibilizam diversas funcionalidades e objetos reutilizáveis.

[4] *Shells* são ferramentas que auxiliam na construção de SBCs, geralmente fornecendo suporte para a Base de Conhecimento, o Motor de Inferência e a Memória de Trabalho.

[5] http://www.java.sun.com.

5 Resultados e Discussões

Conforme mencionado, o conhecimento a ser utilizado no SBC foi extraído utilizando algoritmos de MD. Portanto, a partir dos dois conjuntos de dados CD1 e CD2 construídos foram realizadas duas iterações aplicando o algoritmo de indução de regras. Os resultados dessas iterações são apresentados na Tabela 1, a qual está organizada da seguinte maneira:

- ***Conjunto de Dados***: nome do conjunto de dados que contém os casos utilizados pelo algoritmo de MD;
- ***#Atr***: número de atributos de cada conjunto de dados;
- ***#Regras***: número de regras do modelo completo induzido pelo algoritmo de MD;
- ***EV*** $\pm$ ***EP***: estimativa de erro verdadeiro obtido por meio de 10-*Fold Cross-Validation* (CV[6]) e erro padrão dessa média.

Tabela 1. Resultados da Aplicação do Algoritmo de MD.

Conjunto de Dados	*#Atr*	*#Regras*	*EV* $\pm$ *EP (%)*
CD1	85	15	27,90 $\pm$ 3,40
CD2	63	4	34,10 $\pm$ 1,50
CD1Relief	49	13	27,50 $\pm$ 2,10
CD2Relief	37	4	36,60 $\pm$ 1,90

Na primeira iteração foi aplicado o algoritmo de indução de regras nos conjuntos de dados CD1 e CD2 . Embora a precisão dos classificadores gerados a partir desses conjuntos de dados seja similar a da Classe Majoritária (CM) (36,59%), é importante ressaltar que esse erro é obtido de acordo com número de casos de cada classe presente no conjunto de dados e desse modo, a resposta fornecida com base na precisão da CM é sempre a classe mais freqüente, não sendo baseada nos valores de atributos específicos de cada caso. Assim, dado um novo caso ao sistema, não seria possível fornecer uma explicação sobre a decisão tomada em relação a esse caso se a predição fosse realizada baseada na CM. Contrariamente, uma classificação baseada em um modelo, como o construído por meio de algoritmos de MD que geram árvores e regras de decisão, poderia fornecer uma explicação da resposta de classificação dado um novo caso.

Conforme citado na Seção 3, um dos problemas quando se realiza a extração de conhecimento de uma base de dados está relacionado ao número de atributos

[6] Os dados são selecionados aleatoriamente e divididos em k partições mutuamente exclusivas (*folds*) de exemplos, aproximadamente do mesmo tamanho. O indutor de regras é treinado e testado k vezes. Na primeira vez, o primeiro *fold* é usado para teste e os *k-1 folds* restantes, para treinamento. Na segunda vez, o segundo *fold* é usado para teste e os *k-1 folds* restantes, para treinamento, e assim sucessivamente até o *k-ésimo fold*.

contidos no conjunto de dados. Esse é um dos fatores que influencia no desempenho das predições, pois a maioria dos algoritmos de MD não trabalha bem na presença de um número grande de atributos. Outra questão importante é a compreensibilidade dos modelos construídos, a qual está também relacionada ao número de atributos que são utilizados para construir o modelo.

Desse modo, foi realizada a Seleção de Atributos, utilizando o algoritmo *ReliefF* [11], freqüentemente utilizado para SA, com o objetivo de obter um número menor e relevante de atributos dos conjuntos de dados. A idéia desse algoritmo é estimar a qualidade dos atributos de acordo com quão bem seus valores distinguem entre exemplos da mesma classe e de diferentes classes que estão perto uma da outra. Ao final do processo de SA, o *ReliefF* fornece uma lista dos atributos classificados de acordo com sua relevância em relação à classe. O algoritmo de Seleção de Atributos foi aplicado para ambos os conjuntos CD1 e CD2. A aplicação do *ReliefF* nos conjuntos resultou em um novo conjunto de dados (CD1Relief) com 49 atributos para CD1 e em outro conjunto de dados (CD2Relief) com 37 atributos para CD2.

O próximo passo após a SA foi realizar a aplicação do algoritmo de indução de regras sobre os conjuntos de dados CD1Relief e CD2Relief. Conforme é possível observar na Tabela 1, os resultados para os conjuntos de dados CD1Relief e CD2Relief foram semelhantes aos alcançados na primeira iteração. Portanto, para verificar se houve diferença significativa entre as médias dos erros dos classificadores, construídos com os conjuntos de dados antes e depois da SA (entre CD1 e CD1Relief e entre CD2 e CD2Relief), foi utilizado um teste t não pareado [2].

O resultado do teste t mostrou que não houve diferença estatisticamente significativa entre as médias dos erros dos classificadores construídos a partir de CD1 e CD1Relief (p = 0,9194) e entre CD2 e CD2Relief (p = 0,3177). Desse modo, utilizando apenas os 49 atributos de CD1Relief dos 85 atributos iniciais de CD1, é possível obter um resultado que apresenta, estatisticamente, a mesma precisão. Do mesmo modo, com apenas 37 atributos de CD2Relief dos 63 atributos iniciais de CD2, é possível também obter precisões semelhantes nos resultados. Os conjuntos de atributos selecionados pelo processo de SA, ou seja, representados pelos atributos dos conjuntos de dados CD1Relief e CD2Relief, foram analisados pelos especialistas do domínio. É interessante notar que 35% do total de atributos considerados relevantes pelo método de SA também foram considerados importantes segundo a análise dos especialistas. As regras geradas a partir dos conjuntos de dados CD1Relief e CD2Relief também foram avaliadas e validadas pelos especialistas, tendo sido consideradas relevantes de acordo com o conhecimento do domínio. Desse modo, para a construção do protótipo do sistema *H.pylori*-MIND*Sys*, foram consideradas as regras obtidas a partir do conjunto de dados CD1Relief por esse apresentar melhores resultados nos experimentos realizados.

Após a extração, avaliação e validação do conhecimento, foi realizado o projeto da interface do *H.pylori*-MIND*Sys*. Para a construção da interface levou-se em consideração os atributos utilizados nas regras que foram extraídas pelo al-

goritmo de MD. A interface construída é apresentada na Figura 2. A partir dessa interface, o usuário pode fornecer os valores para os atributos, os quais são utilizados na execução do SBC. O resultado e o grau de confiança da regra ou regras, que permitiram essa conclusão, são apresentados na parte inferior da interface. Também, a interface possibilita ao usuário obter a explicação de como o sistema inferiu uma determinada conclusão.

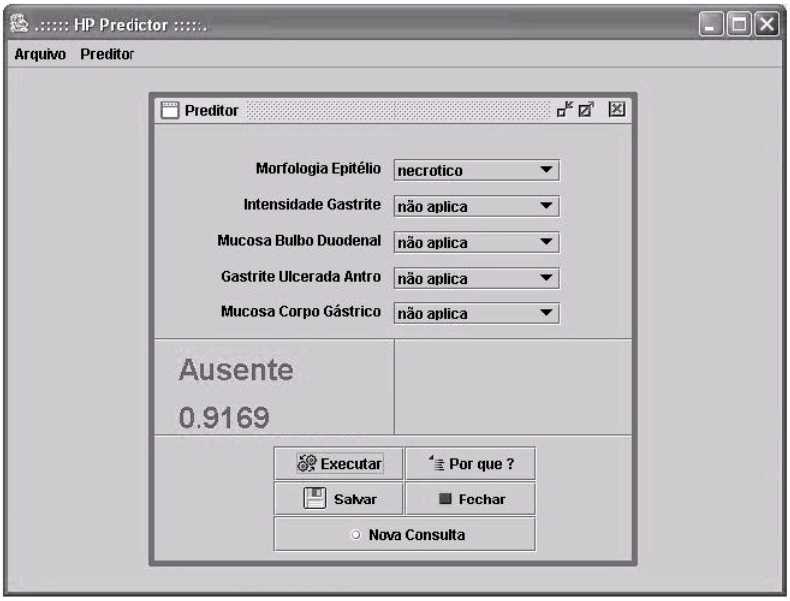

Figura 2. Interface do Sistema *H.pylori*-MIND*Sys*.

Após a validação da interface pelo especialista, o próximo passo foi realizar a codificação da BC do sistema utilizando regras extraídas do conjunto de dados CD1Relief. O conjunto de regras gerado pelo algortimo de indução de regras do *See5* foi convertido para o formato de regras utilizado pela BC do CLIPS.

Na interface do *H.pylori*-MIND*Sys* é necessário o usuário inicialmente preencher informações do paciente. O SBC então realiza o processamento dessas informações e o resultado é apresentado. O usuário pode visualizar as regras que foram disparadas pelo sistema e os resultados podem ser armazenados para consultas posteriores. Neste trabalho foi definido que, havendo conflito[7], a resposta do sistema é fornecida considerando a classe com maior número de regras disparadas. Nesse caso, é fornecido o resultado da Equação 1, a qual utiliza os dois maiores fatores de certeza das duas regras da mesma classe que mais aparecem.

[7] Dependendo das informações fornecidas ao SBC, mais de uma regra da BC, contendo eventualmente respostas distintas, pode ser disparada.

Essa equação foi utilizada em um dos primeiros SBCs, o sistema MyCin [6]. Na Figura 3 é ilustrado um exemplo de conflito.

$$CF3 = (CF1 + CF2) - (CF1 * CF2) \quad (1)$$

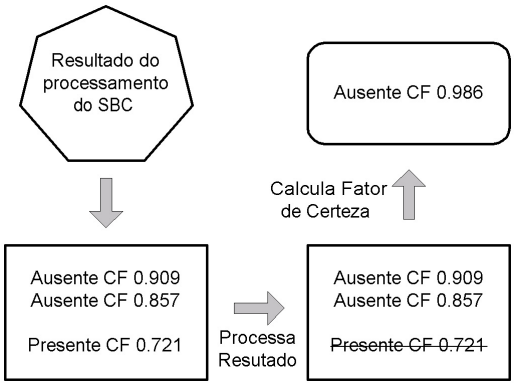

Figura 3. Resolução de Conflito.

Nesse exemplo, duas regras da classe Ausente e uma da classe Presente foram disparadas. Caso ocorra empate no número de regras disparadas para cada classe, é fornecido ao usuário o resultado para as duas classes.

6 Considerações Finais

Neste trabalho foram apresentados o projeto e o desenvolvimento de um protótipo de Sistema Baseado em Conhecimento denominado *H.pylori*-MIND*Sys* para auxílio na predição da existência bactéria *Helicobacter pylori*. Especialistas do domínio consideraram bons os resultados alcançados, tendo sido o conhecimento extraído classificado como relevante de acordo com o conhecimento de domínio. É importante ressaltar que a extração automática de padrões embutidos nos dados pode auxiliar bastante na redução de tempo de construção de SBCs. Outra vantagem é que podem ser identificados padrões que são importantes para a predição da bactéria HP e que podem não ser adquiridos em um processo de aquisição manual. Por fim, os especialistas podem complementar a BC, inserindo regras que não foram identificadas na aquisição automática.

Trabalhos futuros incluem a aplicação do algoritmo de MD sobre outros conjuntos de dados de EDA e com isso, novas regras serem identificadas e avaliadas por especialistas do domínio para que a base de conhecimento do sistema seja aumentada. Outro trabalho inclui a ampliação do sistema para que possa auxiliar o especialista durante a EDA, ou seja, a medida que o exame está sendo realizado e os dados são fornecidos ao sistema, esse, de modo interativo, exibe o

conhecimento embutido em sua BC relacionado com as características correntes do exame. A metodologia proposta também será aplicada a outros domínios.

Agradecimentos. Os autores agradecem Antonio Pietrobom Neto, do Hospital Municipal de Paulínia, por disponibilizar o conjunto de laudos utilizado nos experimentos. Também agradecem ao Professor Juvenal Ricardo Navarro Góes (*in memorian*) pela sua inestimável contribuição. Este trabalho também teve o auxílio do Conselho Nacional de Desenvolvimento Científico e Tecnológico — CNPq.

Referências

1. CLIPS Reference Manual: Basic Programming Guide, http://clipsrules.sourceforge.net/documentation/v630/bpg.pdf (2008)
2. Freedman, D., Pisani, R., and Purves, R.: Statistics. Norton, New York, USA (1998)
3. Giarratano, J. and Riley, G.: Expert Systems: Principles and Programming. PWS Publishing Company, Boston, USA (1998)
4. Han J and Kamber M.: Data Mining: Concepts and Techniques. Morgan Kauf-mann Publishers, California, USA (2006)
5. Lee, H. D.: Seleção de atributos importantes para a extração de conhecimento de bases de dados. Tese de Doutorado, ICMC-USP, `http://www.teses.usp.br/teses/disponiveis/55/55134/tde-22022006-172219/` (2005)
6. Mahoney, J. J. and Mooney, R. J.: Comparing Methods for Refining Certainty-Factor Rule-Bases. The 11th International Conference on Machine Learning, Austin, Texas (1994)
7. Pellicano, R., Fagoonee, S., Palestro, G., Rizzetto, M., Figura, N., Ponzetto, A.: The diagnosis of helicobacter pylori infection: guidelines from the maastricht 2-2000 consensus report. Minerva Gastroenterol Dietol, vol. 50(2), pp. 125—33 (2004)
8. Pyle, D.: Data Preparation for Data Mining. Morgan Kaufmann, California, USA (1999)
9. Quinlan, J. R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, California, USA (1993)
10. Rezende, S. O.: Sistemas Inteligentes: Fundamentos e Aplicações. Editora Manole, Barueri, SP, Brasil (2003)
11. Robnik-Sikonja, M. and Kononenko, I.: Theoretical and Empirical Analysis of ReliefF and RReliefF. Machine Learning Journal, vol. 53, pp. 23 — 69 (2003)
12. Rulequest-Research. See5: An Informal Tutorial, http://www.rulequest.com/see5-win.html (2008)
13. Versalovic, J.: Helicobacter pylori: Pathology and Diagnostic Strategies. American Journal of Clinical Pathology. vol. 119, pp. 403 — 412 (2003)