O Processo de KDD – *Knowledge Discovery in Database* para Aplicações na Medicina

MARIZA FERRO
HUEI DIANA LEE

UNIOESTE - Universidade Estadual do Oeste do Paraná
CECE – Centro de Engenharias e Ciências Exatas
Curso de Ciência da Computação
Cx. Postal 961 – CEP 85.870-900 Foz do Iguaçu (PR)
{mariza98,huei}@dcc.unioeste-foz.br

Resumo: Neste trabalho é apresentado o processo de Knowledge Discovery in Database (KDD), também citado por alguns autores como Análise Inteligente dos Dados no caso de pequenos volumes de dados, como nos estudos de caso que apresentamos aqui. Mostramos como é composto o processo de KDD, dando ênfase a uma das suas etapas que é o data mining, que por sua vez é apoiado pelo Aprendizado de Máquina. Dois estudos de caso serão apresentados, mostrando como o processo de KDD pode ser aplicado para medicina, onde utilizamos duas bases de dados, uma delas para predizer doenças coronárias e outra para o processamento de sêmen, onde na etapa de data mining aplicamos métodos de Aprendizado de Máquina Simbólico Indutivo Supervisionado.

Palavras Chaves: KDD, Extração de Conhecimento, Aprendizado de Máquina.

1 Introdução

No passado, a tecnologia limitada, tanto em relação ao hardware como em relação ao software, permitia o armazenamento de um pequeno volume de dados. Simples consultas aos bancos de dados eram suficientes para a análise destes dados. No presente, grandes avanços tecnológicos, tanto para o hardware quanto para o software, permitem o armazenamento e o processamento de grandes volumes de dados e a necessidade de conhecer e entender a base de dados torna-se cada vez maior. A análise manual se tornou impraticável e métodos eficientes para a análise dos dados auxiliados por computador se tornaram indispensáveis, bem como a extração das informações e padrões que existem nesses dados. A lacuna que existe entre a geração de grandes volumes de dados e a compreensão e extração de novos conhecimentos sobre esses dados vêm crescendo em todas as áreas da atividade humana, porém, cobrir esse buraco é particularmente crucial na medicina. Deste modo, é cada vez mais importante a aplicação de métodos que consigam ajudar no processo de tomada de decisão, descoberta de novos conhecimentos médicos específicos (diagnósticos, prognósticos, monitoramento, etc) sobre um paciente individual ou um conjunto de casos de vários pacientes ao longo do tempo vem se tornando cada vez mais indispensável.

Assim, através do processo de KDD, que envolve desde a preparação de dados, seleção de dados, limpeza de dados, data mining, incorporação de conhecimento anterior apropriado, e interpretação formal dos resultados de mineração, o conhecimento que é derivado dos dados e que pode ser usado para auxiliar nas tomadas de decisões [Fayyad et. al (1996)].

2 O processo de KDD

O termo processo implica que existem vários passos envolvendo preparação de dados, procura por modelos, avaliação de conhecimento e refinamento, todos estes repetidos em múltiplas iterações. O KDD, segundo alguns autores também é citado como Análise Inteligente dos Dados quando se tratar de pequenos volumes de dados, mas que seguem os mesmos passos. Os vários passos do KDD podem ser resumidos como:

1. Conhecimento do domínio da aplicação: inclui o conhecimento relevante e as metas do processo KDD para a aplicação.

2 . Criação de um banco de dados alvo: inclui selecionar um conjunto de dados ou dar ênfase para um subconjunto de variáveis ou exemplo de dados nos quais o "descobrimento" será realizado.

3. Limpeza de dados e pré-processamento : inclui operações básicas como remover ruídos, coleta de informação necessária para modelagem, decidir estratégias para manusear (tratar) campos perdidos, etc. Redução de dados e projeção : inclui encontrar formas práticas para se representar dados, dependendo da meta do processo e o uso de redução dimensionável e métodos de transformação para reduzir o número efetivo de variáveis que deve ser levado em consideração ou encontrar representações invariantes para os dados.

4. Escolha da função de data mining : inclui a decisão do propósito do modelo derivado do algoritmo de data mining (Ex. classificação, regressão e clusterização). Encontrar o algoritmo de data mining: inclui selecionar métodos para serem usados para procurar por modelos nos dados, como decidir quais modelos e parâmetros podem ser apropriados e determinar um método de data mining particular com o modelo global do processo KDD (Ex. o usuário pode estar mais preocupado em entender o modelo do que nas suas capacidades).

5. Interpretação : Inclui a interpretação do modelo descoberto e possível retorno a algum passo anterior como também uma possível visualização do modelo extraído, removendo modelos redundantes ou irrelevantes e traduzindo os úteis em termos compreendidos pelos usuários.

6. Utilização do descobrimento obtido : inclui incorporar este conhecimento na performance do sistema, tomando ações baseadas no conhecimento, ou simplesmente documentando e reportando para grupos interessados. [Fayyad et. al (1996)].

2.1 Data Mining

Data mining é um dos principais passos no processo KDD (Knowledge Discovery in Database), e pode ser definido como um processo de descoberta de padrões nos dados. O processo pode ser automático ou, geralmente, semi-automático. A etapa de data mining pode ser apoiada por basicamente quatro áreas: Aprendizado de Máquina - AM (abordado com mais detalhes a seguir), banco de dados, estatística e visualização.

Data mining utiliza análises estatísticas sofisticadas e Inteligência Artificial - IA, mais especificamente o AM, para descobrir padrões escondidos e relações em bases de dados.

Em data mining, os dados são armazenados eletronicamente e a busca é automatizada por computador que está envolvido com a análise de dados presentes em bancos de dados. Economistas, estatísticos e engenheiros de comunicação têm extenso trabalho com a idéia de que padrões em dados podem ser buscados automaticamente, identificados, validados e usados para predição. O uso de banco de dados em atividades cotidianas traz o data mining para a vanguarda das novas tecnologias empresariais [Witten & Frank (1999)].

2.2 Fundamentos do Data Mining

Muitas técnicas de data mining foram desenvolvidas no passado para extrair informações de dados. Ou seja, data mining é a combinação de diferentes técnicas de sucesso comprovado, como inteligência artificial, estatística e bancos de dados.

Inteligência Artificial – Desde o início dos anos 60, a comunidade de Inteligência Artificial – IA tem pesquisado sistemas que seriam capazes de aprender. Uma classe destes sistemas é chamada de algoritmos de indução. Estes são capazes de induzir um modelo do processo de decisão de um especialista, com base em um conjunto de exemplos.

Métodos Estatísticos - Os algoritmos usados em IA são adequados para descobrir regras e modelos em conjuntos de dados artificiais e relativamente pequenos. A premissa feita é a de que todas as informações necessárias estão disponíveis. Toda a informação utilizada pelo especialista para tomar uma decisão é armazenada na base de dados. Nestas condições, o algoritmo de indução gera modelos que fazem predições corretas para cada exemplo no conjunto de dados e modela corretamente o processo de decisão do especialista. Conjuntos de dados reais não contêm toda a informação necessária para se tomar decisões corretas. Isto não significa que alguns dados têm ruído ou que algumas variáveis são desconhecidas, significa que algumas informações relevantes simplesmente não estão disponíveis. Neste momento a estatística pode entrar. Data mining utiliza inteligência artificial combinada à estatística para gerar bons modelos, mesmo que toda a informação necessária não esteja presente. Testes estatísticos são necessários para validar a qualidade do modelo.

Métodos de Banco de Dados - Data mining é uma tarefa computacionalmente cara. Durante o processo de busca, a qualidade de muitos modelos tem de ser validada. Informações estatísticas sobre os dados são necessárias para avaliar o quão correto é o modelo gerado. Desta forma, data mining tipicamente envolve o envio de milhares de consultas aos bancos de dados da empresa, resultando em altos tempos de resposta e carga de trabalho pesada.

Técnicas de data mining avançadas utilizam métodos de otimização para reduzir a interação com as bases de dados. Por exemplo, o resultado de uma consulta é temporariamente armazenado, de forma que consultas subseqüentes a informações similares

podem ser atendidas sem que a base de dados seja acessada.

2.3 Aprendizado de Máquina

Para aplicar métodos de Aprendizado de Máquina - 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 a forma como o conhecimento adquirido será representado, a estratégia é a forma como o conhecimento é propriamente adquirido.

Diversos paradigmas e estratégias de AM foram propostos. Entre os paradigmas podemos citar: conexionista, genético, estatístico, instanced-based e simbólico [Monard et. al (1997)]. 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 (1973)] 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 (1987a), Quinlan (1987b)].

As estratégias de aprendizado são classificadas de acordo com o grau de complexidade, i.é, dificuldade de aprendizado do conceito por parte do aprendiz. O aprendizado de um novo conceito pode ser realizado de várias maneiras [Monard et. al (1997)]: 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.

Em resoluções de problemas médicos é importante que um sistema de apoio de decisão possa explicar e justificar suas decisões. Especialmente quando se defrontar com uma solução inesperada para um problema novo, o usuário requer explicações e justificativas significativas. Consequentemente a interpretabilidade de conhecimento induzido é uma propriedade importante de sistemas que induzem soluções de dados médicos sobre casos resolvidos no passado; métodos simbólicos de data mining tem esta propriedade (como as árvores de decisão). Já o data mining sub-simbólico, tipicamente não tem esta propriedade, o que dificulta o seu uso em aplicações que explicações são requeridas. Não obstante, quando a precisão do classificador é o critério principal, os métodos sub-simbólicos podem ser muito apropriados, desde que eles tipicamente alcançam precisões que são pelo menos tão boas quanto dos classificadores simbólicos.

A indução de regras é um dos métodos simbólicos de data mining utilizados, onde dado um conjunto de exemplos classificados, um sistema de indução de regras constrói um conjunto de regras da forma IF Condições THEN Conclusão. Um exemplo é coberto por uma regra se o valor do atributo do exemplo cumprir as condições da regra.

Outro método simbólico de data minig utilizado é a indução de árvores de decisão, onde cada nó interior da árvore é etiquetado por um atributo, enquanto ramos que conduzem do nó são etiquetadas pelos valores do atributo.

O processo de construção de árvore é heuristicamente guiado escolhendo o atributo mais informativo a cada passo, com o objetivo de minimizar o número esperado de testes necessários para classificação. Uma árvore de decisão é construída chamando um algoritmo de construção de árvore repetidamente em cada nó gerado da árvore. No nó atual, o conjunto de treinamento atual é quebrado em subconjuntos de acordo com os valores do atributo mais informativo, e recursivamente, uma sub-árvore é construída para cada subconjunto. A construção da árvore pára quando todos os exemplos em um nó são quase da mesma classe. Este nó, chamado folha, é etiquetado por um nome de classe [Lavrac, (1999)].

3 Estudos de Caso

Serão apresentados aqui, segundo os passos do processo de KDD anteriormente descritos, os conjuntos de dados com os quais foram realizados os estudos de caso. O processo se diferenciou para cada um dos conjuntos de dados. O conjunto de dados Heart Disease é um conjunto de dados natural, obtido

no repositório de dados UCI Irvine [Debra, (2000)], o qual está dividido em outros dois conjuntos de dados: processes.switzerland e processed.hungarian, de acordo com a cidade do centro de pesquisa que coletou os dados. O segundo conjunto de dados utilizado é o Processamento de Sêmen, que consiste de dados reais de uma clinica de reprodução humana.

Heart Disease

1. O conhecimento do domínio incluiu extensas consultas bibliográficas sobre cardiopatias coronárias e consulta com especialista para compreender a relevância do problema em questão e compreender a importância da meta de se predizer o número de artérias principais obstruídas que um paciente possuía. As cardiopatias representam hoje uma das causas mais importante de morbidade e mortalidade em todas as nações industrializadas; o diagnóstico de doenças coronárias é baseado nos dados pessoais de um paciente, como a idade e resultados de vários exames médicos como pressão sangüínea e resultados de eletrocardiogramas e dados dos pacientes como sexo, idade, entre outros. O problema é tentar predizer se uma das quatro principais artérias do coração está com seu diâmetro reduzido em mais de 50% .

2. Neste estudo de caso o banco de dados com conjuntos de dados já selecionados estavam prontos, já que são dados naturais coletados do repositório de dados UCI Irvine [Debra, (2000)]; os dados já se encontravam pré-selecionados, sendo que dos 76 atributos apenas 13 estavam disponíveis no conjunto de dados e já estavam em formato eletrônico. O conjunto de Switzerland possui 123 instâncias enquanto o de Hungarian 294.

3. A limpeza dos dados e a pré-seleção não foi realizada na íntegra; não foi necessária a limpeza dos dados e a pré seleção já tinha sido realizada quando os dados foram coletados do repositório. No entanto foi necessário a adequação do formato dos dados ao aceito pelo algoritmo aplicado no passo 4 do processo.

4. Neste estudo de caso foi aplicado o método de Aprendizado de Máquina Indutivo Simbólico Supervisionado utilizando o algoritmo C4.5 para indução de árvores. Buscava-se tanto a extração de novos conhecimentos possivelmente interessantes quanto a construção de classificadores para apoio à tomada de decisão. Esta modelagem dos dados precisou ser retomada várias vezes; os dados apresentavam erros e a descrição dos atributos não eram claros para a adequação dos arquivos. Tanto o passo 3 quanto o 4 tiveram que ser retomados várias vezes para que pudéssemos obter os resultados para a análise do passo 5.

5. Para a análise dos resultados foram aplicados os algoritmos citados acima e juntamente com o especialista os resultados gerados pelas árvores e as regras de decisão foram analisados. A árvore de decisão foi analisada usando-se uma técnica muito utilizada na medicina chamada de *Duplo Cego*, no qual se descrevia atributos presentes na árvore e o especialista emitia sua opinião sobre em qual classe aquela regra deveria se encaixar. Isso era feito sem que ele soubesse qual era a classe atribuída àquela regra pelo indutor. 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 daquela regra. De acordo com os resultados obtidos com as árvores geradas pelo C4.5 a fase 4 não precisou ser retomada.

6. O conhecimento que foi documentado por este estudo caso e reportado pelo especialista para o conjunto processed.switzerland foi de que as regras eram de má qualidade e não puderam ser classificadas através do duplo cego por apresentarem poucas características para se realizar um diagnóstico; as regras são de má qualidade, apesar das taxas de erros apresentadas pelas regras serem baixas, tais como 11,1%, 0% e 6,97%. 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. O erro considerando-se a classe majoritária foi de 61%, enquanto a taxa de erros da árvore para este conjunto de dados foi de 23,6%.

   Para processed.hungarian algumas das regras analisadas foram consideradas muito boas pelo especialista, considerando que merecem futuras investigações mais profundas. Duas das regras analisadas neste conjunto de dados que apresentavam, respectivamente, 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. O erro considerando-se a classe majoritária foi de 36,05%, enquanto que a taxa de erro apresentada pela árvore foi de 12,6%.

Processamento de Sêmen

1. Da mesma forma que o conjunto anterior o estudo sobre o processamento de sêmen também envolveu estudos bibliográficos e consultas com especialistas para a compreensão do domínio e a relevância do estudo em questão. O Processamento de Sêmen é uma etapa bastante importante para a reprodução assistida, que permite quantificar a qualidade do sêmen (processamento diagnóstico) e ainda recuperar a maior quantidade possível de espermatozóides (processamento terapêutico) para que se decida qual técnica de reprodução assistida será utilizada no tratamento; o processamento de sêmen chega a aumentar os custos em até 80% em relação ao valor de um espermograma [Lee et al. (2000)].
2. A criação da base de dados consistiu em uma análise de resultados de exames de pacientes da clínica Androfert, que estavam em formato Word for Windows. Foram selecionados apenas os exames que possuíam os resultados do processamento de sêmen.
3. Foram necessárias análises com o especialista para a pré-seleção de quais seriam os atributos usados para a base de dados de acordo com sua relevância para a escolha da técnica de reprodução assistida. Os dados que estavam em documentos Word foram transferidos para uma planilha eletrônica. Foi realizada uma Seleção de Features e Construção de Features [Lee (2000)]. Várias limpezas foram realizadas nos dados cada vez que os passos seguintes do processo de KDD vem sendo, ainda, aplicados. Este conjunto de dados possui 385 instâncias, que envolvem dados de espermogramas e do exame de processamento de sêmen, tais como concentração total de espermatozóides por ml coletado, percentagem de espermatozóides vivos, entre outros. Os dados faltantes foram tratados sem problemas, já que os algoritmos que foram escolhidos para serem aplicados trabalham bem com dados faltantes ou com ruídos.
4. Para o data mining foram aplicados métodos de Aprendizado de Máquina Indutivo Simbólico Supervisionado. A ferramenta selecionada para a aplicação dos métodos foi o algoritmo C4.5 para indução de árvores de decisão. Buscamos extrair conhecimentos interessantes que não puderam ser percebidos pelos especialistas da área médica, além disso gostaríamos de predizer informações até agora só conseguidas através da realização destes exames para evitar sua aplicação, reduzindo assim os custos. Esta modelagem dos dados vem sendo ainda retomada.
5. A análise dos resultados ainda não foi concluída com a base de dados atual que está sendo trabalhada, porém trabalhos anteriores já mostraram resultados interessantes [Lee et al. (2000)].

4 Conclusões

No decorrer deste trabalho apresentamos o processo de KDD, suas fases e aplicações, com ênfase para a fase de data mining a qual é apoiada por métodos de AM; este processo é chamado de Análise Inteligente dos Dados por alguns autores quando a base de dados é pequena, como em nossos exemplos.

Apresentamos dois exemplos da aplicação do processo de KDD para a medicina, onde a geração de grandes volumes de dados e a extração de conhecimento desses dados de forma manual tem se tornado impraticável. Assim, o processo de KDD, que envolve desde a preparação dos dados, seleção, limpeza, data mining e interpretação dos resultados pode auxiliar nas tomadas de decisões e na descoberta de conhecimentos específicos, tais como diagnósticos e prognósticos.

Agradecimentos: Agradecemos ao Dr. Wu Feng Chung por sua colaboração na análise dos resultados obtidos do conjunto de dados Heart Disease. Agradecemos também ao Dr. Sandro C. Esteves por seu auxílio na análise dos dados Processamento de Sêmen.

5 Referências

Debra J. Richardson. UCI Irvine Repository of ML Databases. http://www.ics.uci.edu/~mlearn/MLRepository.html. 2000.

Fayyad, U.; Shapiro, P. G.; Smyth, P. “Knowledge Discovery and Data Mining: Towards a Unifying Framework”. Disponível em: <http://citeseer.nj.nec.com/cs> Acesso em: 9 jun. 2001. 1996.

Lavrac, N. “Machine Learning for Data Mining in Medicine”. In [Werner et al (1999)]. 1999.

Lee, Huei D. “Seleção e Construção de Features Relevantes para o Aprendizado de Máquina”. Dissertação de Mestrado ICMC-USP, Fevereiro 2000.

Lee, H. D.; Monard, M. C.; Esteves, S.C. “Indução Construtiva Guiada pelo Conhecimento: Um Estudo de Caso do Processamento de Sêmen Diagnóstico”. SBIA – Simpósio Brasileiro de Inteligência Artificial, Atibaia, SP. 2000.

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. & Messenger; 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. “Generating production rules from decision trees”. In Proceedings of the Tenth International Joint Conference on Artificial Intelligence, Italy. 1987[a]

Quinlan, J. R. “Simplifying decision trees”. International Journal of Man-Machine Studies. 1987b.

Werner, H., Yuval, S., Greger, L., Steen, A., Jeremy, W. “Artificial Intelligence in Medicine”. Joint European Conference on AI in medicine and medical decision making, AIMDM’99 proceedings, 1999.