Predictive data mining in clinical medicine: Current issues and guidelines

Resumo:

Antecedentes: A ampla disponibilidade de novos métodos computacionais e ferramentas para análise de dados e modelagem preditiva requer que pesquisadores e profissionais de informática médica selecionem sistematicamente a estratégia mais apropriada para lidar com problemas de previsão clínica. Em particular, a coleção de métodos conhecidos como ‘mineração de dados’ oferece soluções metodológicas e técnicas para lidar com a análise de dados médicos e construção de modelos de previsão. Uma grande variedade desses métodos requer diretrizes gerais e simples que podem ajudar os profissionais na seleção adequada de ferramentas de mineração de dados, construção e validação de modelos preditivos, juntamente com a disseminação de modelos preditivos em ambientes clínicos.

Objetivo: O objetivo desta revisão é discutir a extensão e o papel da área de pesquisa de mineração de dados preditiva e propor uma estrutura para lidar com os problemas de construção, avaliação e exploração de modelos de mineração de dados na medicina clínica.

Métodos: Nós revisamos o trabalho relevante recente publicado na área de mineração de dados preditiva em medicina clínica, destacando questões críticas e resumindo as abordagens em um conjunto de lições aprendidas.

Resultados: o artigo fornece uma revisão abrangente do estado da arte da mineração de dados preditiva em medicina clínica e fornece diretrizes para a realização de estudos de mineração de dados neste campo.

Conclusões: A mineração de dados preditiva está se tornando um instrumento essencial para pesquisadores e profissionais clínicos em medicina.

A compreensão das principais questões subjacentes a estes métodos e a aplicação dos procedimentos acordados e normalizados é obrigatória para a sua implementação e divulgação dos resultados. Graças à integração de dados moleculares e clínicos dentro da medicina genômica, a área recentemente não só ganhou um novo impulso, mas também um novo conjunto de problemas complexos que precisa resolver.

Introdução:

Nos últimos anos, o termo ‘mineração de dados’ tem sido cada vez mais usado na literatura médica. Em geral, o termo não foi ancorado em nenhuma definição precisa, mas em algum tipo de entendimento comum de seu significado: o uso de (novos) métodos e ferramentas para analisar grandes quantidades de dados.

A mineração de dados tem sido aplicada com sucesso em diferentes campos da atividade humana, incluindo marketing, bancos, gestão de relacionamento com o cliente, engenharia e várias áreas da ciência. No entanto, sua aplicação à análise de dados médicos - apesar das grandes esperanças - até recentemente era relativamente limitada. Isso é particularmente verdadeiro para aplicações práticas em medicina clínica que podem se beneficiar de abordagens específicas de mineração de dados que são capazes de realizar modelagem preditiva, explorar o conhecimento disponível no domínio clínico e explicar as decisões propostas, uma vez que os modelos são usados ​​para apoiar as decisões clínicas. O objetivo da mineração de dados preditiva em medicina clínica é derivar modelos que podem usar informações específicas do paciente para prever o resultado de interesse e, assim, apoiar a tomada de decisão clínica. Métodos de mineração de dados preditivos podem ser aplicados à construção de modelos de decisão para procedimentos como prognóstico, diagnóstico e planeamento de tratamento, os quais - uma vez avaliados e verificados - podem ser embutidos em sistemas de informação clínica.

Neste artigo, fazemos uma revisão metodológica da mineração de dados, enfocando seu processo de análise de dados e destacando algumas das questões mais relevantes relacionadas à sua aplicação na medicina clínica. Limitamos o escopo do artigo à mineração de dados preditiva cujos métodos são metodologicamente maduros e muitas vezes

facilmente disponível e pode ser particularmente adequado para a classe de problemas decorrentes da análise de dados clínicos e suporte à decisão.

A mineração de dados é o processo de seleção, exploração e modelagem de grandes quantidades de dados para descobrir padrões ou relacionamentos desconhecidos que fornecem um resultado claro e útil para o analista de dados [1]. Cunhado em meados da década de 1990, o termo mineração de dados tornou-se hoje um sinônimo para "Descoberta de Conhecimento em Bancos de Dados" que, conforme proposto por Fayyad et al. [2], enfatizou o processo de análise de dados ao invés do uso de métodos de análise específicos. Os problemas de mineração de dados são frequentemente resolvidos usando um mosaico de diferentes abordagens extraídas da ciência da computação, incluindo bancos de dados multidimensionais, aprendizado de máquina, computação suave e visualização de dados, e de estatísticas, incluindo testes de hipóteses, agrupamento, classificação e técnicas de regressão. O ofício da mineração de dados reside na escolha e combinação adequadas dessas técnicas para resolver um determinado problema de forma eficiente e confiável.

As tarefas de mineração de dados podem, em geral, ser classificadas em tarefas de descrição e previsão. Enquanto a descrição visa encontrar padrões e associações interpretáveis ​​por humanos, depois de considerar os dados como um todo e construir um modelo de predição procura prever alguma resposta de interesse. Embora os objetivos de descrição e predição possam se sobrepor (os modelos gerados por alguns métodos de predição podem apontar alguns padrões interessantes), a principal distinção é que a predição requer que os dados incluam uma variável de resposta especial. A resposta pode ser categórica ou numérica, portanto, classificando ainda mais a mineração de dados preditiva como, respetivamente, classificação e regressão. Nesta revisão, abordamos problemas de classificação em particular: enquanto a diferença entre os dois reside principalmente no conjunto de métodos usados, o processo de mineração de dados aplicado a problemas de regressão e classificação é bastante semelhante.

Antes de prosseguirmos, usamos um exemplo simples para apresentar os conceitos básicos na mineração de dados preditiva e para mostrar a aplicação de duas técnicas de classificação de dados populares, mas bastante diferentes.

2.2. Métodos de mineração de dados preditivos (classificação)

Os métodos de mineração de dados preditivos originam-se de diferentes campos de pesquisa e frequentemente usam abordagens de modelagem muito diversas.

Eles vêm em vários sabores e podem ser comparados com base em:

• o tratamento de dados faltantes e ruído;

• o tratamento de diferentes tipos de atributos (categóricos, ordinais, contínuos);

• a apresentação de modelos de classificação que podem ou não permitir que o especialista no domínio os examine e compreenda seu funcionamento interno;

• a redução do número de testes [15], ou seja, a redução dos atributos necessários para derivar a conclusão;

• o custo computacional para indução e o uso de modelos de classificação;

• sua capacidade de explicar as decisões alcançadas quando os modelos são usados na tomada de decisão;

• generalização, ou seja, a capacidade de um bom desempenho com casos não vistos.

Listamos aqui alguns dos métodos de mineração de dados preditivos mais comumente usados ​​e os organizamos de acordo com uma classificação recente do pool relevante em KDNuggets (http://www.kdnuggets.com/polls/2006/data mining methods.htm, abril de 2006 ) que pedia aos mineradores de dados para nomear as técnicas que eles usam com mais frequência:

As árvores de decisão usam particionamento de dados recursivo, induz classificadores transparentes cujo desempenho pode sofrer com a segmentação de dados: as folhas nas árvores de decisão podem incluir poucas instâncias para obter previsões confiáveis. A complexidade computacional dos algoritmos de indução é baixa devido a heurísticas poderosas. A maioria dos conjuntos de mineração de dados atuais incluem variantes de C4.5 e seu sucessor e algoritmos de indução de árvore de decisão CART [11,12].

As regras de decisão na forma de 'IF com base na condição em valores de atributos ENTÃO valor de resultado' podem ser construídas a partir de árvores de decisão induzidas como nas regras C4.5 [11], ou podem ser derivadas diretamente dos dados como é o caso com Algoritmos AQ e CN2 [13,14]. Embora em seu desempenho esses algoritmos compartilhem a maioria de suas características com as árvores de decisão, eles podem ser computacionalmente mais caros.

A regressão logística é um método poderoso e bem estabelecido de estatísticas [16]. É uma extensão da regressão comum e pode modelar um resultado de dois valores que geralmente representa a ocorrência ou não ocorrência de algum evento. Como com o classificador bayesiano ingênuo, o modelo subjacente para probabilidade é multiplicativo [17], mas usa um método mais sofisticado baseado em uma estimativa de máxima verossimilhança para determinar os coeficientes em sua fórmula de probabilidade. O tratamento dos valores de atributos ausentes não é simples. O modelo pode ser bem representado por meio de um nomograma [6,8].

As redes neurais artificiais foram até recentemente o algoritmo de modelagem de dados baseado em inteligência artificial mais popular usado na medicina clínica. Isso provavelmente se deve ao seu bom desempenho preditivo, embora possam ter uma série de deficiências [18] que incluem alta sensibilidade aos parâmetros do método - incluindo aqueles que determinam a arquitetura da rede, alto custo computacional em treinamento e indução do modelo que pode - na melhor das hipóteses - ser difícil de interpretar por especialistas no domínio. As redes neurais podem ser capazes de modelar relações não lineares complexas, compreendendo uma vantagem sobre métodos de modelagem mais simples, como o classificador Bayesiano ingênuo ou regressão logística.

Máquinas de vetores de suporte (SVM) são talvez o algoritmo de classificação mais poderoso da atualidade em termos de precisão preditiva [19]. Eles são baseados em fortes fundamentos matemáticos e teoria de aprendizagem estatística [20]. Central para o método é um procedimento que encontra um hiperplano que separa os exemplos de resultados diferentes (ver Fig. 5). Sendo projetados principalmente para problemas de duas classes, os SVMs encontram um hiperplano com uma distância máxima para o ponto mais próximo das duas classes; esse hiperplano é chamado de hiperplano ótimo. Um conjunto de instâncias que está mais próximo do hiperplano ideal é chamado de vetor de suporte. Encontrar o hiperplano ideal fornece um classificador linear. Além desses kernels lineares, as máquinas de vetores de suporte também são frequentemente usadas com outros kernels não lineares que, em essência, transformam o espaço de atributo original em um novo espaço de dimensão superior no qual o classificador linear é inferido.

Funções de kernel populares são, por exemplo, funções polinomiais, de base radial e sigmóides. A escolha do kernel apropriado deve, em princípio, depender das propriedades do conjunto de dados e do domínio do problema. Para conjuntos de dados reais, o hiperplano que separaria claramente os exemplos de classes diferentes na maioria das vezes não existe. Para resolver este problema, foi proposto um método de margem suave [21] onde o hiperplano resultante divide o conjunto de dados em dois conjuntos tão limpos quanto possível, ou seja, onde uma classe prevalece no maior grau possível.

As máquinas de vetores de suporte estão se tornando cada vez mais populares na medicina e, em particular, na bioinformática. Com exceção dos kernels lineares, onde a estrutura do modelo pode ser facilmente revelada por meio dos coeficientes que definem um hiperplano linear, as máquinas de vetores de suporte usam um formalismo que geralmente é inadequado para interpretação por especialistas humanos.

Como tal, e se estivermos apenas interessados ​​na precisão preditiva, as máquinas de vetores de suporte são um contendor sério para as redes neurais artificiais, especialmente porque seu desempenho pode ser mais robusto e depender menos da seleção específica dos parâmetros do método.

O classificador bayesiano ingênuo é uma abordagem que já introduzimos. Apesar de sua simplicidade, seu desempenho é frequentemente pelo menos comparável a outras abordagens mais sofisticadas [15,22]. Devido à rápida indução de um classificador, ele pode ser usado como um algoritmo de linha de base em estudos comparativos. Quando superado em desempenho preditivo por outros algoritmos mais sofisticados, isso geralmente indica a presença de interações não lineares entre atributos.

Redes bayesianas são modelos gráficos probabilísticos que são capazes de expressar convenientemente uma distribuição de probabilidade conjunta sobre uma série de variáveis ​​por meio de um conjunto de distribuições de probabilidade condicional. Uma rede Bayesiana é um gráfico acíclico direcionado onde cada nó representa uma variável estocástica e os arcos representam uma dependência probabilística entre um nó e seus pais. Cada variável xi é considerada independente de seus não descendentes, dado seu conjunto de pais, pa (xi). Sob esta suposição, conhecida como suposição de Markov, a distribuição de probabilidade conjunta de todas as variáveis ​​(x) pode ser escrita seguindo a chamada regra da cadeia:

p (x) = ni = 1

p (xi | pa (xi))

A rede é totalmente especificada por um conjunto de distribuições de probabilidade condicional que quantifica as relações qualitativas entre as variáveis ​​expressas pelo gráfico. Essas distribuições de probabilidade dependem de um conjunto de parâmetros, como as entradas das tabelas de probabilidade condicional para variáveis ​​discretas ou a média e a variância da distribuição gaussiana para variáveis ​​contínuas. Embora as redes bayesianas tenham sido tradicionalmente aplicadas na medicina como um instrumento para realizar o raciocínio probabilístico [23-26], vários algoritmos e ferramentas estão disponíveis hoje em dia para aprender a estrutura do grafo e o modelo probabilístico subjacente a partir dos dados [27-30]. Redes bayesianas podem ser facilmente aplicadas em problemas de classificação, onde podem ser vistas como uma generalização do classificador Bayesiano ingênuo, modelando as interações entre as variáveis ​​do problema. Eles são agora cada vez mais usados ​​em mineração de dados preditiva e em bioinformática [31,32]. As principais desvantagens desse método residem no aprendizado da estrutura do grafo, que pode exigir um grande conjunto de dados, e na interpretação das causalidades inferidas [28].

Os algoritmos para o aprendizado de redes Bayesianas a partir de dados são baseados no framework de seleção de modelos Bayesianos. O objetivo é aprender a estrutura S com a maior distribuição de probabilidade posterior, dado um conjunto de dados x. Essa distribuição de probabilidade posterior pode ser calculada como:

p (S | x) = p (x | S) p (S)

p (x) ∝ p (x | S) p (S)

A posterior é proporcional ao produto de dois termos, a saber, a verossimilhança marginal p (x | S), que mede a probabilidade do modelo com relação aos dados x, e a probabilidade anterior da estrutura p (S). A verossimilhança marginal é obtida como a média da verossimilhança sobre os valores do conjunto de parâmetros. A verossimilhança marginal pode ser calculada de forma aproximada apenas quando as variáveis ​​são discretas [29] e quando o modelo é condicionalmente gaussiano [33].

A comparação da distribuição posterior das diferentes estruturas requer a exploração do espaço de todas as estruturas possíveis, um problema que se mostra intratável com uma abordagem de força bruta. Para lidar com este problema, muitos algoritmos heurísticos têm sido propostos na literatura. O mais amplamente aplicado é o algoritmo de busca guloso K2, descrito na Ref. [29], mas algoritmos genéticos [34] e técnicas de Cadeia de Markov Monte Carlo [35] também foram aplicadas com sucesso.

O algoritmo de k-vizinhos mais próximos é inspirado na abordagem frequentemente adotada por especialistas de domínio que tomam decisões com base em casos semelhantes vistos anteriormente [17]. Para uma determinada instância de dados, o classificador de k-vizinhos mais próximos pesquisa as k instâncias de treinamento mais semelhantes e classifica com base em sua classe predominante. A pesquisa pelas instâncias mais semelhantes pode

ser lento e requer a recuperação de um conjunto completo de treinamento no momento da classificação.

Os métodos listados acima costumam ser parte integrante da maioria dos conjuntos de mineração de dados modernos e, sozinhos ou em combinação com o pré-processamento, costumam ter um desempenho bom e suficientemente rápido.

As maiores diferenças no tratamento de dados clínicos podem surgir no desempenho preditivo e na interpretabilidade dos resultados.

Ao longo desta revisão, defendemos que ambos são importantes e se os métodos funcionam de forma semelhante com relação à precisão, aqueles que oferecem uma explicação e modelos interpretáveis ​​devem ser preferidos.

Os padrões de mineração de dados preditivos são esparsos, mas emergentes. Os que recentemente estão ganhando atenção e ampla aceitação são CRISP-DM, SEMMA e PMML. Os dois primeiros são padrões que definem o processo de mineração de dados. O CRISP-DM foi elaborado pelo Processo Padrão entre Indústrias para o Grupo de Interesse em Mineração de Dados (www.crisp-dm.org), que no final dos anos 1990

definiu o chamado Modelo de Processo CRISP-DM [36]. O CRISP-DM divide a mineração de dados em várias fases: negócios e compreensão de dados, preparação de dados, modelagem, avaliação e implantação. Ele define as entradas, saídas e estratégias gerais a serem aplicadas em cada fase. SEMMA (amostra, explorar, modificar, modelar, avaliar) é uma metodologia de mineração de dados proposta pelo SAS Institute e usada em seu poderoso pacote de mineração de dados. Enquanto o CRISP-DM fornece um modelo de gerenciamento de projeto abrangente, SEMMA se concentra principalmente na aplicação de estatísticas exploratórias e técnicas de mineração de dados baseadas em visualização.

A linguagem de marcação de mineração de dados preditiva (PMML, www.dmg.org) é muito relevante para a comunicação, compartilhamento e implantação de modelos preditivos. PMML é um padrão aberto independente de fornecedor emergente para definir modelos de mineração de dados. Ele define uma linguagem de marcação baseada em XML para a codificação de muitos modelos de mineração de dados preditivos que incluem árvores de decisão e regras, Classificadores Bayesianos Nave e modelos de regressão logística. A versão mais recente de suítes populares de mineração de dados oferece suporte a esse padrão, sendo capaz de exportar e importar modelos codificados em arquivos XML compatíveis.

Relacionados à mineração de dados preditivos estão os padrões para apresentação e codificação de dados, como a Nomenclatura Systemized of Human and Veterinary Medicine (SNOMED) e o Unified Medical Language System (UMLS). Mas como esses padrões são empregados em sistemas de gerenciamento de banco de dados clínicos que os mineradores de dados usam para obter seus dados, e como ajudam na uniformidade e consistência dos conjuntos de dados [37], eles (ainda?) Não se tornaram uma parte explícita de qualquer sistema de mineração de dados que aborda a análise de dados médicos. Uma exceção a isso, mas que está além do escopo desta revisão, é a mineração de texto médico, onde, por exemplo, o UMLS foi usado para relacionar conceitos médicos e resumos de artigos citados no Medline [38,39].

**2.4. Mineração de dados e estatísticas preditivas**

Em sua breve história, a mineração de dados foi, no início, um tanto enganosamente associada apenas a métodos de análise de dados provenientes de campos diferentes da estatística. As características expostas desses métodos são que eles podem abordar grandes quantidades de dados, fazer uso de diferentes tipos de dados (vários tipos de atributos, mineração de texto), são muito flexíveis na modelagem (por exemplo, inclusão de não linearidade) e podem automatizar a maioria do processo de análise. O sucesso inicial das abordagens em áreas como análise de cesta de mercado e mineração de texto, juntamente com a ênfase exagerada de abordagens de aprendizado de máquina e reconhecimento de padrões em suítes de mineração de dados emergentes, provocou vários estatísticos para incentivar sua comunidade a se envolver e contribuir o campo [40]. Desde então, o campo amadureceu substancialmente e seu amadurecimento também se reflete na forma como a mineração de dados atual se relaciona com as estatísticas. Os conjuntos de mineração de dados estão se tornando parte de grandes pacotes estatísticos, os principais livros sobre mineração de dados foram escritos por estatísticos [17,41], enquanto muitos desenvolvimentos recentes em mineração de dados se concentraram em estatísticas de ponte, visualização e abordagens de análise de dados de vários campos da engenharia . Uma distinção muito enfatizada entre estatística clássica e mineração de dados envolve o tamanho dos dados tratados pela última [40]. A mineração de dados lida com a análise secundária.

Lá, os dados não são coletados propositalmente para testar algumas hipóteses de pesquisa, mas são obtidos de bancos de dados legados ou data warehouses onde o volume de dados pode ser muito maior. Neste artigo, porém, argumentamos que, para aplicações em dados clínicos

análise de outros aspetos da mineração de dados pode ser apenas ou até mais relevante. Mais importante, isso inclui tornar o conhecimento descoberto a partir dos dados explícito e comunicável aos especialistas do domínio, o fornecimento de uma explicação ao implantar e usar esse conhecimento com novos casos e a capacidade de codificar e usar o conhecimento do domínio no processo de análise de dados.

Conjuntos de dados - incluindo aqueles retirados da medicina clínica - são frequentemente propensos a diferentes fontes de ruído, eles podem incluir vários tipos de recursos preditivos (por exemplo, nominal, valor real, vêm de uma série temporal, etc.), podem incluir um substancial número de valores de recursos ausentes e podem ser regidos por

processos não lineares subjacentes. A mineração de dados moderna e os métodos estatísticos são frequentemente poderosos o suficiente para lidar com a maioria desses casos, com a principal diferença na abordagem para a descoberta de modelos preditivos. A análise exploratória de dados em estatísticas na maioria das vezes envolve uma pesquisa manual e a modelagem de, por exemplo, não linearidades e interações de atributos, ao passo que, ao usar a mineração de dados, deve-se primeiro confiar em técnicas automáticas, como indução construtiva [42], descoberta de interação de atributos [ 43] e abordagens de modelagem não linear que pesquisam sistematicamente os dados e o espaço de atributos. O conhecimento existente em algum domínio do problema influenciaria nas estatísticas a composição do conjunto de dados a ser coletado, enquanto - quando apropriadamente codificado - ajudaria a mineração de dados a se concentrar e relatar apenas os padrões relevantes para o problema encontrados na análise de dados secundários. Para algoritmos de mineração de dados, os dados são apenas uma fonte de informação e outras incluem qualquer conhecimento adicional que pode ser obtido, codificado e tornado útil na análise.

2,5. Mineração de dados preditiva e medicina genômica

Nos últimos anos, a mineração de dados preditiva recebeu um forte impulso da pesquisa em biologia molecular. Métodos de mineração de dados, como agrupamento hierárquico [44] ou máquinas de vetores de suporte [45], são rotineiramente aplicados na análise de dados de alto rendimento provenientes de microarrays de DNA ou espectrometria de massa. Curiosamente, nos últimos anos, vários artigos destacaram o potencial dos dados preditivos

mineração para inferir modelos clinicamente relevantes de dados moleculares e, portanto, fornecer suporte à decisão no novo campo da medicina genómica [46]. Hoje em dia, três tipos diferentes de dados moleculares podem estar disponíveis para os médicos: (i) dados do genótipo, muitas vezes representados por uma coleção de um único nucleotídeo

polimorfismos (SNPs), variações da sequência de DNA que ocorrem quando um único nucleotídeo na sequência do genoma é alterado; uma vez que cada indivíduo tem muitos SNPs, sua sequência forma um padrão único de DNA para aquela pessoa; (ii) dados de expressão gênica, que podem ser medidos com microarrays de DNA para obter um instantâneo da atividade de todos os genes em um tecido em um determinado momento ou com técnicas que dependem de uma reação em cadeia da polimerase (PCR) e PCR em tempo real quando a expressão de apenas alguns genes precisa ser medida com maior precisão; (iii) dados de expressão de proteínas, que podem incluir um conjunto completo de perfis de proteínas obtidos com tecnologias de espectros de massa ou alguns marcadores de proteínas que podem ser medidos com ensaios ad hoc. A maioria dos trabalhos publicados na área de mineração de dados preditivos para medicina genômica tem como objetivo analisar dados de expressão gênica provenientes de microarranjos de DNA, constituídos por milhares de genes para cada paciente, com o objetivo de diagnosticar (sub) tipos de doenças e para obter um prognóstico que pode levar a decisões terapêuticas individualizadas.

Os trabalhos publicados estão principalmente relacionados à oncologia, onde há uma forte necessidade de definir estratégias terapêuticas individualizadas [47]. Um artigo seminal dessa área é o de Golub et al. [48] ​​e enfoca o problema do diagnóstico diferencial precoce de leucemia mieloide aguda (LMA) e leucemia linfoblástica aguda (LLA). Eles foram capazes de derivar um modelo de classificação baseado em uma abordagem de votação ponderada baseada em uma lista de cerca de 50 genes. Hoje, há muitos relatórios que mostram a utilidade potencial dos dados de microarray de DNA para uma previsão de resultado no tratamento do câncer [49-51 ]. Para melhorar a precisão da classificação e a relevância clínica dos modelos prognósticos, alguns autores propuseram uma integração de dados clínicos e de expressão gênica. Nevins e col. [52] propôs uma abordagem baseada em árvore de decisão em que os genes são primeiro agrupados em "metágenos" e, em seguida, usados em uma árvore de decisão em conjunto com dados clínicos, como o status do nódulo linfático, para prever a sobrevivência de um paciente. Uma abordagem diferente foi proposta por Futschik et al. [53], onde os dados clínicos e de microarray são usados para construir dois modelos separados para a previsão do resultado de linfoma difuso de grandes células B. Os modelos empregados são um classificador Bayesiano e uma Rede Neural Fuzzy. A previsão final é obtida por meio de um classificador de ensemble cujos parâmetros também são inferidos dos dados de treinamento. Parâmetros clínicos e expressões gênicas também foram combinados na modelagem de regressão de Cox na estratificação de risco de meduloblastomas [54].

Nos últimos anos, surgiram algumas críticas contra as abordagens baseadas na expressão gênica para construir modelos preditivos e derivar listas de genes úteis para a previsão de resultados [55]. Embora vários grupos tenham publicado listas de genes preditivos com desempenho preditivo muito bom, foi observado que eles podem variar amplamente de estudo para estudo. Essa variabilidade pode estar relacionada à falta de robustez devido ao pequeno número de casos clínicos em relação ao número de atributos. Um comentário recente de Berrar et al. [56] aponta que muitos artigos de mineração de dados em genômica e proteômica são afetados pelo chamado viés de seleção, uma vez que a seleção de recursos é frequentemente (incorretamente) realizada em todo o conjunto de dados antes da validação cruzada. Este procedimento adapta muito o classificador ao conjunto de dados. O mesmo problema, com referência a publicações iniciais proeminentes sobre a modelagem de classificação de dados de expressão de genes do câncer, foi observado por Simon et al. [57]. Ein-Dor et al. [58] usaram uma análise teórica para mostrar que milhares de pacientes podem ser necessários para obter listas de genes confiáveis. A integração do conhecimento sobre a função dos genes e sobre os processos biomédicos com dados clínicos e de expressão gênica e a fusão de dados provenientes de diferentes estudos [59] são direções promissoras para melhorar a robustez e o impacto prático desses estudos.

Graças à possibilidade de medir espectros de massa do soro, perfis proteômicos extraídos de técnicas de espectrometria de massa [45] foram analisados ​​para derivar modelos preditivos. Neste caso, o conjunto de recursos é representado por algumas centenas ou milhares de relações massa / carga, na dependência da resolução da técnica de medição. Abordagens de mineração de dados preditivos nesta área foram aplicadas para prever os resultados dos pacientes no caso de câncer de próstata e ovário [45,60,61]. Nessas aplicações, a fase de pré-processamento é crucial. Por exemplo, Yu et al. [45] desenvolveram uma estratégia baseada em uma combinação de filtragem de recursos com o teste de Kolmogorov-Smirnov, análise de wavelet e máquinas de vetores de suporte para definir o modelo preditivo. O alto número de recursos, combinado com a necessidade de pré-processamento dos espectros brutos, torna o problema de aprendizado de modelos robustos muito difícil. Como os dados proteômicos são caracterizados por muitos recursos e muito menos casos, o risco de overfitting é ainda maior do que com conjuntos de dados de microarray. Várias propostas de procedimentos sistemáticos para extrair modelos preditivos de dados de espectrometria de massa foram recentemente propostas para evitar esses problemas [62,63]. Os marcadores de expressão de proteína também são amplamente usados ​​para construir modelos de prognóstico em câncer, embora recentemente tenha havido grande interesse na aplicação de modelagem estatística e mineração de dados para a análise de dados de microarranjos de tecido, que são uma nova ferramenta de alto rendimento para o estudo de padrões de expressão de proteínas em tecido [64,65]. Podemos esperar que esta área dê origem a várias aplicações de mineração de dados nos próximos anos.

Outra área onde a mineração de dados preditiva foi

aplicada é a análise de dados sobre polimorfismos de nucleotídeo único (SNPs). Os estudos de associação do genoma geralmente incluem várias centenas de pacientes e controles e consideram várias centenas de milhares de SNPs, com o objetivo de identificar aqueles para os quais o risco da doença é aumentado.

O objetivo ambicioso é usar informações de SPN para encontrar a base genética das chamadas características complexas, ou seja, aquelas características que não seguem estritamente a herança de Mendel. A definição de um modelo prognóstico multivariada é uma tarefa típica de mineração de dados que tem sido estudada em vários artigos. Por exemplo, Sebastiani et al. [32] usam redes bayesianas para extrair as relações entre SNPs e o risco de um acidente vascular cerebral em pacientes que sofrem de anemia falciforme. Eles foram capazes de extrair um modelo com um pequeno número de genes que foram validados em um conjunto de dados separado, prevendo a ocorrência de um acidente vascular cerebral em 114 indivíduos com 98,2% de precisão. Curiosamente, neste artigo a seleção de genes e SNPs foi realizada integrando o conhecimento prévio no processo de análise de dados. A construção de modelos baseados em SNPs não é trivial uma vez que tem que enfrentar os mesmos problemas de dimensionalidade em proteômica e genômica mencionados acima. Quando é importante modelar (ou descobrir) interações SNP, geralmente é necessário limitar a análise a várias dezenas de SNPs devido à disponibilidade de dados e à complexidade da análise [66]. Para melhorar a escalabilidade, o progresso no campo dependerá do uso de análise de interação, indução construtiva e visualização [67]. Além disso, há uma necessidade de integrar a análise estatística padrão com base em pedigrees e um desequilíbrio de ligação para reduzir o número de SNPs.

Com as interações gene-gene desempenhando um papel importante na suscetibilidade e progressão de doenças comuns e resposta ao tratamento, e com os estudos de caso-controle emergentes que coletam dados SNP de todo o genoma, o próximo passo lógico é um gene de todo o genoma análise de interação gênica. No entanto, as ferramentas de mineração de dados que poderiam considerar centenas de milhares de SNPs e perfis de expressão de genes e proteínas de milhares de pacientes ainda não existem. Um grande desafio para os cientistas da computação é, portanto, disponibilizar essas ferramentas e projetar heurísticas eficientes para superar a pesquisa exaustiva e proibitivamente complexa por interações gênicas. O desafio de projetar tal software é fornecer uma interface de análise exploratória interativa que forneça aos usuários que não são cientistas da computação um suporte contínuo na descoberta de interação e na formação de novas hipóteses a serem testadas em um laboratório molhado.

3. Contribuição da mineração de dados para a modelagem preditiva em medicina clínica

Modelos preditivos em medicina clínica são "... ferramentas para ajudar na tomada de decisões que combinam dois ou mais itens de dados do paciente para prever resultados clínicos" [68]. Esses modelos podem ser usados ​​em vários contextos clínicos por médicos e podem permitir uma reação imediata a situações desfavoráveis ​​[69]. A mineração de dados pode contribuir efetivamente para o desenvolvimento de modelos preditivos clinicamente úteis graças a pelo menos três aspectos inter-relacionados:

(a) uma abordagem abrangente e objetiva para a análise de dados que envolve a aplicação de métodos e abordagens extraídas de diferentes áreas científicas;

(b) a capacidade explicativa de tais modelos;

(c) a capacidade de usar o conhecimento do domínio (histórico) no processo de análise de dados.

3.1. Um processo sistemático e integrado

Como uma disciplina de engenharia, a mineração de dados conta com seu modelo de processo associado que, sendo tão importante para o campo, foi recentemente considerado com muita atenção e para o qual vários padrões foram desenvolvidos. A vantagem, mas também a dificuldade da mineração de dados, é que ela é uma estrutura que integra várias abordagens diferentes de diversas disciplinas. Seguindo as etapas padrão no estudo de um problema, a análise e implantação de dados podem ajudar os pesquisadores a fazer uso sistemático dessas várias ferramentas e escolher apropriadamente entre as técnicas disponíveis. Assim como os protocolos na medicina, os padrões de processo na mineração de dados ajudam seus usuários, orientando-os no processo de análise, expondo os aspectos que poderiam ser esquecidos ou negligenciados. Recentemente, uma série de grandes suítes de mineração de dados, como o SPSS 'Clementine (www.spss.com/spssbi/clementine) e o SAS' Enterprise Miner (www.sas.com/products/miner), fizeram uso do processo padrões explícitos: ali, o usuário escolhe a fase que deseja abordar e só é mostrado um conjunto de ferramentas aplicáveis ​​a essa fase.

3.2. Explicação

A mineração de dados inclui abordagens que podem desempenhar um papel duplo:

eles podem ser usados ​​para derivar uma regra de classificação e para entender quais informações estão contidas nos dados disponíveis.

Inspirado nos primeiros sistemas especialistas como Mycin [70] e Internist [71] que estavam bastante enraizados em aplicações médicas, a comunicação explícita do conhecimento descoberto a partir dos dados e a explicação subsequente das decisões quando esse conhecimento é usado na classificação de novos casos é o que é enfatizado por uma série de técnicas de mineração de dados. No exemplo introdutório, já demonstramos que as árvores de classificação podem revelar padrões interessantes em dados observacionais.

Exemplos em que tal análise levou à descoberta de novos conhecimentos médicos incluem estudos de lesão cerebral [72], geriatria [73] e atendimento ao trauma [74].

Outro formalismo para representar modelos de classificação que permite uma fácil explicação dos resultados são as redes Bayesianas. Uma aplicação interessante de aprendizagem de rede Bayesiana no contexto de mineração de dados preditiva tem sido

publicado recentemente por Sierra e Larranaga [75]. Em seu trabalho, eles comparam a precisão de uma abordagem bayesiana ingênua na previsão da sobrevivência de pacientes com melanoma cutâneo maligno com a de três diferentes redes bayesianas induzidas a partir dos dados. A Fig. 6A mostra um exemplo de uma rede Bayesiana induzida reproduzida a partir do artigo original. Notamos que, por exemplo, a variável sexo não é considerada útil para classificar os casos, enquanto as variáveis ​​‘número de nós positivos’, ‘espessura’ e ‘estágio’ são consideradas dependentes uma da outra. A Fig. 6B mostra o modelo de Bayes ingênuo do mesmo problema. Uma vez que a estrutura da rede permanece fixa após a fase de aprendizagem, a saída gráfica reflete apenas as suposições a priori sobre as relações de variáveis, enquanto o conhecimento aprendido está oculto nas tabelas de probabilidade.

3.3. A utilidade do conhecimento de domínio

Juntamente com a capacidade de explicar, alguns algoritmos de mineração de dados podem levar em consideração o chamado "conhecimento de fundo". Literalmente, o conhecimento prévio é a "informação que é essencial para a compreensão de uma situação ou problema" [76]. No processo de construção de um modelo preditivo, usar o conhecimento de fundo significa ser capaz de levar em conta informações que já são conhecidas e não devem ser redescobertas a partir dos dados. Esta questão pode ser particularmente importante na análise de dados médicos [77].

O conhecimento prévio pode ser expresso em diferentes formatos: exemplos podem ser encontrados nas áreas de regras de decisão [78], Modelos Bayesianos [79], conjuntos fuzzy [80] e hierarquias de conceito [3,81]. Entre outros, um método que pode ser particularmente apropriado para lidar e codificar o conhecimento de fundo envolve redes bayesianas. Em redes bayesianas, o conhecimento prévio é explorado para definir a estrutura da rede, ou seja, o número de variáveis, arcos e direções do arco.

Além disso, seguindo o paradigma Bayesiano, as probabilidades anteriores nas tabelas de probabilidade condicional são fornecidas para levar em consideração o conhecimento prévio disponível sobre as relações entre as variáveis ​​do problema. Probabilidades anteriores permitem que um modelo seja derivado mesmo quando a informação proveniente dos dados é fraca, e pode ajudar a evitar overfitting, onde o modelo derivado reproduz os dados muito próximos e falha em classificar corretamente os casos novos e invisíveis [82].

Um exemplo dessa abordagem é um estudo em uma rede bayesiana projetada para avaliar a profilaxia da doença do enxerto contra o hospedeiro após o transplante de medula óssea em crianças [79].

A estrutura da rede foi avaliada com base no conhecimento prévio disponível, enquanto as probabilidades foram primeiro definidas por especialistas e então atualizadas com base em um conjunto de dados de cinquenta pacientes [79]. O uso de conhecimento prévio na construção de redes bayesianas e na obtenção de suas probabilidades é uma área ativa de pesquisa [82-84], onde a mineração de dados e a engenharia do conhecimento frequentemente combinam seus esforços e resultados.

O conhecimento prévio também pode ser facilmente explorado na construção de regras de classificação: por exemplo, um conjunto incompleto de regras de classificação fornecidas pelo especialista pode ser refinado e aumentado com base nos dados disponíveis, enquanto a pesquisa de regras pode ser conduzida por um certo número de restrições de monotonicidade [85,86].

4. Processo de mineração de dados preditivo: tarefas

e orientações

A mineração de dados costuma ser a aplicação de várias técnicas diferentes de várias disciplinas com o objetivo de descobrir padrões interessantes de dados. Dada a grande variedade de técnicas disponíveis e campos interdisciplinares, não é surpresa que a mineração de dados seja frequentemente vista como uma arte difícil de aprender e ainda mais difícil de dominar.

Como mencionamos, vários modelos e padrões de processos foram propostos para introduzir princípios de engenharia, sistematizar o processo e definir tarefas típicas de mineração de dados. Na seção sobre padrões, apresentamos o CRISP-DM, um

padrão de processo de mineração que parece estar ganhando a mais ampla aceitação. Embora o CRISP-DM enumere uma série de métodos que podem ser usados ​​para realizar tarefas de mineração de dados, não se destina a fornecer diretrizes precisas sobre quais técnicas, esquemas de avaliação e estatísticas usar. Ou seja, todos eles devem ser específicos para um domínio de problema, tarefas particulares de mineração de dados e o tipo de dados em consideração. A mineração de dados preditiva em medicina clínica é um exemplo de uma tarefa específica e diretrizes que abordam diferentes aspectos da análise de dados médicos podem ser fornecidas para acompanhar o modelo CRISP-DM e torná-lo mais útil neste domínio. Na seguinte descrição do processo de mineração preditiva de dados, geralmente aderimos ao esquema CRISP-DM, mas também tentamos ser específicos e listar uma série de problemas, recomendações e diretrizes que podem se aplicar à mineração preditiva médica e que foram propostas e avaliadas por pesquisadores e desenvolvedores ativos na área.

4.2. Preparação de dados

Para mineração de dados, os dados clínicos geralmente vêm de bancos de dados dedicados (ou mesmo de formulários em papel) que foram coletados propositalmente para estudar um problema clínico específico. Embora ainda não esteja amplamente disponível, outra fonte importante de dados clínicos são os armazéns de dados. Atualmente, os algoritmos de mineração de dados mais amplamente usados ​​requerem que os dados sejam colocados em uma forma tabular que inclua fatores preditivos e resultados e seja construída por meio da consulta de um ou vários bancos de dados dedicados [91].

Uma regra importante na construção e avaliação de

modelos preditivos é que eles nunca devem ser construídos e testados no mesmo conjunto de dados. Para isso, técnicas como validação cruzada são usadas (consulte a próxima seção), mas também pode ser uma boa ideia nesta fase dividir os dados em dois conjuntos: o primeiro, muitas vezes referido como o conjunto de aprendizagem, é usado para compare diferentes algoritmos de mineração de dados, estime seu desempenho usando algumas métricas estatísticas, encontre o melhor conjunto de parâmetros para classificação de recursos, seleção e métodos de aprendizagem e, finalmente, selecione a técnica de modelagem com melhor desempenho. Usando essa técnica, um modelo final deve ser desenvolvido a partir de um conjunto de aprendizado completo e testado em um segundo conjunto de dados, comumente referido como um conjunto de validação. A divisão de dados pode ser arbitrária ou baseada no tempo ou rótulo de origem das instâncias de dados. Conjuntos separados de aprendizagem e validação são necessários para avaliar objetivamente o desempenho preditivo. Os modelos de mineração de dados podem ser complexos e, em casos extremos, podem "lembrar" cada instância de dados com a qual aprenderam. Esses modelos funcionam perfeitamente em dados que foram usados ​​para aprendizagem, mas mal com qualquer novo caso que não corresponda a alguma instância de dados dos dados de aprendizagem. Diz-se que tais modelos se generalizam mal devido ao sobreajuste. A maioria das técnicas contemporâneas de mineração de dados incluem mecanismos eficientes para evitar overfitting, como poda para árvores de decisão, limitando a complexidade da rede neural e a seleção apenas das regras mais significativas para modelagem de regra de decisão, mas é apenas a avaliação de dados independentes conjunto que pode garantir que o bom desempenho não resultou de overfitting.

4.3. Modelagem e avaliação

Uma vez que os dados são divididos em um conjunto de aprendizado e validação, agora é hora de empregar nossas técnicas de modelagem e ajustar seus parâmetros ao conjunto de dados de aprendizado. O objetivo desta fase é determinar qual algoritmo de mineração de dados tem o melhor desempenho para que possamos usá-lo para gerar nosso modelo preditivo de destino. Os modelos preditivos podem ser avaliados com base em seu desempenho preditivo e compreensibilidade. Dos dois, o desempenho preditivo é mais fácil de quantificar e as estatísticas típicas incluem métricas como sensibilidade, especificidade, precisão de classificação [92], área sob a curva ROC [93] e escore de Brier [94]. A compreensibilidade é uma medida subjetiva que é avaliada por especialistas do domínio participantes. Embora possa ser proibitivamente difícil quantificar a compreensibilidade, modelos preferíveis podem ser encontrados respondendo a perguntas como: ‘dados os dois modelos, qual é mais fácil de entender? qual explica melhor as decisões? qual deles os especialistas têm mais confiança usando? 'Se a compreensibilidade e a explicação estão em jogo, os algoritmos de mineração de dados podem ser classificados primeiro usando as estatísticas de desempenho preditivas escolhidas e, em seguida, dos poucos modelos de classificação superior, os especialistas de domínio podem selecionar o modelo final com base na sua compreensibilidade e capacidade de explicação.

Conforme mencionado na seção anterior, para estimar as estatísticas que avaliam o desempenho preditivo, uma abordagem desejável é aplicar a chamada estratégia de retenção: um subconjunto do conjunto de aprendizagem, o conjunto de treinamento, é usado para construir o modelo enquanto outro subconjunto, o conjunto de teste, é usado para estimar a precisão do modelo. No entanto, o procedimento de validação faz um uso bastante ineficiente dos dados: a estratégia típica é aprender com dois terços dos dados e, então, testar no terço restante da amostra. Tal estratégia pode não ser aplicável com um pequeno número de dados, uma vez que os algoritmos para aprendizado do modelo prognóstico podem ter problemas devido ao conjunto de dados reduzido para aprendizado, enquanto o teste pode ainda ser insuficiente para atingir os limites de intervalo de confiança desejados. Um método contemporâneo popular a ser usado na resolução dos problemas mencionados acima é a validação cruzada k-fold. Com a validação cruzada, os dados são divididos em um número (k) de subconjuntos de dados que contêm aproximadamente um número igual de instâncias de dados e correspondem aproximadamente à distribuição de resultados do conjunto de aprendizagem (validação cruzada estratificada). Normalmente, o conjunto de dados de aprendizagem é dividido em dez subconjuntos de dados (validação cruzada de 10 vezes). Em seguida, os dados dos nove subconjuntos são usados ​​para modelagem, enquanto o subconjunto restante é usado para testar o modelo resultante e avaliar as estatísticas. O processo de treinamento e teste é repetido 10 vezes, cada vez usando um subconjunto de teste diferente. As estatísticas médias são então relatadas e caracterizam o método de modelagem. Além da validação cruzada, outras abordagens de divisão de dados podem ser usadas, como validação cruzada "não-excluída", amostragem aleatória, bootstrap, etc. [95,96].

Atenção especial deve ser dada à estimativa dos parâmetros. A maioria dos métodos de mineração de dados depende de um conjunto de parâmetros que definem o comportamento do algoritmo de aprendizagem e influenciam direta ou indiretamente a complexidade dos modelos resultantes.

Por exemplo, o grau de poda pode ser definido para indução da árvore de decisão, o número de unidades em uma camada oculta pode ser definido para modelos de rede neural feed-forward e o nível necessário de significância estatística pode ser definido para regras de decisão. Embora a descoberta do melhor conjunto de parâmetros possa ser

caracterizado como uma busca no espaço de parâmetros que emprega alguma técnica de otimização de última geração, os profissionais frequentemente definem um conjunto de valores de parâmetros mais prováveis ​​e, novamente por meio de validação cruzada, avaliam cada conjunto separadamente para encontrar o vencedor. A avaliação dos métodos de mineração de dados produz não apenas a classificação das técnicas de mineração de dados, mas também identifica o conjunto de parâmetros apropriado a ser usado. Observe também que a classificação de recursos, seleção de subconjunto e construção podem ter seus próprios parâmetros, que também requerem otimização.

4,4. Construção do modelo preditivo alvo

As técnicas de avaliação descritas acima fornecem os fundamentos para classificar métodos de mineração de dados e identifica um conjunto adequado de parâmetros.

Agora podemos usar um conjunto de dados de avaliação completo e o método mais bem classificado para construir nosso modelo preditivo. O modelo resultante é então avaliado no conjunto de dados de validação e, se seu desempenho preditivo for aceitável, este é agora nosso modelo preditivo de destino. Observe que, ao relatar as qualidades preditivas do modelo, apenas as estatísticas obtidas usando o conjunto de dados de validação têm mérito e devem importar; relatar os resultados dos conjuntos de dados de aprendizagem pode ser enganoso, pois eles estão sujeitos a sobreajuste. Se a tarefa da mineração de dados é observar as relações entre os recursos e os recursos e o resultado, agora é hora de escrutinar, analisar e visualizar o modelo resultante.

5. Discussão

Em comparação com a mineração de dados nos negócios, marketing e economia, as aplicações de mineração de dados médicos têm várias características distintas [104]. O mais importante é que a medicina é um contexto crítico de segurança [105] no qual as atividades de tomada de decisão devem sempre ser apoiadas por explicações.

Isso significa que o valor de cada dado pode ser maior do que em outros contextos: os experimentos podem ser caros devido ao envolvimento do pessoal e ao uso de instrumentação cara e devido ao potencial desconforto dos pacientes envolvidos.

Na mineração clínica, os conjuntos de dados podem ser pequenos e relatar situações não reproduzíveis. Os dados podem ser ainda afetados por várias fontes de incerteza, como erros de medição ou dados ausentes ou erros na codificação das informações enterradas em relatórios textuais. Médicos e pesquisadores lidam com essas dificuldades explorando seu conhecimento do domínio. Da mesma forma, a mineração de dados pode lidar com esses problemas aplicando cuidadosamente a seleção de variáveis ​​e modelos, avaliando corretamente os modelos resultantes e codificando explicitamente esse conhecimento e usando-o na análise de dados [106]. Atualmente, a mineração de dados é um campo muito diversificado, com uma série de técnicas que podem servir ao mesmo propósito e se comportar igualmente bem. Pode não ser prático explorar todos os métodos alternativos ao minerar um determinado conjunto de dados, enquanto a escolha de quais técnicas usar costuma ser guiada pelos instintos de mineradores de dados especialistas. Embora seja improvável que, com a variedade atual de abordagens, a comunidade possa criar livros de receitas e receitas, tentamos fornecer algumas descrições gerais de tarefas e um conjunto simples de diretrizes que podem ser aplicadas à construção de modelos preditivos clínicos usando técnicas de mineração de dados .

No geral, as idéias que apresentamos podem ser resumidas na seguinte lista:

• Defina os critérios de sucesso com antecedência. Definir intervalos aceitáveis

de estatísticas de avaliação antes da modelagem.

• Se possível e para referência, compare os resultados de desempenho com os obtidos a partir da modelagem estatística clássica.

• Probabilidades de modelo, não associação de classe nítida. Prefira métodos que relatem intervalos de confiança.

• Evite overfitting. Nunca teste modelos em dados que foram usados ​​em sua construção. No caso de pequenos conjuntos de dados, use validação cruzada ou técnicas semelhantes para obter avaliação

Estatisticas.

• Se possível, teste o modelo resultante em um conjunto de dados separado e independente.

• Relate as pontuações de desempenho com intervalos de confiança.

• Prefira técnicas de modelagem que expõem relações e podem

apresentá-los de uma forma legível. Se a descoberta de relacionamentos é um objetivo da mineração de dados, evite modelos de caixa preta.

• Se ainda tiver um desempenho aceitável, prefira uma modelagem simples

técnicas, possivelmente aquelas que derivam modelos que podem ser

revisado e criticado por especialistas.

• Classificação de recursos, seleção de recursos, indução construtiva e assim por diante, juntamente com qualquer estimativa de parâmetro, são todos parte da modelagem e devem ser testados dentro da validação cruzada. Usá-los no pré-processamento que ocorre antes da validação cruzada leva ao sobreajuste.

• O projeto não é concluído quando um bom modelo é encontrado.

Pense em como incluir seu modelo em algumas informações clínicas ou sistema de suporte à decisão. Se possível, realize um estudo de custo / benefício.

• Avalie explicitamente a aplicabilidade do modelo e seu potencial de generalização. Aqui, considere em particular o tipo de coleta de dados (retrospectiva, prospectiva, derivada de um ensaio clínico ou de rotina clínica), o número de dados disponíveis e o desempenho do modelo.

Essas diretrizes estão relacionadas a questões emergentes na medicina personalizada e genómica. Hoje, a construção de modelos preditivos confiáveis ​​pode exigir a integração de dados extraídos de fontes heterogêneas que incluem dados clínicos, laboratoriais, genéticos, genómicos e proteômicos. A disponibilidade total de repositórios de dados e armazéns capazes de fornecer simultaneamente essas informações sobre um único paciente e os métodos para integrá-los a um sistema de suporte à decisão são questões que ainda precisam ser resolvidas.

6. Conclusão

No momento, muitos métodos de mineração de dados preditivos maduros foram aplicados com sucesso a uma variedade de problemas práticos em medicina clínica. Conforme sugerido por Hand [40], a mineração de dados é particularmente bem-sucedida onde os dados são abundantes.

Para a medicina clínica, isso inclui a análise de data warehouses clínicos, estudos epidemiológicos e estudos emergentes em genómica e proteômica. Crucial para esses dados são as abordagens de mineração de dados que permitem o uso do conhecimento de fundo, descobrir relacionamentos interpretáveis ​​e não triviais interessantes, construir modelos baseados em regras e outros modelos simbólicos que podem ser revisados ​​e examinados por especialistas, descobrir modelos que oferecem uma explicação quando usado para previsão e, finalmente, a descoberta do modelo de ponte e suporte à decisão para implantar modelos preditivos na prática clínica diária. Com as promessas oferecidas pela medicina genómica e as necessidades futuras de integrar dados moleculares e clínicos, a mineração de dados e outras abordagens computacionais intensivas em conhecimento estão se tornando necessárias para o avanço do estado da arte de aplicações de pesquisa e da vida real [107,108].

Por último, mas não menos importante, a mineração de dados clínicos lida com "problemas do lado da cama, isto é, com modelos que prevêem o resultado do paciente. A tomada de decisão que usa um modelo de previsão específico deve, portanto, também levar em consideração as questões de ética e o custo da previsão enquanto se preocupa com a análise dos resultados.