1ª JORNADA CIENTÍFICA DA UNIOESTE – 24 a 26/10/2001 – CASCAVEL/PR

APLICAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA PARA EXTRAÇÃO DE CONHECIMENTO E CONSTRUÇÃO DE CLASSIFICADORES: ESTUDO DE CASO DE BASES DE DADOS MÉDICAS

Ferro, M; Lee, H.D; Chung, W.F.

Universidade Estadual do Oeste do Paraná/Foz do Iguaçu/Centro de Engenharia e Ciências Exatas, Ciência da Computação/mariza98, huei@dcc.unioeste-foz.br; Universidade Estadual de Campinas/Faculdade de Ciências Médicas/Departamento de Cirurgia

PALAVRAS CHAVE: Aprendizado de máquina, extração de conhecimento, bases de dados médicas

ÁREA DO CONHECIMENTO: Ciências Exatas e da Terra

RESUMO: Neste trabalho mostramos como as técnicas de Aprendizado de Máquina Simbólico Indutivo Supervisionado podem ser aplicados sobre conjuntos de dados médicos, com o objetivo de se extrair conhecimento e construir classificadores/preditores. Serão apresentados aqui os métodos de Aprendizado de Máquina para se alcançar esses objetivos, as ferramentas aplicadas e como realizamos os estudos de caso sobre bases de dados naturais, do domínio médico e os resultados que obtivemos com nossas aplicações.

INTRODUÇÃO: A aplicação de Técnicas de Aprendizado de Máquina – AM – a conjuntos de dados pode ser vista sob dois aspectos: a extração de conhecimentos e a construção de classificadores/preditores. Neste trabalho é enfocada a extração de conhecimento de bases de dados. São descritos dois estudos de caso, realizados sobre conjuntos de dados naturais encontrados no repositório de dados UCI (Debra, 00). Ambas as bases de dados são estudadas utilizando métodos de AM simbólico para a extração de conhecimento. Essas bases de dados contém dados de exames médicos cardiológicos e a classificação do número de artérias reduzidas um paciente possui.

MATERIAL E MÉTODOS: A metodologia aplicada para a pesquisa seguiu por quatro fases distintas: Coleta dos dados, preparação dos dados, modelagem dos dados e análise dos resultados. A primeira fase engloba a coleta dos conjuntos de dados do repositório de dados UCI (Debra, 00). O conjunto de dados utilizado é o Heart Disease, que está subdividido em outras duas, de acordo com a cidade do centro de pesquisa que coletou os dados: processed.switzerland com 123 casos e processed.hungarian com 294 casos. A segunda fase é a preparação destes dados para que se possa aplicar as técnicas de Aprendizado de Máquina. A terceira fase consiste da aplicação das técnicas de AM e finalmente, a análise dos resultados obtidos com a aplicação das técnicas, que constitui a quarta fase. A fase de coleta dos dados pode ser retomada à medida que os dados são modelados ou os resultados são obtidos, sendo que na última fase a participação do especialista da área médica é indispensável. **1. Aprendizado de Máquina:** Para aplicar métodos de AM, tanto para a extração de conhecimento como a construção de classificadores é preciso primeiramente definir o paradigma e a estratégia que será aplicada. Enquanto o paradigma diz respeito à forma como o conhecimento adquirido será representado, a estratégia é a forma como o conhecimento é propriamente adquirido. Diversos paradigmas e estratégias de AM já foram propostos. Entre os paradigmas podemos citar: conexionista, genético, estatístico, instanced-based e simbólico (Monard et al., 97). No paradigma simbólico os sistemas de aprendizado buscam aprender construindo representações simbólicas de um conceito através da análise de exemplos e contra-exemplos desse conceito. As representações simbólicas mais comumente utilizadas são as árvores e regras de decisão. É atribuído a Morgan e Messeger (Morgan & Messeger, 73) a criação do programa para a indução de árvores de decisão. Os trabalhos com indução de regras de decisão iniciaram-se com a simples tradução das árvores de decisão para regras e evoluindo para técnicas que empregam, por exemplo, generalização (Quinlan 87a, Quinlan 87b). As estratégias de aprendizado são classificadas de acordo com o grau de complexidade, i.é, dificuldade de aprendizado do conceito por parte do aprendiz. Em qualquer processo de aprendizado, o

aprendiz usa o conhecimento acumulado para obter novo conhecimento. Este novo conhecimento é então lembrado para uso posterior. O aprendizado de um novo conceito pode ser realizado de várias maneiras (Monard et al., 97): aprendizado por hábito, instrução, dedução, analogia e a que exige maior complexidade de inferência, o Aprendizado por Indução; este é caracterizado pelo raciocínio que parte do específico para o geral. É a forma de inferência lógica que permite obter conclusões gerais a partir de exemplos particulares. É por aprendizado indutivo que nós humanos, a partir de um conjunto de observações chegamos a características particulares, ou que a partir de fatos chegamos a algumas generalizações. Por exemplo, se dissermos: todos os felinos observados tinham coração. Por indução: todos os felinos têm coração. Existem duas formas de aprendizado por indução: Observação e Descoberta, no qual o aprendiz analisa entidades fornecidas e tenta determinar se existe algum subconjunto que pode ser agrupado em certas classes de maneira útil (não existe um professor que tenha o conhecimento) e por Exemplos ou Supervisionado, no qual o aprendiz induz a descrição de um conceito formulando uma regra geral a partir de exemplos e contra-exemplos fornecidos pela base do conhecimento (por exemplo, um professor), por isso pode ser chamado de supervisionado. Neste trabalho enfocamos o AM Simbólico Indutivo Supervisionado. 2. **Algoritmo -** Nesta seção é descrito o algoritmo indutor de AM: C4.5 (Quinlan, 93) para árvores de decisão.Este algoritmo "aprende" conceitos representados por árvores de decisão. A escolha desse algoritmo foi influenciada, principalmente por três questões: a grande aceitação desse algoritmo entre a comunidade acadêmica, ser de uso livre e gerar resultados que podem ser mais facilmente interpretáveis por especialistas de outros domínios, neste caso específico os médicos. C4.5 - Os algoritmos para aprendizado através de árvores de decisão, geralmente usam uma heurística para estimar qual é o melhor atributo. O C4.5 utiliza um critério de decisão baseado na medida de entropia, i.e., grandeza que mede desordem tanto em objetos físicos como informações. Esta medida é um valor entre 0 e 1, onde 0 indica um conjunto uniforme (só um valor de classe está presente para todas as instâncias daquela característica) e 1 indica um conjunto onde a probabilidade é igual a todos os valores de classe presentes. O critério de divisão procura minimizar a entropia. No C4.5 cada nó folha assume a forma: <Ci> (N) ou <Ci> (N/E) onde Ci é o nome de uma classe, N é a soma fracionária das instâncias que alcançam aquele nó, e é o número de instâncias que pertencem a classes diferentes da classe Ci para a árvore, (Baranauskas e Monard, 00), porém foram classificadas como pertencentes a Ci. 3. **Estudos de Caso -** Serão descritos nessa sessão os estudos de caso realizados neste trabalho. O objetivo é extrair conhecimento por meio da aplicação de um método de Aprendizado de Máquina Simbólico Indutivo Supervisionado, como a indução de árvores de decisão sobre conjuntos de dados do repositório de dados UCI (Debra, 00). O conjunto de dados utilizado é o Heart Disease, que está subdividido em outras duas: processed.switzerland e processed.hungarian, de acordo com a cidade do centro de pesquisa que coletou os dados. O conjunto de dados refere-se ao diagnóstico de doenças coronárias, no qual estão presentes atributos referentes a exames médicos, tais como resultados de eletrocardiogramas, idade, pressão sangüínea, tipo de dor torácica (angina) entre outros. A classe prediz doenças coronárias em relação à obstrução de artérias, assumindo valores entre 0 – 5 de acordo com o número de artérias coronárias reduzidas em mais que 50%.

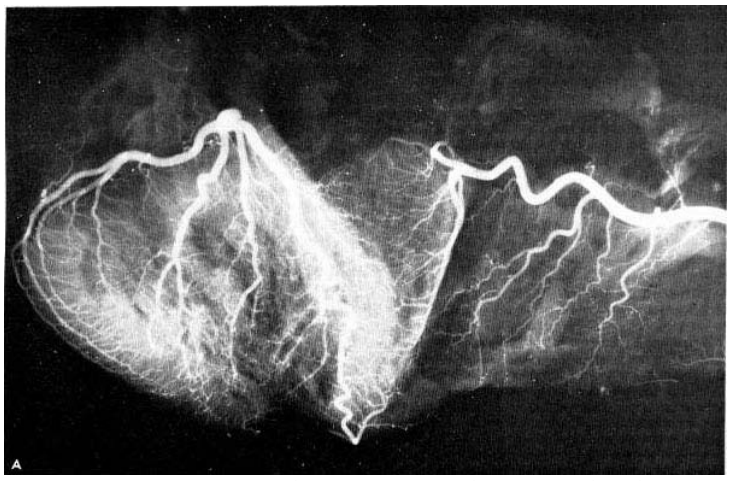

Figura A – Angiograma de um coração aberto, saudável.

A Figura A mostra um coração aberto com artérias coronárias preenchidas pela massa radiopaca. À direita está a artéria coronária principal (horizontal) com pequenos ramos descendentes. À esquerda vêem-se o ramo descendente principal esquerdo e o ramo circunflexo principal esquerdo (horizontal). Entre esses dois há diversos grandes ramos. O calibre dos vasos diminui progressivamente, sem obstruções ou estreitamentos. Na Figura B, angiograma em um coração aberto com coronárias arterioscleróticas gravemente estreitadas e ocluídas. A massa radiopaca não conseguiu penetrar em um grande segmento da artéria coronária direita. Podemos observar os estreitamentos e as tortuosidades do ramo descendente esquerdo e do ramo circunflexo esquerdo (Robbins et al., 96).

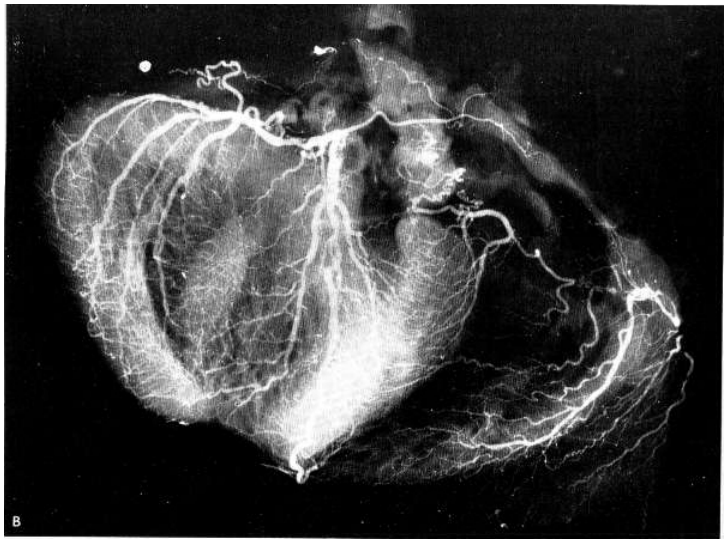

Figura B – Angiograma em um coração aberto com coronárias arterioscleróticas

Este estudo de caso envolveu quatro fases, sendo a primeira fase a coleta dos dados do repositório, seguida pela preparação dos dados, no qual os dados foram transformados para formatos aceitos pelo algoritmo C4.5 (Quinlam, 1993). A terceira fase foi a modelagem dos dados, que foi a aplicação do algoritmo C4.5. A última fase foi a análise dos resultados.

RESULTADOS E DISCUSSÃO: Nos experimentos realizados nesses estudos de caso foi utilizado o algoritmo C4.5, citado anteriormente. Os objetivos propostos foram: verificar se um paciente com algumas características, representadas pelos atributos, possuía artérias obstruídas predizendo doenças coronárias e extrair e avaliar novos conhecimentos gerados pela aplicação de métodos de AM. A análise foi realizada sobre os resultados gerados a partir da aplicação do algoritmo (árvores de decisão) nas bases processed.switzerland e processed.hungarian; nesta etapa é imprescindível a ajuda de um especialista do domínio. Foi feita a análise das árvores de decisão. O processo de análise foi feito a partir da técnica de duplo cego, no qual se descrevia atributos presentes na árvore e o especialista emitia sua opinião sobre qual classe aquela regra deveria se encaixar. O critério usado para a escolha das regras que seriam utilizadas no duplo cego foi olhar para a proporção de casos de erro em relação aos acertos das regras. A) Processed.Switzerland - Uma das regras analisadas apresentava 50% de erro em relação aos acertos, além de baixa qualidade ela também foi considerada ruim pelo especialista, sob a afirmação de que é impossível classificar um paciente apenas por ele apresentar uma angina atípica. Algumas outras regras foram analisadas, tendo 11,1%, 0% e 6,97% de erros, apesar das porcentagens de erros serem baixas o especialista considerou todas elas de má qualidade e não puderam ser classificadas através do duplo cego por apresentar poucas características para se realizar um diagnóstico. O erro considerando-se a classe majoritária foi de 61%, enquanto a taxa de erro apresentada pela árvore para esse conjunto de dados foi de 23,6%. B) Processed.hungarian - Duas das regras analisadas neste conjunto de dados e que apresentavam 13,5% e 18,51% de erros em relação aos acertos foram consideradas ótimas regras pelo especialista, pois elas analisam os mesmos atributos que um especialista analisaria para diagnosticar um paciente. Outra regra analisada, apesar de apresentar 28,95% de erro, foi classificada corretamente pelo especialista durante o duplo cego. Outras duas regras com 2,15% e 9,8% de erros também foram consideradas muito boas pelo especialista e foram classificadas corretamente. Algumas regras analisadas foram consideradas muito boas pelo especialista, mas considerou que merecem uma investigação mais profunda. O erro considerando-se a classe majoritária foi de 36,05%, enquanto a taxa de erro apresentada pela árvore para esse conjunto de dados foi de 12,6%.

CONCLUSÃO: O conhecimento extraído na forma das árvores de indução para o conjunto processed.hungarian foi analisado pelo especialista e considerado bastante coerente. Uma parte do conhecimento gerado pela base processed.hungarian foi visto pelo especialista com grande interesse nos incentivando a continuar o trabalho sobre esse conjunto de dados utilizando a estratégia de Aprendizado de Máquina Simbólico.Este trabalho mostrou que a interação entre diferentes áreas de conhecimento pode produzir resultados interessantes. Assim, para que a aplicação do Aprendizado de Máquina possa gerar frutos é necessário que dois grupos de pesquisadores estejam unidos: aqueles que conhecem métodos de AM e aqueles com o conhecimento no domínio da aplicação para o fornecimento de dados e avaliação do conhecimento adquirido.

REFERÊNCIAS BIBLIOGRÁFICAS:

Baranauskas, J. A.;Monard, M.C. An Unified Overview of Six Supervised Symbolic Machine Learning Inducers. Technical Report 103, ICMC-USP, Fevereiro 2000. (ftp://ftp.icmc.sc.usp.br/pub/BIBLIOTECA/rel_tec/RT_103.zip).

Debra J. Richardson. UCI Irvine Repository of ML Databases. http://www.ics.uci.edu/~mlearn/MLRepository.html.

Monard, M. C.;Batista, G. E. A. P. A.;Kawamoto, S.;Pugliesi, J. B. Uma Introdução ao Aprendizado Simbólico de Máquina por Exemplos. Notas do ICMC, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, Outubro 1997.

Morgan, J.; Messeger; R.THAID: A sequential Search Program for the Analysis of Nominal Scale Dependent Variables. Technical report, Institute Social Research, University of Michigan, 1973.

Quinlan, J. R. (1987a). Generating production rules from decision trees. In *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, Italy.

Quinlan, J. R. (1987b). Simplifying decision trees. *International Journal of Man-Machine Studies*.

Quinlan, J. R. (1993). C4.5: *Programs for Machine Learning*. Morgan Kaufmann. San Francisco, CA.

Robbins, S. L.; Ramzi, S. C.; Venay K. Patologia Estrutural e Funcional. 5ª edição. Rio de Janeiro, 1996.