DOI: 10.5433/1981-8920.2014v19n1p249

PONTOS DE ATENÇÃO PARA O USO DA MINERAÇÃO DE DADOS NA SAÚDE

PUNTOS DE ATENCIÓN EN MINERÍA DE DATOS DE APOYO A LAS DECISIONES EN SALUD

Deborah Ribeiro Carvalho - ribeiro.carvalho@pucpr.br
Doutora em Computação de Alto Desempenho pela Universidade Federal Do Rio Janeiro (UFRJ). Professor da Pontifícia Universidade Católica do Paraná (PUCPR).

Leandro Fabian Almeida Escobar - l.escobar72@gmail.com
Mestrando do Programa de Pós-Graduação em Ciência e Gestão da Informação da Universidade Federal do Paraná (UFPR). Professor da Universidade Positivo (UP).

Denise Tsunoda - dtsunoda@gmail.com
Doutora em Engenharia Elétrica e Informática Industrial pela Universidade Tecnológica Federal do Paraná (UFPR). Professora da UFPR.

RESUMO

Introdução: A Mineração de Dados representa uma alternativa para apoiar a decisório, mas não constitui prática regular nas atividades da Saúde.
Objetivo: Revisar a literatura, identificando pontos de atenção na utilização da Mineração de Dados na área da Saúde.
Metodologia: Os pontos de atenção foram obtidos mediante a análise de publicações em periódicos, sistematizados em quadros de referência cruzada e foram descobertas de regras associação e respectivas exceções.
Resultados: Foram identificados 14 pontos, formando um conjunto de dados dos quais obtiveram-se 345 regras gerais e respectivas exceções. Os pontos mais citados são “previsão de eventos” e “auxílio ao planejamento” e os menos citados são “apropriação de protocolo específico” e “detecção de padrões em tempo real”. Percebe-se associação entre o “auxilio ao planejamento” e a “não detecção de padrões em tempo real”. Exceção ocorre, quando o “auxilio ao planejamento” é citado juntamente com “explicações causais”, associando-se à “detecção de padrões em tempo real”.

Conclusões: Os pontos de atenção indicam critérios para a adoção da Mineração de Dados na Saúde. Destaca-se “explicações causais” que, citado em 8 artigos, determina exceções nas associações entre os demais pontos, indicando que a seleção da tarefa de mineração passa pela adequação às expectativas dos usuários.

Palavras chave: Mineração de Dados em saúde. Pontos de atenção.

1 INTRODUÇÃO

A Mineração de Dados constitui uma alternativa para processar grandes volumes de dados dos Sistemas de Informação em Saúde, dada a sua capacidade de descobrir padrões úteis, novos e surpreendentes, possibilitando o apoio em análises complexas sobre dados clínicos.

O KDD (*Knowledge Discovery in Databases*), por sua vez, é um processo geral para a conversão de dados brutos em informações úteis (FAYYAD; PIATETSKI-SHAPIRO; SMYTH, 1996) e tipicamente constituído pelas etapas de pré-processamento, que envolve a de seleção, limpeza e preparação dos dados; processamento, que trata da descoberta de padrões mediante algoritmos de Mineração; pós-processamento, que refina os resultados obtidos durante o processamento, seja compondo novos padrões ou avaliando seu interesse, e interpretação dos padrões extraídos, culminando na obtenção de conhecimentos antes ocultos.

Vários são os desafios para a utilização da Mineração de Dados que se inicia com a preparação das bases de dados, seleção dos algoritmos que melhor se adequam ao problema proposto e por fim análise dos resultados, ou seja, dos padrões descobertos. Por isso mesmo, apesar de todo desenvolvimento já realizado, esta área continua sendo objeto de pesquisas por novas soluções que se aproximem do real interesse dos potenciais usuários. As razões pela constante pesquisa se devem ao fato de que, apesar da existência de várias experimentações quanto ao uso de Mineração de Dados, ainda é baixa a sua adoção nos processos de tomada de decisão diários (MARISCAL; MARBÁN; FERNANDEZ, 2010), por conta da pouca familiaridade dos especialistas com a metodologia, vantagens e dificuldades (MEYFROIT et al., 2009), a prevalência de estudos estatísticos com objetivo de revelar relações lineares simples entre os fatores de saúde (CRUZ-RAMIRES et al*.,* 2012) ou, ainda, a exploração comparativa de técnicas de Mineração de Dados, deixando de lado etapas relativas à interpretação e avaliação de resultados (BLOMBERG, 2010).

Portanto, para que a Mineração de Dados seja efetivamente incorporada ao processo de decisão em Saúde, faz-se necessário identificar quais os pontos de atenção durante o processo de descoberta de conhecimento nas bases de dados de forma a cumprir com os critérios de aceitação determinados pelos especialistas. Parte-se da hipótese, então, de que existem elementos e situações que requerem tratamento criterioso por parte do pesquisador para que a Mineração de Dados seja adotada como ferramenta efetiva de apoio à decisão em Saúde.

Este artigo se propõe, desta forma, a pesquisar e discutir possíveis pontos de atenção que venham a facilitar uma utilização mais intensiva da Mineração de Dados na Saúde. A pesquisa se baseia em relatos sobre experiências de aplicação da Mineração de Dados na Saúde, buscando identificar questões que facilitem ou não a sua adoção.

2 ENCAMINHAMENTOS METODOLÓGICOS

Uma pesquisa aplicada, qualitativa, exploratória e bibliográfica foi conduzida nas bases de Periódicos Capes, Scielo, *PubMed* e Domínio Público. Para a pesquisa foram utilizados os seguintes descritores: *Data Mining*, Mineração de Dados, KDD e *Acceptance*, KDD e *Obstacle*, *Data Mining* e *Acceptance* e *Health*.

Após a leitura e análise dos textos que atenderam ao descritor foram selecionados aqueles cujo tema relata a aplicação da Mineração de Dados e a identificação de estratégias para a efetividade no uso dos padrões descobertos para a área da Saúde. Os critérios de inclusão adotados foram: data de publicação posterior a 2005; apresentar a aplicação do KDD, bem como uma avaliação da efetividade do uso da Mineração de Dados discutindo benefícios e/ou dificuldades.

Foram identificados os aspectos citados pelos autores como importantes, tanto para a realização dos experimentos, como também para a devida incorporação na rotina dos especialistas/gestores. Estes aspectos foram agrupados por similaridade, conforme a natureza do problema, dificuldade ou solução caracterizando pontos de atenção.

Os dados coletados foram, sistematizados em uma tabela, sendo os pontos de atenção representados nas linhas os artigos contendo os relatos nas colunas. Cada célula (cruzamento das linhas e colunas) foi preenchida com S(im) ou N(ao), indicando presença ou ausência O conjunto de dados oriundo desta tabela submetido à tarefa de Descoberta de Regras de Associação.

Para a Mineração de Dados foi utilizado o algoritmo Apriori do ambiente WEKA (WITTEN; FRANK; HALL, 2011), que descobre regras de associação do tipo X → Y (se X então Y). Sendo X e Y representando itens de dados do conjunto de entrada. Para cada regra de associação descoberta são determinados o suporte da regra, indicando a ocorrência de X e Y em relação à base de dados e a confiança da regra, que indica a quantidade de registros que possuem X e também possuem Y. (AGRAWAL; SRIKANT, 1994).

As regras descobertas foram pós-processadas, para a identificação de possíveis exceções, pois tendem a ser mais interessantes que a respectiva regra geral (HUSSAIN et al., 2000).

Regra Geral: X →Y (se X então Y)
Regra de Exceção: X, A → Y (se X e A então não Y)

Para a descoberta das regras de exceção foi utilizado o DRE – Descobre Regras de Exceção (MILANI; CARVALHO, 2013).

3 RESULTADOS

Foram encontradas 660 publicações (Tabela 1), das quais 625 publicações foram excluídas por não estarem de acordo com os critérios de inclusão. Os 35 artigos remanescentes foram analisados, sendo novamente excluídos 17, por se limitarem a apresentar e explicar o KDD, não discutirem estratégias ou critérios para efetividade no uso da Mineração de Dados ou por não terem foco na Área da Saúde.

Resultando assim 18 publicações que estavam de acordo com os critérios de elegibilidade (Tabela 1).

Tabela 1 - Número de Artigos Encontrados e Analisados, segundo Bases de Periódicos.

Base Consultada	Artigos encontrados	Artigos analisados
Domínio Público	52	4
Periódicos da Capes	398	10
Pubmed	15	0
Scielo	195	4
TOTAL	**660**	**18**

Fonte: Produzido pelos autores

Analisando as 18 publicações percebe-se que 83% relatam experimentos envolvendo a área clínica e apenas 27% envolvendo gestão da saúde. Sobre as três grandes etapas do KDD, todos os autores relatam Mineração de Dados, 78% também discutem a etapa de Pré-Processamento e apenas 39% o Pós-Processamento.

A relação de pontos de atenção obtida a partir da leitura e análise das 18 publicações é**:**

- **Apropriação de protocolo específico**: atenção aos processos específicos de diagnóstico ou tratamento de acordo com indicação formal da especialidade médica a partir de protocolos ou diretrizes amplamente aceitos pela comunidade especializada.
- **Avaliação da qualidade dos dados**: trata da atenção e crítica à consistência, disponibilidade e tratamento dos dados. Este ponto de atenção está diretamente relacionado às etapas de seleção e preparação dos dados. Ainda que a qualidade dos dados faça do problema do desenvolvimento de Sistemas de Informação em geral, sua presença se justifica pela necessidade de se melhorar a qualidade da coleta de dados, mesmo ocorrendo em sistemas específicos ou ocasião diferente do momento da tomada de decisão em Saúde.
- **Integração de diferentes bases de dados**: Enriquecimento da base de dados a partir de diferentes fontes, sejam elas distribuídas, em formatos diversos ou ainda pertencentes à organizações diferentes da organização objeto do estudo. Um exemplo é a necessidade de complementar dados sobre atendimentos com dados sobre o perfil demográfica das regiões amostras do estudo.
- **Desenvolvimento de funcionalidade especifica**: Construção de software específico para uso ou exploração dos modelos gerados. Desenvolvimento de sistema ou interface de usuário que suporte o processo de tomada de decisão tanto para a visualização dos dados quanto para a navegação nos padrões encontrados.
- **Obediência a processo de trabalho específico**: O sistema de coleta de dados e de apoio à decisão reproduz ou segue um conjunto de procedimentos encadeados logicamente e definidos como modo de trabalho dentro da organização em saúde. Trata-se da reprodução de processo de trabalho particular.
- **Utilização de modelo de informações especificas**: Adoção de conjunto de informações específicas e previamente definidas seja por protocolos ou pela

própria organização Saúde. Trata-se da determinação de quais informações são relevantes sem que sejam indicadas as sequencias lógicas dos procedimentos ou ainda a ordem das atividades de atendimento.

- **Avaliação (subjetiva ou objetiva) da relevância dos padrões encontrados**: Avaliar os padrões extraídos determinando o quanto são relevantes ou interessantes para o especialista em saúde.
- **Combinação de diferentes tipos de tarefas de Mineração de Dados**: Uso de mais de uma tarefa de Mineração de Dados para a realização do experimento ou identificação de padrões.
- **Detecção de padrões em tempo real**: Extrair e apresentar padrões simultaneamente à ocorrência dos eventos que são fontes dos dados.
- **Auxílio ao planejamento em saúde**: Apoiar ações com base nos modelos extraídos pela Mineração de Dados.
- **Representação visual dos resultados**: Criação de recursos que facilitem a compreensão dos modelos extraídos mediante a visualização dos resultados. Trata da necessidade de comunicação amigável dos padrões encontrados para os especialistas em Saúde envolvido nos estudos.
- **Descrição de eventos ocorridos numa determinada população**: Descrever padrões aos quais uma determinada amostra está sujeita, explicando as relações encontradas entre os dados minerados.
- **Previsão da ocorrência de eventos numa determinada população**: Antecipar eventos de acordo com os padrões identificados nas bases de dados.
- **Fornecimento de subsídios para explicações causais**: Demonstração das cadeias de eventos, suas causas e consequências.

3.1 Síntese dos Estudos Analisados

Trindade et al. (2012) aplicou o KDD para a identificação de padrões de comportamento das Hepatites Virais nas bases de dados do SINAN (Sistema de Informação de Agravos e Notificações) do Sistema Único de Saúde – Governo Federal do Brasil, objetivando subsidiar ações de controle e prevenção da doença.

Foi aplicado o algoritmo C4.5 (QUINLAN, 1993) na descoberta de padrões e o C4.5*Rules* para pós-processar os padrões que caracterizam as Hepatites Virais.

Recursos visuais, como gráficos e mapas, foram utilizados para facilitar o entendimento dos padrões encontrados para o especialista.

A baixa qualidade dos dados foi um fator determinante. Dos 134 atributos selecionados, 65 (48%) puderam ser utilizados. Dentre os problemas destacados, está a ausência do dado, sugerindo falhas durante coleta dos dados.

West et al. (2005) comparam a capacidade de generalização de 24 diferentes algoritmos de aprendizado supervisionado com duas bases de dados publicas, com o objetivo de testar a acurácia de seus resultados na previsão da ocorrência do câncer de mama e de identificar suas principais causas. Em todas as execuções, é utilizada validação cruzada para os resultados mediante 100 fragmentos das bases de dados.

Segundo suas conclusões, os modelos cujo treinamento e testes foram mais instáveis (pela diversificação dos conjuntos de dados) atingem melhores níveis de acurácia, reforçando a importância da qualidade dos dados de entrada dos experimentos.

Ainda, os autores afirmam que modelos resultantes de combinações de vários algoritmos obtém maior acurácia na previsão de eventos, em detrimento da seleção de modelos únicos.

Helma e Kazius (2005) organizam um estudo para construção e validação de um modelo preditivo para processos bioquímicos e toxicidade para uso em estudos clínicos, epidemiológicos ou sobre reações adversas, com base nas orientações da OECD (Organization for Economic Cooperation and Developement).

Embora o estudo utilize uma base de dados experimentais, o texto aborda os impactos da baixa qualidade de dados, destacando a duplicidade de registros, a ausência de dados e o erro humano de preenchimento. Os autores citam o nível de agregação dos dados, por interferir nos resultados da Mineração, uma vez que os atributos utilizados podem ser generalizados ou específicos, alterando a definição dos conjuntos de dados por conta da duplicação de valores proporcionada.

A atenção a protocolos regulatórios é evidente, por conta da necessidade em validar os padrões mediante definições formais tanto para os dados quanto para os modelos preditivos, mediante o conhecimento corrente da área específica.

Kobus (2006) aplicou o algoritmo Apriori (AGRAWAL; SRIKANT, 1994) em uma base de pacientes com mais de 40 anos de idade e com ao menos um registro relativo a doenças cardiovasculares entre 2002 e 2004, a fim de identificar associações que indiquem padrões de risco e, portanto, possíveis tratamentos de alta complexidade e alto custo.

Os resultados da Mineração foram classificados pelos especialistas quanto a sua relevância, irrelevância ou insignificância. Destaque para o fato de que regras cuja confiança é de 50%, denotando aleatoriedade (o que as classificaria como irrelevantes sob uma perspectiva objetiva) foram dadas como relevantes pelos especialistas, por conta da combinação de elementos que demonstram.

A autora discorre sobre a qualidade dos dados nos Sistemas de Informação de Saúde, o que pode inviabilizar a utilização da Mineração de Dados com recurso cotidiano na Saúde, por conta do tempo requerido na preparação e limpeza.

Steiner et al. (2006) aplicaram a Mineração de Dados em uma base de 118 pacientes apresentando quadro de Icterícia por câncer ou cálculo biliares. Seus experimentos se basearam no uso de técnicas de análise exploratória dos dados, realizando o pré-processamento para diferentes tarefas de Mineração de Dados visando obter aquela que discriminasse padrões com a máxima acurácia, ampliando o respaldo à tomada de decisão dos especialistas médicos quanto aos seus diagnósticos.

Destacam que a geração de árvores de decisão, obstante ser a técnica menos precisa, é aquela que conquistou maior eficiência junto aos especialistas, dada a facilidade em comunicar os padrões encontrados, deixando claro quais são os atributos discriminadores e seus respectivos pontos de corte quando comparados com os protocolos específicos de diagnóstico e tratamento e mediante julgamento do especialista. Característica interessante, segundo o estudo, para a adoção sistemática da Mineração de Dados no diagnóstico e tratamento da Icterícia.

Ramon et al. (2007) utilizam Mineração de Dados com o objetivo de primeiramente descobrir problemas em 1.548 pacientes de Unidades de Tratamento Intensivo (UTI) e, desta forma, encontrar padrões que apóiem o especialista intensivista no diagnóstico e no tratamento a ser aplicado em tempo real. Os autores buscam, ainda, prever riscos aos quais os pacientes estejam expostos, utilizando algoritmos de Árvore de Decisão, *First Random Forests* e Redes Bayesianas, todos avaliados mediante validação cruzada.

Os experimentos prevêem o tempo de permanência e as chances de sobrevivência, bem como o desenvolvimento de estados de risco de vida, mediante a apropriação de critérios clínicos de avaliação de severidade APACHE II (*Acute Physiology and Chronic Health Evaluation II*), com base na resposta inflamatória sistêmica ou insuficiência renal, acompanhados de modelos específicos de informação clínica.

Os autores relatam dificuldades relativas a ruído nos dados, tamanho das bases de dados, existência de diferentes bases de dados a serem integradas e características

individuais dos pacientes, que apresentam diferentes respostas a medicamentos ou diferentes parâmetros de saúde no momento em que ingressam na UTI.

Stein Junior (2008) aplicou o KDD em uma base de dados sobre riscos em micro regiões na cidade de Curitiba, obtida mediante um questionário de entrevista com especialistas em Saúde Coletiva, do qual foram retirados os atributos que foram processados com o algoritmo J48, com objetivo de especificar um sistema de informações para monitoramento de condições em micro áreas urbanas e apoio ao planejamento de ações para melhoria da qualidade de vida da população.

O autor declara que o fato de ter adotado atributos determinados pelos especialistas entrevistados pode ter causado uma ausência de padrões novos. Reforçando a necessidade de gerar associações que surpreendam o tomador de decisão sem a contaminação dos repertórios do especialista.

O alinhamento das regras extraídas com um protocolo prévio e relativo ao domínio em questão apóia a tomada de decisão. Por exemplo, regras que associavam condições de precariedade da urbanização de uma micro área com baixo risco para a população foram descartadas pelos especialistas, justamente porque o critério de avaliação adotado determina que uma área somente pode ser considerada de baixo risco se não apresentar precariedades.

Bodini Junior (2009) utiliza a base de dados do Sistema de Movimento de Autorização de Internação Hospitalar do Sistema Único de Saúde (AIH/SUS) na busca de *outliers*, por sua dissimilitude ou inconsistência e para uso em auditorias. Destaca que os dados, além de volumosos, são multidimensionais, cujos atributos não podem ser considerados separadamente. O autor coloca a Mineração de Dados como uma técnica indicada para a seleção do escopo de auditoria, uma vez que considera a ótica multidimensional dos dados.

Um aplicativo foi implementado para a execução da tarefa de agrupamento nebuloso. Para a Classificação de Classe Única mediante SVM – One Class, o autor se aproveitou do aplicativo LIBSVM, desenvolvido por Chang e Lin (2001).

Por fim, a tarefa de Mineração é facilitada por um aplicativo com interface gráfica de usuário, que recebe os arquivos de dados, transforma-os e submete-os aos aplicativos responsáveis pelos algoritmos de Mineração. Este aplicativo aplica uma função de similaridade média aos registros normalizados da base de dados, classificando-os a seleção de uma linha de corte que evidencie os registros anômalos.

Dallagassa (2009) aplicou Árvores de Decisão C4.5 a uma base de dados de beneficiários de uma operadora privada de saúde com o objetivo de identificar relações que indicassem a possibilidade de desenvolvimento de Diabetes Melitus do tipo 2, tendo como base o protocolo de diagnóstico deste mal.

As regras encontradas foram classificadas quanto ao seu interesse utilizando as medidas objetivas de interesse Taxa de Acerto e Cobertura, agrupando-as em três classes.

Em seguida, especialistas em Saúde as assinalaram quanto a concordando, concordando parcialmente ou discordando da regra. Atestando, assim, se as regras confirmam ou não seus conhecimentos prévios.

Uma interface para navegação e consulta aos padrões encontrados foi desenvolvida pelo autor e, em conclusão, todas as etapas realizadas são organizadas de forma a propor um processo para desenvolvimento de soluções baseado no KDD.

Kuretzki (2009) integra a Mineração de Dados ao módulo Analisador, do SINPE® (Sistema Integrado de Protocolos Eletrônicos). São construídas duas funcionalidades para a seleção de dados a partir dos protocolos presentes no SINPE, executar a sua classificação mediante o algoritmo ID3 (QUINLAN, 1986) e executar o algoritmo Apriori (AGRAWAL; SRIKANT, 1994) sobre tal base, de forma a integrar os resultados obtidos para o apoio à decisão em saúde.

Conforme sua pesquisa, 53,85% (Cinquenta e três vírgula oitenta e cinco por cento) dos entrevistados consideram excelente a possibilidade de possuir um ferramenta de Mineração de Dados no sistema SINPE, destacando a independência e facilidade de uso e leitura dos padrões por parte dos usuários. A satisfação em relação ao aplicativo é de 80% (oitenta por cento), dos quais, 40% declararam excelência do software.

Meyfroidt et al. (2009) utilizam diferentes algoritmos para avaliação de resultados em tempo real dos tratamentos impostos a pacientes internados em Unidades de Tratamento Intensivo. Apresentando os benefícios de diferentes tarefas de aprendizado supervisionado, a fim de prever o estado geral, exposição a riscos ou tempo de permanência dos pacientes, bem como o auxílio ao planejamento de tratamentos. Os resultados de Árvores de Decisão, Florestas Aleatórios (*Random Forests*), Redes Neurais, Redes Bayesianas, *Support Vector Machines* e Processo Gausiano são comparados com protocolos de aferição clínica quanto às suas porcentagens de acerto ou erro nas previsões e, em todos os casos, vantagens na geração de modelos preditivos são apresentadas e discutidas.

Já que nenhuma das técnicas se mostrou superior às demais, os autores recomendam a aplicação de múltiplos algoritmos sempre que possível, na busca de resultados mais acurados.

Os estudos citam a confidencialidade, a quantidade de dados, a necessidade de integração de diferentes bases de dados, sua organização e qualidade como barreiras no uso de bancos de dados clínicos para pesquisa e desenvolvimento de soluções. Ainda, afirmam que a representação simplificada dos resultados amplia o entendimento por parte do especialista, contribuindo para a adoção cotidiana dos resultados da Mineração de Dados.

Ayed et al. (2010) constroem um sistema de apoio à decisão aplicando o KDD em Unidades de Tratamento Intensivo com o objetivo de apoio aos intensivistas no entendimento das relações causais, prevenção e previsão de infecções hospitalares.

Os autores destacam que o sucesso de um Sistema de Suporte à Decisão baseado no KDD depende da devida análise das necessidades do tomador de decisões, determinação das atividades preparatórias para apoio à decisão, manipulação de dados relevantes e visualização adequada dos resultados. Evidenciando a importância de se correlacionar os dados com as atividades humanas envolvidas na tomada de decisão.

Pregam, portanto, uma abordagem que coloque o usuário no centro do processo de desenvolvimento dos sistemas de apoio à decisão baseados no KDD, planejando e avaliando tanto as atividades quanto os resultados obtidos desde o início e até o final do processo de desenvolvimento, pois um sistema de apoio à decisão difícil de utilizar é geralmente abandonado pelo usuário.

Vianna et al. (2010) aplicam tarefas de Classificação para descoberta das características da mortalidade infantil, formas de evita-la e respectivas causas, em bases de dados de saúde pública no estado do Paraná.

Todo o procedimento foi baseado nos protocolos de saúde infantil e apoiados por especialistas em Saúde para validação dos padrões encontrados.

A integração de dados oriundos de três diferentes bases de dados foi fundamental para realização das tarefas de Mineração de Dados. Em paralelo, a qualidade das informações também é considerada, uma vez que dados inconsistentes foram encontrados e corrigidos.

Le et al. (2011) tratam a incerteza sobre a dosagem e a frequência do tratamento com prótons para o câncer de próstata mediante a Mineração e Dados, e desenvolvem um aplicativo para planejamento de tratamentos em tempo de atendimento, integração os

dados dos pacientes, oriundos de diferentes fontes e a validação dos padrões encontrados conforme os protocolos de conformidade do tratamento com prótons.

O autor cita distribuição dos dados dos pacientes em diferentes sistemas de informação como a dificuldade mais restritiva ao trabalho.

Destaca-se a revisão do protocolo de terapia com prótons para câncer de próstata, o que deu origem a um modelo de informações clínicas e um fluxo de dados específico para o processo em questão.

Carvalho et al. (2012) demonstram o potencial das técnicas de Mineração de Dados aplicando algoritmos de Classificação, descoberta de regras de associação e de agrupamento em uma base de dados clínicos sobre pacientes de fisioterapia a partir de dados sobre perfil profissional e informações clínicas. Os autores propõem que a decisão sobre a terapia aplicada seja apoiada pelos padrões encontrados, uma vez que o perfil da atividade profissional pode inferir os tipos de lesões possíveis, indicando as recomendações de prevenção ou de tratamento.

O artigo destaca as dificuldades relacionadas à qualidade das informações, sejam pela diferença de linguagem para preenchimento de campos textuais abertos, dados incompletos ou informações contraditórias. No contexto estudado pelas autoras, fica evidente a ausência de um processo apoiado por sistemas de informação para o trabalho do especialista em Saúde, o fisioterapeuta neste caso, o que resulta na perda das informações sobre pacientes e sobre a evolução do próprio tratamento.

Martínez e Bermúdez (2012) utilizam Redes Bayesianas e Árvores de Decisão para previsão e diagnóstico médico em doenças cardiovasculares com bases de dados cujos atributos são discretos.

Os autores defendem a combinação de diferentes algoritmos para a aplicação de Mineração de Dados no apoio ao especialista médico. No estudo apresentado, apresentam um aplicativo construído especificamente e defendem a simplicidade de navegação nas regras encontradas para o especialista em Saúde, uma vez que este não tem condições de dedicar muito tempo à análise dos padrões.

Concluem que o auxílio ao planejamento de tratamentos pode ser apoiado com Mineração de Dados desde que mediante a representação visual dos padrões de maneira personalizada e simplificada.

Moghimi et al. (2012) constroem um sistema de apoio à decisão baseado na análise de riscos e previsão de resultados em cirurgias para implante de quadril e joelho, tendo como base o KDD, levando em conta que a decisão em realizar tais procedimentos

deve considerar várias dimensões e os riscos sobre a qualidade de vida dos pacientes, a inovação das técnica a ser empregada no implante, as condições de saúde do paciente, especialmente dos ossos a serem operados e os custos relativos ao tratamento em questão.

Para tanto, os autores adotam o fluxo de trabalho em cirurgias ortopédicas e a estratégia de encontrar as relações entre os fatores de riscos em si, mas também entre os riscos cirúrgicos e os resultados dos procedimentos, implementando um sistema de detecção de riscos em tempo real, baseado em Redes Neurais e Regras de Associação.

Os dados relativos aos resultados após as cirurgias são adicionados ao banco de dados e os padrões são atualizados por conta da atualização da base de dados.

Cruz-Ramírez et al. (2013) aplicam algoritmos evolucionários com multiobjetivos e Redes Neurais para a determinação das melhores combinações doador-recipiente e para a previsão do sucesso e sobrevida em pacientes como suporte à decisão em procedimentos de transplante de fígado.

Os estudos foram conduzidos em bases de dados relativas a 11 unidades de transplante de fígado espanholas e recipientes com idade acima de 18 anos que receberam transplantes de doadores mortos, cujos dados foram obtidos durante o processo de transplante.

Os autores adotam um modelo de informações que represente os dados tanto para avaliação de compatibilidade entre doador e recipiente, quanto para a avaliação do estado de saúde do doador, o que permitiu o acompanhamento dos resultados dos transplantes até o decorrer de um ano após o transplante, a perda do enxerto por rejeição ou a morte do recipiente.

Os pontos de atenção indicados em cada publicação estão organizados no Quadro 1.

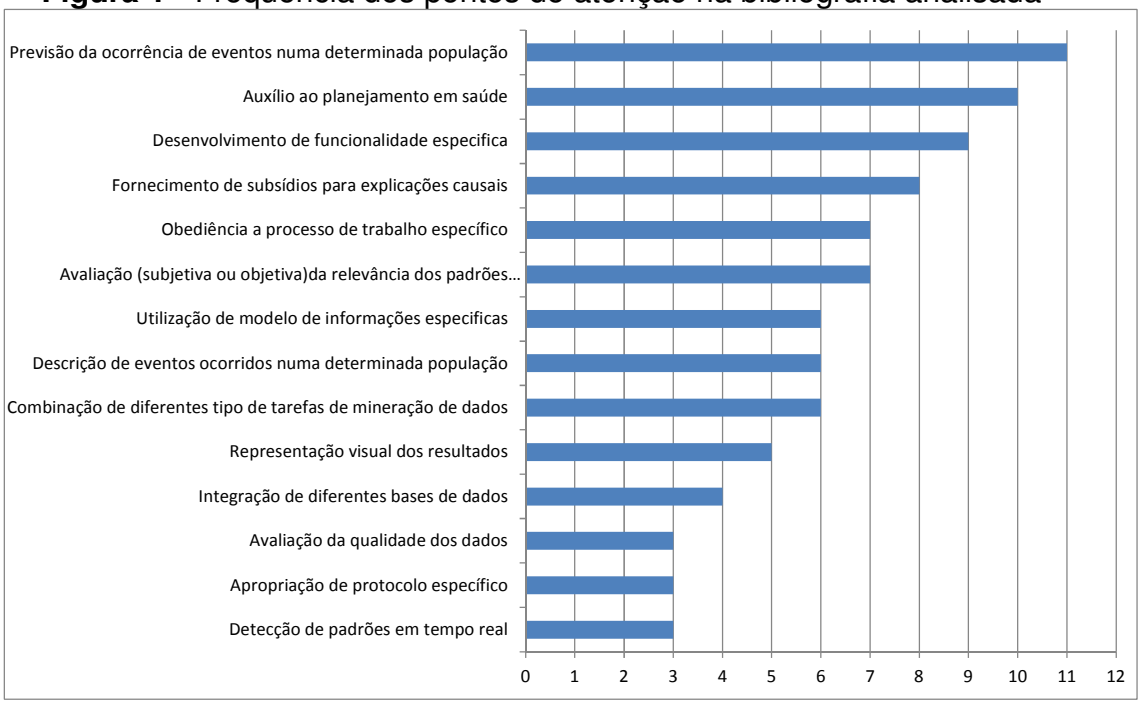
Quadro 1 - Pontos de atenção para o uso efetivo de resultados da Mineração de Dados conforme as publicações analisadas

Pontos de atenção	West et al (2005)	Helma e Kasius (2006)	Kobus (2006)	Steiner et al. (2006)	Ramon et al (2007)	Stein Junior (2008)	Bodini Junior (2009)	Dallagassa (2009)	Kuretski (2009)	Meyfroidt (2009)	Ayed et al (2010)	Vianna et al (2010)	Lee at al (2011)	Carvalho et al (2012)	Martínes e Bermúdez (2012)	Moghimi et al (2012)	Trindade (2012)	Cruz-Ramírez (2013)
Apropriação de protocolo específico		X			X			X		X		X	X					
Avaliação da qualidade dos dados		X	X		X					X		X		X			X	
Integração de diferentes bases de dados					X					X		X		X				
Desenvolvimento de funcionalidade especifica							X		X		X		X		X	X		
Obediência a processo de trabalho específico				X									X			X		
Utilização de modelo de informações especificas		X			X	X							X					X
Avaliação (subjetiva ou objetiva) da relevância dos padrões encontrados			X	X				X										
Combinação de diferentes tipo de tarefas de mineração de dados	X						X		X					X	X	X	X	
Detecção de padrões em tempo real					X					X	X							
Descrição de eventos ocorridos numa determinada população			X	X			X	X	X		X	X	X	X			X	
Previsão da ocorrência de eventos numa determinada população	X	X			X	X		X		X	X			X	X	X		X
Fornecimento de subsídios para explicações causais		X	X		X	X		X		X	X				X	X		
Auxílio ao planejamento em saúde			X		X			X		X					X	X	X	X
Representação visual dos resultados	X							X		X	X				X		X	

Fonte: Produzido pelos autores.

A partir da figura 1 é possível observar a frequência absoluta da presença dos pontos de atenção nas publicações analisadas.

Figura 1 - Frequência dos pontos de atenção na bibliografia analisada



Fonte: Produzido pelos autores

Extração de Regras de Associação. Os dados foram tratados com o algoritmo *Apriori* (AGRAWAL; SRIKANT, 1994), e foram obtidas 39114 regras de associação.

Pós processamento. O conjunto de regras de associação foi tratado com o algoritmo DRE (MILANI; CARVALHO, 2013), obtendo-se 345 regras gerais acompanhadas de suas respectivas regras de exceção.

Dentre os pares de regras gerais e respectivas exceções encontradas, é possível destacar que o ponto de atenção Detecção de Padrões em Tempo Real não relacionado com o Auxilio ao Planejamento em Saúde:

Se Auxilioplanejamentosaude = SIM *então* Deteccaotemporeal = NÃO

(Suporte = 44,4%; Confiança = 75%)

Entretanto, a regra acima possui 88 exceções, identificando que a detecção de padrões em tempo real está presente nos relatos que incluíram integração de

diferentes bases, avaliação da qualidade dos dados, apropriação de protocolo específico, explicações causais e previsão de eventos, como demonstrado no Quadro e 3.

Quadro 2 - Frequência dos pontos de atenção associados à detecção de padrões em tempo real

Pontos de Atenção	**Frequência nas regras de exceção**
Integração de diferentes bases = SIM	52%
Avaliação da qualidade dos dados = SIM	36%
Apropriação de protocolo específico = SIM	34%
Explicações Causais = SIM	24%
Previsão de Eventos = SIM	22%

Produzido pelos autores

Quadro 3 - Exemplos das Regras de exceção envolvendo detecção de padrões em tempo real

SE Explicações causais = SIM e Auxilio ao planejamento em saúde = SIM ENTÃO Detecção de padrões em tempo real = SIM (Suporte 11,1%; Confiança 100,0%)
SE Avaliação da qualidade dos dados = SIM e Apropriação de protocolo específico = SIM e Auxilio ao planejamento em saude = SIM ENTÃO Detecção de padrões em tempo real = SIM (Suporte 11,1%; Confiança 100,0%)

Produzido pelos autores

Muito embora a Representação Visual dos Resultados não seja citada por 12 dos 18 estudos analisados, as regras de exceção obtidas demonstram que está presente quando explicações causais é um ponto de atenção citado pelos autores, estando presente em 45% das regras de exceção encontradas. (

Quadro 2 e 5).

Quadro 2 - Frequência dos pontos de atenção que levam á presença da representação visual dos resultados

Pontos de atenção	Frequência nas regras de exceção
Explicações causais SIM	45%
Previsão de eventos SIM	27%
Diferentes tarefas de DM SIM	24%
Apropriação de protocolo SIM	18%
Detecção em tempo real SIM	17%
Descrição de eventos ocorridos SIM	5%
Auxilio ao planejamento em saúde SIM	2%
Avaliação da qualidade dos dados SIM	1%

Produzido pelos autores

Quadro 3 - Exemplos das Regras de exceção envolvendo detecção de Representação visual dos resultados

SE Modelo de informações especificas = NÃO e Diferentes tarefas de DM = NÃO e Explicações causais = SIM e Previsão de eventos = SIM ENTÃO Representação visual dos resultados = SIM (Suporte 16,7%; Confiança 100,0%)
SE Modelo de informações especificas = NÃO e Diferentes tarefas de DM=NÃO e Explicações causais=SIM ENTÃO Representação visual dos resultados=SIM (Suporte 22,2%; Confiança 75,0%)

Produzido pelos autores

4 DISCUSSÃO

A predominância de publicações encontradas (87%), que relatam a adoção de Mineração de Dados para a área clínica, evidencia um espaço importante da utilização desta técnica na rotina clínica para potencializar a eficiência do tratamento. Obstante a baixa frequência (23%) de relatos usando a Mineração de Dados na gestão da Saúde, esta pode proporcionar benefícios sobre critérios para a aplicação de recursos e elaboração de estratégias para melhor entender a população.

Por outro lado, o entendimento dos pontos de atenção para o uso da Mineração de Dados na Saúde permite a elaboração de táticas para aumento da eficiência e maior aproveitamento dos benefícios gerados. A necessidade por cuidados criteriosos na elaboração e execução de projetos de Mineração de Dados na Saúde fica evidente à medida que os materiais pesquisados demonstram que,

assim como os padrões descobertos, a forma com que tais padrões são comprovados ou disponibilizados para os especialistas influencia sua aceitação e uso rotineiro. A existência de elementos específicos da Área da Saúde requer que requisitos de conformidade, qualidade ou de usabilidade sejam atendidos de maneira a aumentar a possibilidade de adoção intensiva dos resultados da Mineração de Dados.

Ainda que a descrição de padrões e a previsão de eventos sejam a natureza da Mineração de Dados, o fato de 17 estudos relatarem ambos como pontos de atenção mostra que o apoio à decisão em saúde requer aplicações que combinem capacidades descritivas e preditivas para a devida colaboração para o especialista, enfatizando possíveis causas, relações e consequências.

A prevalência dos objetivos primários da Mineração de Dados cercados por pontos relativos à representação visual dos resultados ou à detecção de padrões em tempo real deixa evidente que os algoritmos disponíveis ainda não atendem plenamente aos critérios de aceitação da área da Saúde. Um exemplo claro desta limitação está na necessidade de representar os resultados visualmente, mostrando que a saída dos algoritmos não é adequada para o contexto e exige tratamento posterior. Por outro lado, a existência de pontos de atenção relativos à integração de diferentes bases ou a avaliação da qualidade dos dados, também sugere que as entradas dos algoritmos atuais não são suficientemente implementadas, requerendo um esforço na seleção, avaliação da qualidade e preparação dos dados.

Tais limitações impõem ao processo de Mineração de Dados maior custo operacional, pela necessidade de mais trabalhos e mais tempo, tanto para a preparação dos dados – pré-processamento – quanto para a representação dos resultados – pós processamento e, ao mesmo tempo, abrem oportunidades para a construção de algoritmos ou até mesmo ferramentas para pré e pós processamento, ampliando as funcionalidades e otimizando o processo, reservando tempo e espaço para a análise e interpretação dos padrões, contribuindo para a tomada de decisão mais eficiente.

O ponto de atenção Auxílio ao Planejamento em Saúde, por sua vez, mostra que a Mineração de Dados pode contribuir o ganho de eficiência e eficácia na decisão a médio e longo prazos, baseada em grandes volumes de informação.

O ponto Desenvolvimento de Funcionalidade Específica indica uma forte tendência a implantação de Sistemas de Apoio à Decisão elaborados

especificamente e de forma amigável para o especialista em Saúde, aproximando-o dos possíveis benefícios da Mineração de Dados. Mas, esta possibilidade requer do pesquisador e desenvolvedor da solução atenção aos Protocolos Específicos, que orientam os procedimentos clínicos e ainda validam os resultados obtidos na própria Mineração, representando um desafio no entendimento mais profundo da área de aplicação.

A apropriação de protocolos específicos também indica a necessidade da área da Saúde em seguir princípios e procedimentos amplamente aceitos em cada especificidade médica. Este ponto de atenção, além de servir com base para testes e validações dos resultados, traz o desavio para o pesquisador e para o especialista de não permitirem que o conhecimento corrente determine os resultados, impedindo, assim, o surgimento de conhecimentos surpreendentes, mas sim, facilitando a validação dos modelos, sejam descritivos ou preditivos. Ou seja, o cuidado em não viciar os padrões obtidos com a Mineração de Dados amplia a complexidade da aplicação na área.

Muito embora os pontos de atenção Detecção de Padrões em Tempo Real e Apropriação de Protocolo Específico tenham as menores frequências nas publicações analisadas, o que evidencia afastamento da rotina médica, sua presença mostra que, além de importantes para a Mineração de Dados na Saúde, existem estudos para aproximar o KDD do dia a dia da Área da Saúde, indicando ricas possibilidades para exploração.

A associação entre os pontos de atenção identificados serve como base para estratégias que ampliem os benefícios da Mineração de Dados na Saúde mediante sua combinação. Por exemplo, a forte associação entre Detecção de Padrões em Tempo Real, Integração de Diferentes Bases de Dados e Avaliação da Qualidade dos Dados permite antecipar quais tarefas serão realizadas durante o processo de descoberta de conhecimentos e quais funcionalidades devem ser implementadas para acurácia e maior aceitação e uso na Área da Saúde.

Partindo do pressuposto de que a decisão pela combinação de pontos de atenção pode ser apoiada pela associação levantada neste estudo, é possível induzir uma série de alternativas, compondo diferentes caminhos para o aumento da efetividade da Mineração de Dados na Saúde.

5 CONCLUSÃO

Utilizar a Mineração de Dados para analisar os grandes volumes de informações clínicas disponíveis pode auxiliar os especialistas em Saúde na melhora das condições dos pacientes e no ganho de eficiência no uso dos recursos médicos. Entretanto, mesmo sendo amplamente aplicada a problemas médicos, há uma necessidade evidente em lançar mão de estratégias e boas práticas que aumentem as chances da Mineração de Dados ser adotada na rotina médica.

Ainda que a adoção de Sistemas de Informação na Saúde seja evidente, a avalanche de informações clínicas geradas requer apoio computacional para a tomada de decisões. Há fortes evidências na literatura sobre a eficiência da Mineração de Dados na descoberta de conhecimentos, proporcionando eficiência, melhores tratamentos e ganho de qualidade de vida para os pacientes.

O planejamento efetivo de ações de prevenção ou de tratamento, mediante modelos que permitam ao especialista avaliar alternativas previamente (e automaticamente) elaboradas e indicar resultados futuros com alta acurácia, com explicações causais em tempo real, mais complexas e, por isso, mais completas, agiliza a resposta clínica, aumenta as chances de cura e reduz riscos para os pacientes. Embora seja possível identificar pontos de atenção para o uso da Mineração de Dados na Saúde, a aplicação de tais questões está ligada ao contexto e deve ser detalhadamente planejada e avaliada.

Por outro lado, o uso de Regras de Exceção para obter a associação entre os pontos identificados permite traçar estratégias mais complexas, atendendo um maior número de requisitos, para o uso rotineiro da Mineração de Dados.

Embora haja inúmeras experiências e relatos que indiquem quais podem ser os pontos de atenção para o sucesso no uso da Mineração de Dados na Saúde, fica evidente a ausência de um guia ou conjunto de padrões práticos para tal fim, o que denota uma série de oportunidades ricas e interessantes para a pesquisa sobre a aceitação de aplicação da Mineração de Dados em sistemas de apoio à decisão médica.

REFERÊNCIAS

AGRAWAL, Rakesh; SRIKANT, Ramakrishnan. Fast algorithms for mining association rules in large databases. In: International conference on very large data bases, 20., 1994, San Francisco. **Proceeding…** San Francisco: Morgan Kaufmann, 1994. p. 478-499.

AYED, Mounir Ben et al. A user-centered approach for the design and implementation of KDD-based DSS: a case study in the healthcare domain. **Decision Support Systems**, Amsterdam, v. 50 p. 64-78, 2010.

BLOMBERG, Luciano Costa. **Gestão de métricas e indicadores de doenças em saúde bucal suportado por um ambiente de descoberta de conhecimento em banco de dados**. 2010. 107f. Dissertação (Mestrado) - Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre.

BODINI JUNIOR, Antonio Carlos. **Utilização de técnicas de data mining na detecção de outliers em auxílio à auditoria operacional com um estudo de caso com dados do sistema de informações hospitalares**. 2009. 122f. Tese (Doutorado) - Universidade Federal do Rio de Janeiro, Rio de Janeiro.

CARVALHO, Deborah et al. Mineração de dados aplicada à fisioterapia. **Fisioterapia e Movimento**, Curitiba, v. 25, n. 3, p. 595-605, jul./set. 2012.

CRUZ-RAMÍREZ, Manuel et al. Memetic pareto differential evolutionary neural network used to solve an unbalanced liver transplantation problem. **Soft Computing**, New York, v. 17, n. 2, p. 275-284, 2013.

DALLAGASSA, Marcelo Rosano. **Concepção de uma metodologia para identificação de beneficiários com indicativos de diabetes melitus tipo 2**. 2009. 105 f. Dissertação (Mestrado) - Pontifícia Universidade Católica do Paraná, Curitiba.

FAYYAD, Usama; PIATETSKI-SHAPIRO, Gregory; SMYTH, Padhraic. The KDD process for extracting useful knowledge from volumes of data. **Communications Of The Acm**, New York, v. 39, n. 11, p. 27-34, 1996.

HELMA, Christoph; KAZIUS, Jeroen. Artificial intelligence and data mining for toxicity prediction. **Current Computer Aided Drug Design**, San Francisco, v. 2, n. 2, p. 123-133, dez. 2005.

HUSSAIN, Farhad et al. Exception rule mining with relative interestingness measure. **PAKDD**, Kyoto, n. 1, v.1805, p. 86-97, 2000.

KOBUS, Luciana Schleder Gonçalves. **Aplicação da descoberta de conhecimento para identificação de usuários com doenças cardiovasculares elegíveis para programas de gerenciamento de caso**. 2006. 145f. Dissertação (Mestrado) - Pontifícia Universidade Católica do Paraná, Curitiba.

KURETZKI, Carlos Henrique. **Técnicas de mineração de dados aplicadas em bases de dados da saúde a partir de protocolos eletrônicos**. 2009. 98f. Dissertação (Mestrado) - Universidade Federal do Paraná, Curitiba.

LE Anh H et al. Intelligent epr system for evidence-based research in radiotherapy: proton therapy for prostate cancer. **International Journal Of Computer Assisted Radiology And Surgery**, Heidelberg, v. 6, n. 6, p. 769-784, 2011.

MARISCAL, Gonzalo; MARBÁN, Óscar; FERNANDEZ, Covadonga. A survey of data mining and knowledge discovery process models and methodologies. **The Knowledge Engineering Review**, Cambridge, v. 25, n. 2, p. 137 – 166, 2010.

MARTÍNEZ, Guillermo Roberto Salarte; BERMÚDEZ, Yanci Viviana Castro. Modelo híbrido para el diagnóstico de enfermedades cardiovasculares basado en inteligência artificial. **Tecnura**, Colombia, v. 16, n. 33, p. 35-53, 2012.

MEYFROIDT Geert et al. Machine learning techniques to examine large patient databases. **Best Practice & Research Clinical Anaesthesiology**, Amsterdam, v. 1, n. 23, p 127–143, 2009.

MILANI, Cristian. Simioni; CARVALHO, Deborah Ribeiro. Pós-processamento em kdd. **Revista de Engenharia e Tecnologia**, Ponta Grossa, v. 5, n. 2, p. 151-162, 2013.

MOGHIMI, Hoda et al. Incorporating intelligent risk detection to enable superior decision support: the example of orthopaedic surgeries. **Heath and Techonology**, Berlin, v. 2, n. 1, p. 33-41, 2012.

QUINLAN, John Ross. **C4.5**: programs for machine learning. São Francisco: Morgan Kaufmann Publishers, 1993

QUINLAN, John Ross. **Introduction to decision trees**. São Francisco: Machine Learning, 1986.

RAMON Jan. et al. Mining data from intensive care patients. **Advanced Engineering Informatics**, Oxford, v. 21, n. 3, p. 243–256, 2007.

STEIN JUNIOR, Altair Von. **Descoberta de regras por meio de kdd para a classificação de micro áreas homogêneas de risco**. 2008. 106f. Dissertação (Mestrado) - Pontifícia Universidade Católica do Paraná, Curitiba.

STEINER, Maria Terezinha Arns et al. Abordagem de um problema médico por meio do uso do processo KDD com ênfase em análise exploratória dos dados. **Revista Gestão e Produção**, Florianópolis, v. 13, n. 2, p. 325-337, 2006.

TRINDADE, Carla Machado et al. Technology in health: knowledge discovery in public health databases: study of viral hepatitis in the state of Paraná, Brazil. **Iberoamerican Journal of Applied Computing**, Ponta Grossa, v. 2, n. 2, 2012.

VIANNA, Rossana. Mineração de dados e características da mortalidade infantil. **Caderno Saúde Pública**, Rio de Janeiro, v. 26, n. 3, p. 535-542, 2010.

WEST, David et al. Ensemble strategies for a medical diagnostic decision support system: a breast cancer diagnosis application. **European Journal of Operational Research**, Amsterdam, v. 162, n. 2, p. 532-551, 2005.

WITTEN, Ian; FRANK, Eibe; HALL, Mark. **Data mining**: practical machine learning tools and techniques. San Francisco: Morgan Kaufmann, 2011.

Title

Attention points in data mining for decision support in health

Abstract

Introduction: Data Mining is an alternative to support the decision making process, but it is not a regular practice in health care activities.
Objective: Review of the literature to identify points that deserve attention in the use of Data Mining in Health Care.
Methodology: These points were obtained by analysis of publications in journals, and organized in cross-reference tables. Association rules and exceptions were discovered.
Results: Fourteen points were identified, forming a set of data from which 345 general rules and their respective exceptions were obtained. The points most often cited were “prediction of events” and “support to planning” and the least cited points were “appropriation of specific protocol” and “real-time pattern recognition”. An association was observed between “support to planning” and “non real-time recognition of patterns”. An exception occurs when “support to planning” is cited along with “causal explanations”, being associated to “real-time pattern recognition”.
Conclusions: The points of attention indicate criteria for the adoption of Data Mining in Health Care. The point “causal explanations” deserves mention. It was cited in 8 articles and determines exceptions in the associations among other points, indicating that the selection of the mining task involves the adaptation to users’ expectations.

Keywords: Data Mining in Health Care. Points of attention.

Título

Puntos de atención en minería de datos de apoyo a las decisiones en salud

Resumen

Introducción: La Extracción de Datos representa una alternativa para apoyar la decisión, pero no constituye práctica regular en las actividades de la Salud.
Objetivo: Revisar la literatura, identificando puntos de atención en la utilización de la Extracción de Datos en el área de la Salud.

Metodología: Los puntos de atención fueron obtenidos mediante el análisis de publicaciones en periódicos, sistematizados en cuadros de referencia cruzada y fueron descubiertas de reglas asociación y respectivas excepciones.
Resultados: Fueron identificados 14 puntos, formando un conjunto de datos de los que se obtuvieron 345 reglas generales y las respectivas excepciones. Los puntos más citados son "previsión de eventos" y "auxilio a la planificación" y los menos citados son "apropiación de protocolo específico" y "detección de estándares en tiempo real". Se percibe asociación entre el "auxilio a la planificación" y la "no detección de estándares en tiempo real". Excepción ocurre, cuando el "auxilio a la planificación" es citado juntamente con "explicaciones causales", asociándose a la "detección de estándares en tiempo real".
Conclusiones: Los puntos de atención indican criterios para la adopción de la Extracción de Datos en la Salud. Se destacan "explicaciones causales" que, citado en 8 artículos, determina excepciones en las asociaciones entre los demás puntos, indicando que la selección de la tarea de extracción pasa por la adecuación a las expectativas de los usuarios.

Palabras clave: Extracción de Datos en la salud. Puntos de atención

Recebido em: 28.10.2013
Aceito em: 17.02.2014