ISSN 0103-5150
Fisioter. Mov., Curitiba, v. 25, n. 3, p. 595-605, jul./set. 2012

Mineração de Dados aplicada à fisioterapia

Data Mining applied to physiotherapy

Deborah Ribeiro Carvalho[a], Auristela Duarte Moser[b], Verônica Andrade da Silva[c], Marcelo Rosano Dallagassa[d]

[a] Doutora em Computação pela Universidade Federal do Rio de Janeiro (UFRJ), docente dos cursos de Ciência e Engenharia da Computação da Pontifícia Universidade Católica do Paraná (PUCPR), Curitiba, PR - Brasil, e-mail: ribeiro.carvalho@pucpr.br
[b] Fisioterapeuta, doutora em Ergonomia pela Universidade Federal de Santa Catarina (UFSC), docente do curso de Fisioterapia da Pontifícia Universidade Católica do Paraná (PUCPR), Curitiba, PR - Brasil, e-mail: auristela.lima@gmail.com
[c] Discente do curso de Fisioterapia da Pontifícia Universidade Católica do Paraná (PUCPR), Curitiba, PR - Brasil, e-mail: veronicalirio@hotmail.com
[d] Mestre em Tecnologia Aplicada em Saúde pela Pontifícia Universidade Católica do Paraná (PUCPR), Curitiba, PR - Brasil, e-mail: dallagassa@unimedpr.com.br

Resumo

Introdução: Com o aumento da quantidade de dados armazenados na prática da Fisioterapia e da área de saúde em geral, amplia-se, também, a possibilidade de obtenção de informações importantes no apoio ao processo decisório dos profissionais de saúde. Porém, muitas vezes, o volume de dados gerados é tão grande que dificulta sua utilização, demandando processos mais sofisticados para a manipulação de tais dados. **Objetivo**: Este artigo se propõe a apresentar e discutir o potencial de utilização do processo KDD sobre um conjunto de dados de acompanhamento fisioterapêutico de pacientes, bem como sua utilidade na tomada de decisões terapêuticas ou profiláticas. **Metodologia**: Selecionou-se um subconjunto de dados, referentes a prontuários disponíveis em uma clínica de fisioterapia, do qual foram extraídos três grandes grupos-alvo de tarefas de Mineração de Dados: associação, classificação e agrupamento, explicitados no texto. **Resultados**: Foram extraídos padrões a partir dos dados, de tal forma que se permitisse ao leitor entender passo a passo o processo, ampliando sua compreensão dos resultados obtidos. Foram descobertos padrões em diversos formatos, os quais evidenciaram as possíveis relações entre as variáveis disponíveis. Em seguida, não apenas os padrões foram discutidos, mas, também, a importância da qualidade dos dados coletados. **Conclusões**:

As etapas de classificação, descoberta de regras de associação e agrupamento dos dados oportunizou melhor entendimento das especificidades de pacientes atendidos pela clínica em questão, ampliando, assim, o conhecimento do profissional na identificação das condutas a serem adotadas.

Palavras-chave: Descoberta de conhecimento. Mineração de Dados.
Acompanhamento de processos fisioterapêuticos. Apoio à decisão.

Abstract

Introduction*: With the increasing amount of data stored in the practice of physiotherapy and health area in general, expands the possibility of obtaining important information to decision support of health professionals. However, many times the volume of generated data is so great that their use is difficult, requiring more sophisticated procedures for data manipulation.* ***Objective****: This article aims to present and discuss the potential use of the KDD process on a set of monitoring data for physical therapy patients, as well as its usefulness in decision-making therapeutic or prophylactic.* ***Methods****: We selected a subset of data, referring to records available in a physical therapy clinic, from which were extracted three major groups of data mining tasks: association, classification and clustering.* ***Results****: Knowledge was extracted from the data in such a way that allows the reader to understand step-by-step process, broadening their understanding of the results. Knowledge was discovered in various formats, which showed the possible relationships among the variables available. Not only the knowledge was discussed, but also the importance of quality of data collected.* ***Conclusions****: The tasks of classification, association rules and clustering allowed a better understanding of the patient's characteristics seen by the clinic in question, thus expanding the knowledge of professionals in the identification of actions to be adopted.*

Keywords*: Knowledge discovery. Data Mining. Process monitoring in physiotherapy. Decision support.*

Introdução

A "era da informação" é caracterizada pela crescente expansão no volume de dados gerados e armazenados (1), situação que também é identificada quando do acompanhamento dos pacientes em atendimentos fisioterapêuticos. Se considerarmos os dados referentes a cadastro pessoal, diagnóstico clínico e fisioterapêutico, anamnese, exame físico, exames complementares e evolução, eles geram aproximadamente 80 variáveis por paciente. Considerando que para os dois primeiros conjuntos de dados existe uma única ocorrência por paciente e que os demais podem ocorrer várias vezes, percebe-se o grande volume de dados gerados e armazenados.

Esse grande volume de dados demanda técnicas e ferramentas que, com eficiência, potencializem a transformação dos dados em informação útil e oportuna, viabilizando sua efetiva utilização pelo fisioterapeuta, como apoio na identificação das melhores práticas a serem adotadas no tratamento em questão. Vale destacar que as boas práticas estão fortemente dependentes do estabelecimento do diagnóstico.

Para potencializar a recuperação e o uso desses dados, uma das alternativas é o processo de KDD – *Knowledge Discovery in Database*, o qual permite descobrir relações entre os dados armazenados mais facilmente do que com as técnicas tradicionalmente utilizadas. Pesquisadores da área de Computação têm pesquisado e desenvolvido diversos métodos e programas computacionais que, constantemente, são incorporados ao processo KDD.

O processo KDD foi originalmente desenvolvido com objetivos comerciais, mas tem sido amplamente adotado por outras áreas, inclusive a da saúde, podendo ser identificados vários exemplos, tais como os seguintes:

- Kobus (2) propôs um modelo, baseado em KDD, para apoiar a indicação de usuários de planos de saúde com doenças cardiovasculares elegíveis para o ingresso em programas de gerenciamento de casos.
- Lopes (3) comparou o desempenho entre métodos de descoberta de padrões mais simples *versus* métodos mais sofisticados, sobre 15 bases de dados da saúde.

- Vianna et al. (4) integraram bases de dados de três diferentes sistemas de informação (SINASC, SIM e SIMI), no período de 2000 a 2004, demonstrando não apenas a viabilidade, mas, também, a utilização do KDD, minimizando, assim, a subjetividade (o viés) do gestor no contexto da mortalidade infantil.
- Von Stein et al. (5) adotaram o KDD para apoiar as equipes de saúde na identificação e na delimitação das microáreas de risco, objetivando priorizar ações com vistas a modificar suas condições de vida.
- Dallagassa (6) propõe uma metodologia, baseada em KDD, para a detecção precoce de pacientes com propensão a determinadas doenças. As experimentações realizadas validam a metodologia para a predição de pacientes com diabetes mellitus tipo 2.
- Kuretzki (7), objetivando aperfeiçoar o uso dos dados coletados a partir do Sistema Integrado de Protocolos Eletrônicos (SINOE), propôs a utilização do KDD para avaliar até que ponto pacientes que, no exame físico, apresentaram sinal de Blumberg, passando por cirurgia do tipo apendicectomia com dor abdominal, estão propensos a terem apendicite aguda etc.
- Fengyinga et al. (8), a partir da análise de agrupamento, avaliam a segurança alimentar em províncias chinesas, bem como o estado de pobreza, entre 2002 e 2007, utilizando as perspectivas de disponibilidade de alimentos, acesso, consumo, nutrição e vulnerabilidade à insegurança alimentar.
- Machado (9) adotou o processo KDD para a identificação de características desenvolvimentais em relação ao desenvolvimento humano, com base em estudos sobre a imagem corporal interna.

Esse processo é composto de diversas etapas, que podem ser agrupadas em três grandes grupos: a preparação dos dados; a Mineração de Dados propriamente dita; e o pós-processamento dos resultados obtidos pela mineração. Na etapa de pré-processamento, o foco principal está em sistematizar as diversas fontes de dados e compatibilizá-los. O pós-processamento contempla a depuração e/ou a síntese das relações descobertas, que usualmente são denominadas de "padrões descobertos". Na maioria das experimentações KDD, a etapa de pós-processamento se justifica, pois o volume de conhecimento descoberto é tão grande que dificulta sua análise, chegando ao limite de inviabilizar seu uso no apoio à tomada de decisão. Isso se deve a vários fatores, como padrões redundantes ou relações irrelevantes, entre outros (10).

Na etapa de Mineração de Dados, podem ser identificados programas computacionais correspondentes a três tarefas principais: classificação, descoberta de regras de associação e agrupamento, as quais são comentadas a seguir.

O principal objetivo da classificação é que o programa computacional construa um classificador que represente a descoberta de algum tipo de relação entre as variáveis disponíveis no conjunto de dados de entrada. Nessa tarefa, deve ser identificada, entre as diversas variáveis de entrada, aquela que representa a classe a ser predita pelo classificador, a qual se denomina "variável classe" (11).

Segundo Breiman et al. (12), um classificador extraído de um conjunto de dados serve a dois propósitos: predição de um valor e entendimento da relação existente entre aquela indicada como a classe e as demais variáveis do conjunto disponível. Como exemplo, pode-se ter uma aplicação na área da saúde, na qual um médico poderia classificar alguns de seus pacientes em duas classes: "tem" ou "não tem" determinado diagnóstico. Para cumprir o segundo propósito, é exigido do classificador que ele não apenas classifique, mas, também, explicite as relações entre as variáveis, baseado em conhecimentos extraídos da base de dados, de forma compreensível.

Um classificador permite verificar a relação entre variáveis como profissão, idade e diagnóstico. Dessa forma, todas as relações descobertas ajudarão a caracterizar o diagnóstico (variável classe) com base nas demais disponíveis. Um exemplo dessa afirmação será apresentado no Quadro 2, tópico "Resultado".

Na segunda tarefa – descoberta de regras de associação –, os padrões descobertos são apresentados na forma de regras do tipo $A \rightarrow C$ (lidas como: SE (A), ENTÃO (C)), em que A e C representam, respectivamente, o Antecedente e o Consequente da regra. Ao contrário da tarefa de classificação, para o processo de descoberta de regras de associação não se elege uma variável como sendo a variável classe a ser predita, ou seja, todas as regras descobertas podem associar indistintamente os atributos disponíveis no conjunto de dados de entrada. Por exemplo, uma regra de associação descoberta poderia ser: SE fraqueza muscular,

ENTÃO encurtamento muscular. Em geral, os programas computacionais disponíveis para descobrir regras de associação permitem identificar o quanto o padrão descoberto é representativo no conjunto de dados oferecido como entrada. Dentre as diversas formas de avaliar as regras descobertas, pode-se utilizar a probabilidade. Por exemplo, 18,8% dos indivíduos representados nos dados apresentam fraqueza muscular; dentre eles, em torno de 1/3 (66,7%) também apresentam encurtamento muscular.

A tarefa de agrupamento de dados procura agrupar o conjunto de registros disponíveis em dois ou mais grupos, com base em alguma medida de proximidade ou de semelhança entre os valores das variáveis que constituem cada um dos registros. Dá-se o nome de "modelo de agrupamento", ou simplesmente "agrupamento", ao conjunto de grupos formados a partir dos registros da base de dados oferecida como entrada para o programa computacional. O objetivo é que registros pertencentes a um mesmo grupo sejam bastante similares e, ao mesmo tempo, distintos de registros pertencentes a outros grupos.

Qual seria uma possível aplicabilidade da criação de grupos a partir dos registros de dados disponíveis? Por exemplo, um conjunto de pacientes pode ser agrupado em vários grupos baseados nas similaridades dos seus sintomas, e os sintomas comuns aos pacientes de cada grupo podem ser usados para descrever a qual grupo um novo paciente pertencerá. Assim, um dado paciente seria inserido em um grupo cujos pacientes têm sintomas similares aos dele. Dessa forma, a tarefa de agrupamento tem como resultado a criação de uma nova variável que represente o grupo (ou classe) ao qual o registro mais se assemelhe.

Existem alguns experimentos de KDD que relacionam a tarefa de agrupamento com a classificação, sendo a etapa de agrupamento uma fase anterior à de classificação quando, originalmente, não se tem no conjunto de dados uma variável que represente naturalmente a classe na qual o registro se enquadre. Assim, a variável classe pode ser criada, a partir da identificação dos grupos, para posterior busca pela explicação sobre os critérios determinantes, envolvendo as variáveis para a criação dos grupos com base na descoberta de classificadores (13).

Muitas vezes, o conjunto de padrões/regras descobertos é tão grande que chega a inviabilizar a análise por parte do especialista, demandando formas que facilitem esse trabalho. Uma das formas mais simples de pós-processar as regras descobertas é a eliminação das regras ditas redundantes. Por exemplo, na eventualidade do algoritmo, descobrir duas regras, tais como:

$A \rightarrow C$ (Se A, então C);

$A, B \rightarrow C$ (Se A e B, então C).

Nesse caso, a segunda regra não acrescenta nenhum conhecimento em relação à primeira, e, dessa forma, o especialista poderia apenas se limitar à avaliação da primeira, evitando, assim, um sobre-esforço em avaliar ambas as regras.

Hussain et al. (14) apresentam outro método, o qual objetiva identificar, dentre as regras descobertas, aquelas com mais chance de serem úteis ao especialista, que identifica, com base em um conjunto descoberto, um subconjunto de regras que representam regras de exceção. A Figura 1 mostra um par de regras que representam a regra geral e a(s) respectiva(s) regra(s) de exceção. O símbolo "¬" denota a negação lógica, a qual pode representar a simples alteração do valor possível da variável que se apresenta no consequente da regra. É importante observar que uma regra de exceção é uma especialização de uma regra geral, e uma regra de exceção associa um consequente distinto em comparação com a regra geral. Esse método assume que regras gerais representam padrões conhecidos pelo usuário, tendo em vista que elas têm uma grande cobertura, ao contrário das regras de exceção, que, em geral, tendem a ser mais interessantes, uma vez que têm baixa cobertura. Entende-se como bque satisfazem a condição que compõe o antecedente da regra. Sendo assim, as regras de exceção tendem a ser surpreendentes, dado o fato de representarem uma contradição em relação à regra de senso comum (regra geral). É importante observar que a regra geral auxilia na explicação da causa da regra de exceção.

Tendo em vista a pouca apropriação do processo KDD pela fisioterapia, este artigo objetiva exemplificar e discutir a exploração de dados oriundos de acompanhamentos de pacientes utilizando Mineração de Dados e o pós-processamento dos padrões (regras) descobertos. A contribuição decorre do fato de buscar não apenas apresentar os conceitos, mas, também, demonstrá-los com exemplos de fácil compreensão, permitindo

$A \rightarrow C$ regra do senso comum (alta cobertura e precisão)
$A, B \rightarrow \neg C$ regra de exceção (baixa cobertura e alta precisão)

Figura 1 - Estrutura da regra de exceção diante da regra de senso comum

ao leitor replicar o comportamento dos algoritmos e melhor entender o processo, identificando possibilidades de aplicação em seu dia a dia profissional.

Materiais e métodos

Para demonstrar a aplicação de cada uma das três tarefas de Mineração de Dados, foi utilizado um conjunto de dados que registra o acompanhamento de 16 pacientes em uma clínica de fisioterapia. Para essa clínica, especificamente, existem dados disponíveis desde 2005, porém, para facilitar a demonstração e o entendimento dos resultados obtidos, foram selecionados, dentre os registros com o preenchimento mais completo, aleatoriamente, os registros de 16 pacientes. Como o artigo tem como objetivo demonstrar a potencialidade dessas novas tecnologias, caso o número de registros fosse muito grande, não seria possível ao leitor entender passo a passo o processo, nem mesmo replicá-lo.

Para o experimento em relação à tarefa de classificação, foi adotado o programa computacional C4.5 (15). Esse programa foi escolhido por atender a duas recomendações: predição e explicitação da relação entre as variáveis e a classe predita. O classificador descoberto é representado por uma estrutura hierárquica, na qual as ramificações internas representam as condições que compõem o antecedente das regras, por exemplo: se profissão = estudante e idade > = 15 anos, então a situação prevista é lesão.

Para demonstrar a tarefa de descoberta de regras de associação, foi utilizado o programa computacional Apriori (16). Essa escolha se deve ao fato de os dados coletados apresentarem várias ocorrências para uma mesma variável. Por exemplo: a variável diagnóstico fisioterapêutico pode apresentar um ou mais valores possíveis para um mesmo paciente, situação que não é permitida por outros programas computacionais que desempenham essa tarefa.

Para exemplificar a tarefa de agrupamento, foi utilizado o ambiente Weka (17), por se tratar de um ambiente bastante utilizado, de fácil operacionalização.

O uso de distintos programas computacionais permite ao leitor ampliar seu conhecimento sobre as diversas alternativas disponíveis.

Resultados

Como exemplo de base de dados sobre a qual é possível descobrir padrões na forma de classificadores, pode-se considerar a Tabela 1, que representa o cadastro de dados pessoais, bem como o respectivo diagnóstico de um grupo de pacientes de uma clínica de fisioterapia. As variáveis idade, sexo, peso, altura, estado civil, profissão, jornada, intervalo entre as horas trabalhadas, dieta alimentar ou atividade física são ditas variáveis "previsoras", e a variável diagnóstico clínico corresponde à variável classe. A presença do ponto de interrogação ("?") representa a ausência de valor para esse registro na variável correspondente, sendo que não é permitida a ausência de valor para a variável classe.

Desse conjunto de dados (Tabela 1), é possível extrair o classificador representado no Quadro 1.

A fim de contribuir para a compreensibilidade dos padrões descobertos, essa estrutura de árvore de decisão pode ser transformada em um conjunto de regras "se"... (condições), "então"... (classe), cuja interpretação é: "se" os valores das variáveis previsoras satisfazem as condições da regra, "então" a situação a ser classificada pertence à classe (valor da variável classe) prevista pela regra (consequente).

Dessa forma, a árvore de decisão representada no Quadro 1, que contém 12 ramificações distintas, é passível de ser transformada em 12 regras (R), podendo-se destacar algumas, tais como:

- R1: Se profissão = motorista ou representante comercial, então diagnóstico = lesão (cobertura de três exemplos).
- R2: Se profissão = auxiliar de limpeza, então diagnóstico = escoliose (cobertura de um exemplo).
- R3: Se profissão = estudante e idade > 15, então diagnóstico = lesão (cobertura de dois exemplos).
- R4: Se profissão = estudante e idade < 12, então diagnóstico = cifose (cobertura de dois exemplos).
- R5: Se profissão = estudante e idade < 15 e > 12, então diagnóstico = cifose (cobertura de dois exemplos).

O número entre parênteses (árvore de decisão) representa a cobertura da regra, ou seja, o número de exemplos da base de dados que satisfazem as condições que correspondem ao antecedente da regra. Para melhor identificar, o oitavo e o nono exemplos de dados (Tabela 1) são cobertos pela regra R1. Porém, o classificador descoberto pode, também, apresentar o que se denomina *erro de classificação*;

Tabela 1 - Cadastro de pacientes e seus respectivos diagnósticos

Idade	Sexo	Peso	Altura	Estado Civil	Profissão	Jornada	Intervalo	Horas	Dieta	Atividade física	Diagnóstico
12	F	40	1,58	Solteiro	Estudante	0	Sim	2	?	Sim	Cifose
15	F	50	1,57	Solteiro	Estudante	0	Sim	2	Não	Não	Escoliose
32	F	71	1,65	Casado	Auxiliar de limpeza	8	Sim	1	?	?	Escoliose
14	M	57	1,72	Solteiro	Estudante	0	Sim	2	?	?	Escoliose
31	M	61	1,71	Casado	Técnico eletrônico	8	Sim	1	Não	Sim	Cervicobraquialgia
36	M	90	1,83	Solteiro	Chaveiro	11	Não	0	Não	Sim	Fratura
20	M	78	1,69	Solteiro	Estudante	0	Sim	2	?	Sim	Lesão
50	M	88	1,70	Casado	Motorista	12	Sim	2	Não	Sim	Lesão
43	F	58	1,70	Divorciado	Representante comercial	44	Sim	2	?	?	Lesão
31	M	89	1,84	Casado	Representante comercial	8	Sim	1	Sim	Sim	Lesão
22	M	85	1,75	Solteiro	Estudante	0	Sim	2	Não	Sim	Lesão
45	M	78	1,76	Casado	Técnico em manutenção	?	Sim	2	Não	Não	Hérnia de disco
24	M	79	1,70	?	Atleta	?	Sim	2	?	Sim	Osteíte púbica
76	F	63	1,58	Viúvo	Do lar	44	Não	0	Não	Não	Tendinose
70	M	88	1,74	Casado	Professor	4	Sim	2	Não	Não	Ruptura de tendão do quadríceps
11	M	60	1,59	Solteiro	Estudante	0	Sim	2	Não	Sim	Fratura

Fonte: Dados da pesquisa.

Quadro 1 - Exemplo de classificador descoberto e representado por uma árvore de decisão

```
Árvore de Decisão:

    Profissao in {motorista, representante comercial}: lesao (3.0)
    Profissao = auxiliar de limpeza: escoliose (1.0)
    Profissao = tecnico eletronico: cervicobraquialagia (1.0)
    Profissao = chaveiro: fratura (1.0)
    Profissao = do lar: tendinose (1.0)
    Profissao = tecnico em manutencao: hernia de disco (1.0)
    Profissao = atleta: osteite pubia (1.0)
    Profissao = estudante
      | Idade > 15: lesao (2.0)
      | Idade <= 15:
            | | Idade <= 12: cifose (2.0/1.0)
            | | Idade > 12: escoliose (2.0)
```

Fonte: Dados da pesquisa.

por exemplo: nem todos os registros cobertos pela regra que compõem a base de dados apresentam como valor da variável classe aquele valor predito pelo consequente da regra. Um exemplo de erro é evidenciado a partir da regra R4, quando dois registros, representados pela primeira e pela última linhas da Tabela 1, satisfazem as duas condições do antecedente, mas apenas um desses (o primeiro) tem como valor para a variável classe – diagnóstico – o valor predito cifose, o que caracteriza um erro de predição da regra em questão.

Como exemplo de base de dados para os experimentos de descoberta de regras de associação, para a tarefa de Mineração de Dados, pode-se considerar os dados de diagnóstico clínico e fisioterapêutico para cada um dos pacientes em questão (Quadro 2).

Quadro 2 - Diagnósticos clínicos e fisioterapêuticos dos pacientes

Diagnóstico clínico	Diagnóstico fisioterapêutico		Resultado
Cifose torácica	Lombalgia	Encurtamento muscular	Diminuição da ADM
Escoliose toracolombar	Lombalgia	Encurtamento dos paravertebrais	Diminuição de força
Escoliose toracolombar	Algia toracolombar	Diminuição da ADM	Impotência funcional
Escoliose toracolombar	Alteração postural	Fraqueza muscular	Encurtamento muscular
Cervicobraquialgia	Algia	Perda da ADM	Fraqueza muscular
Fratura do úmero	Capsulite adesiva	Diminuição da ADM	
Lesão de LCM	Algia de joelho	Diminuição da ADM	Instabilidade articular
Lesão talofibular anterior	Diminuição da ADM	Edema de tornozelo	Algia
Lesão parcial de LCM	Diminuição da ADM	Fraqueza muscular	Encurtamento muscular
Lesão	Diminuição da ADM	Algia	Encurtamento muscular
Lesão do menisco medial	Algia de joelho	Edema	Estiramento do LCA
Hérnia de disco lombar	Lombalgia		
Osteíte púbica	Pubalgia	Diminuição da ADM	Encurtamento muscular
Tendinose do supraespinhoso	Diminuição da ADM	Perda de funcionalidade	
Ruptura de tendão do quadríceps	Encurtamento muscular	Diminuição da ADM	Atrofia muscular
Fratura de tornozelo	Diminuição da ADM	Diminuição de força	Edema

Fonte: Dados da pesquisa.

Conforme pode ser observado no Quadro 2, cada paciente apresenta, obrigatoriamente, um diagnóstico clínico e pelo menos um diagnóstico fisioterapêutico. Ao contrário do que ocorreu na preparação dos dados para a descoberta de padrões na forma de classificadores, os dados não apresentam formatação estruturada, podendo ter um ou mais valores para uma mesma variável, e não se faz necessária a indicação de valor ausente com a utilização do ponto de interrogação ("?").

Com base nos dados apresentados no Quadro 2, é possível descobrir uma série de regras de associação, como:

- R6 – SE encurtamento muscular, ENTÃO diminuição da ADM (37,5%, 83,3%).
  - Onde se lê que: 37,5% dos pacientes apresentam encurtamento muscular, dos quais 83,3% também apresentam diminuição da ADM.
- R7 – SE diminuição da ADM, ENTÃO encurtamento muscular (68,8%, 45,5%).
  - Onde se lê que: 68,8% dos pacientes têm diminuição da ADM, dos quais 45,5% também apresentam encurtamento muscular.

A aplicação do programa computacional Apriori (16) nos dados do Quadro 2 resultou em 278 regras descobertas. Como forma de facilitar a etapa de avaliação dessas quase três centenas de regras, inicialmente foram eliminadas as regras ditas redundantes, restando, assim, apenas 109, o que representa 39% das regras originalmente descobertas.

Outra forma de facilitar a avaliação das regras descobertas, potencializando a chance de identificação daquelas que venham a ser úteis, bem como das que agreguem conhecimento ao que o especialista já conhece, é a busca por situações que expressem exceções, tais como:

- Regra Geral 1:
  - SE encurtamento muscular, ENTÃO diminuição da ADM (37,5%, 83,3%). Onde se lê que: 37,5% dos pacientes têm encurtamento muscular, dos quais 83,3% também apresentam diminuição da ADM.
- Regra de Exceção 1.1:
  - SE encurtamento muscular E ruptura de tendão do quadríceps, ENTÃO atrofia muscular (6,3%, 100,0%). Isso indica que, se o

encurtamento muscular estiver associado a ruptura de tendão do quadríceps, então os pacientes também apresentam atrofia muscular.

- Regra de Exceção 1.2:
  - SE encurtamento muscular E lesão parcial de LCM, ENTÃO fraqueza muscular (6,3%, 100,0%). Isso indica que, se o encurtamento muscular estiver associado a lesão parcial de LCM, então os pacientes também apresentam fraqueza muscular.
- Regra de Exceção 1.3:
  - SE encurtamento muscular E cifose torácica, ENTÃO lombalgia (6,3%, 100,0%). Ou seja, se o encurtamento muscular estiver associado a cifose torácica, o diagnóstico fisioterapêutico se altera para lombalgia.

Ampliando a complexidade dos dados a serem trabalhados, poder-se-ia complementar os dados do Quadro 2 com outros, relacionados à anamnese e a testes específicos, possibilitando a descobertas das seguintes regras:

- Regra Geral 2:
  - SE teste gaveta anterior E teste gaveta posterior, ENTÃO diminuição da ADM (25,0%, 100,0%).
- Regra de Exceção 2.1:
  - SE teste gaveta anterior E teste gaveta posterior E lesão de LCM, ENTÃO instabilidade articular (6,3%, 100,0%).
- Regra de Exceção 2.2:
  - SE teste gaveta anterior E teste gaveta posterior E teste patrickfabere, ENTÃO lombalgia (6,3%, 100,0%).

Onde se lê que 25% dos pacientes realizam testes de gaveta anterior e posterior, sendo que todos apresentam diagnóstico fisioterapêutico de diminuição de ADM. Porém, se esses mesmos testes estiverem combinados ao diagnóstico clínico de lesão de LCM, o diagnóstico fisioterapêutico se altera para instabilidade articular, comprovando as consequências da sobreposição de quadros clínicos que acarretam consequências limitantes. Da mesma forma, se esses mesmos testes estiverem combinados ao teste patrickfabere, o diagnóstico fisioterapêutico se altera para lombalgia.

Finalmente, para demonstrar a terceira e última tarefa da Mineração de Dados, pode-se aplicar o programa computacional que agrupa os registros de dados conforme características de maior similaridade entre os valores disponíveis para cada uma das variáveis da Tabela 2, que resultou no agrupamento dos 16 registros em apenas dois grupos, rotulados como grupo 0 e grupo 1.

A partir da execução do programa que identifica os grupos e rotula, para cada registro, a qual grupo ele pertence, pode-se perceber que os 16 registros foram distribuídos em dois grupos, um com sete e o outro com nove registros.

Conforme já foi mencionado anteriormente, com a identificação do grupo ao qual pertence cada registro de dados, é possível construir um classificador (Quadro 3), considerando como variável classe a própria identificação do respectivo grupo e buscando, assim, explicitar quais variáveis presentes no conjunto de dados tiveram participação efetiva na identificação desses grupos.

A partir do classificador descoberto na forma de árvore de decisão, é possível perceber que, a despeito de o conjunto de dados disponibilizar 12 variáveis previsoras além da própria variável classe, a única que de fato determinou a criação dos grupos foi a variável profissão. Vale destacar que, quando da construção do primeiro classificador demonstrado no Quadro 1, a mesma variável foi determinante na construção do classificador em questão.

Discussão

Para melhor entender o processo de descoberta de classificadores na forma de árvore e sua posterior transformação em regras, note que os registros de dados (Tabela 1) de sequência 8, 9 e 10 apresentam, como profissão, motorista ou representante comercial, e que o respectivo valor para a variável classe é lesão. Ou seja, a regra R1 expressa exatamente essa relação, ou seja, se a profissão for motorista ou representante comercial, então o diagnóstico predito é lesão. Pode-se inferir que essas ocupações apresentam características que devem ser mais exploradas para efeito de adequações de rotinas de trabalho ou fomento a comportamentos de autocuidado, além de outras intervenções, nos níveis primários, secundários ou terciários, na área da saúde.

Essa relação num primeiro momento pode parecer trivial e facilmente identificada em um conjunto de dados de pequena magnitude, porém em geral as clínicas dispõem de milhares de dados que refletem o acompanhamento de pacientes por longos períodos

Tabela 2 - Cadastro de pacientes, respectivos diagnósticos e grupo

Idade	Sexo	Peso	Altura	Estado civil	Profissão	Jornada	Intervalo	Horas	Dieta	Atividade física	Diagnóstico	Grupo
12	F	40	1,58	Solteiro	Estudante	8	Sim	2	?	Sim	Cifose	1
15	F	50	1,57	Solteiro	Estudante	8	Sim	2	Não	Não	Escoliose	1
32	F	71	1,65	Casado	Auxiliar de limpeza	9	Sim	1	?	?	Escoliose	0
14	M	57	1,72	Solteiro	Estudante	8	Sim	2	?	?	Escoliose	1
31	M	61	1,71	Casado	Técnico eletrônico	8	Sim	1	Não	Sim	Cervicobraquialgia	0
36	M	90	1,83	Solteiro	Chaveiro	11	Não	0	Não	Sim	Fratura	1
20	M	78	1,69	Solteiro	Estudante	8	Sim	2	?	Sim	Lesão	1
50	M	88	1,70	Casado	Motorista	12	Sim	2	Não	Sim	Lesão	0
43	F	58	1,70	Divorciado	Representante comercial	44	Sim	2	?	?	Lesão	1
31	M	89	1,84	Casado	Representante comercial	8	Sim	1	Sim	Sim	Lesão	1
22	M	85	1,75	Solteiro	Estudante	0	Sim	2	Não	Sim	Lesão	1
45	M	78	1,76	Casado	Técnico em manutenção	?	Sim	2	Não	Não	Hérnia de disco	0
24	M	79	1,70	?	Atleta	?	Sim	2	?	Sim	Osteíte púbica	0
76	F	63	1,58	Viúvo	Do lar	44	Não	0	Não	Não	Tendinose	0
70	M	88	1,74	Casado	Professor	4	Sim	2	Não	Não	Ruptura de tendão do quadríceps	0
11	M	60	1,59	Solteiro	Estudante	8	Sim	2	Não	Sim	Fratura	1

Fonte: Dados da pesquisa.

Quadro 3 - Classificador gerado a partir do conjunto de dados e de seus respectivos grupos

Árvore de Decisão:

Profissao in {motorista, auxiliar de limpeza, tecnico eletronico, tecnico em manutencao, atlete professor, do lar}: 0(7.0)
Profissao in { estudante, chaveiro, representante comercial}: 1(9.0)

Fonte: Dados da pesquisa.

de tempo. Desnecessário dizer que para estes conjuntos de maior magnitude a identificação não será possível sem contar com o apoio de programas computacionais tais como o C4.5.

Vale também destacar que essas regras descobertas podem alimentar um sistema de apoio à identificação de diagnósticos durante a atividade do fisioterapeuta. Durante a consulta, esse profissional colhe uma série de dados do paciente, os quais, fornecidos ao sistema, poderão retornar como resposta na forma dos possíveis diagnósticos preditos, inclusive indicando a margem de segurança de acerto de cada um deles. Uma ferramenta como essa, além de facilitar a atividade do fisioterapeuta, potencializa a utilização das bases de dados que estão disponíveis.

Sobre a potencialidade de uso dos padrões descobertos na forma de regras de associação, bem como a identificação de situações de exceção, é importante ressaltar que a relação (Regra Geral 1) não é necessariamente de causalidade linear, mas, sim, decorrente de uma combinação de condições disfuncionais que podem apresentar congruência ou incongruência. As incongruências, quando ocorrem, podem ser reveladoras de incompatibilidade de linguagens para o registro dos dados, preenchimento incorreto ou, ainda, necessidade de que se refaça o processo diagnóstico ou de que outros exames sejam pedidos.

Já observando a Regra de Exceção 1.2, se o encurtamento muscular estiver combinado à lesão parcial

de LCM, o diagnóstico fisioterapêutico se altera para fraqueza muscular. Nesse caso, percebe-se a sobreposição de condições de origem crônica e aguda que, somadas, podem resultar em consequência negativa para a função: a fraqueza muscular por desuso. Poderia, além disso, ser inferido que a condição crônica de encurtamento muscular, ao comprometer a postura estática e dinâmica, predispõe em médio e longo prazo a uma diminuição da capacidade funcional.

A oportunidade de adoção de descobertas de regras de associação sobre os dados de pacientes submetidos à fisioterapia pode oportunizar melhor entendimento das especificidades que podem ocorrer com o grupo de pacientes atendido pela clínica em questão, ampliando, assim, o conhecimento do profissional na identificação das condutas a serem adotadas.

Sobre a tarefa de agrupamento, foi possível perceber que, a despeito das diversas variáveis disponíveis para caracterizar os pacientes envolvidos na amostra, aquela que melhor caracterizou não apenas a variável classe, mas, também, o agrupamento por similaridades foi a variável profissão. Esse tipo de informação poderia ser importante na identificação do critério de agrupamento dos clientes, para o desenvolvimento de alguma atividade para a qual fosse interessante reunir os pacientes mais "parecidos entre si", como é o caso das práticas de educação em saúde ou das terapias em grupo.

Outra contribuição pode ser a importância da variável profissão para a avaliação, apesar de nem sempre ser devidamente explorada em anamnese, o que poderia elucidar a gênese do problema e os fatores desencadeantes da dor, além do *deficit* de funcionalidade e da recorrência dos distúrbios.

Nesse exemplo específico em Mineração de Dados, a visibilidade dada ao atributo profissão gerou, também, possibilidades aos fisioterapeutas de, com a devida apropriação desse conjunto de dados, identificar novos padrões de associação de variáveis, as quais possam dar significado às ações diagnósticas e terapêuticas na área de Fisioterapia do Trabalho.

Conclusão e trabalhos futuros

O processo KDD envolve uma série de etapas, desde a preparação dos dados, a descoberta de padrões até a avaliação do quanto esses padrões agregam valor ao que o especialista já conhece sobre o problema em questão. Este artigo, além de apresentar essas etapas, detalhou o comportamento de alguns programas computacionais em relação a conjuntos de dados de pequena magnitude, permitindo, assim, que o leitor entenda um pouco melhor o seu funcionamento. Para a etapa de Mineração de Dados, foram simulados programas para a descoberta de padrões, na forma de regras de associação, classificadores e agrupamento. Para a etapa de pós-processamento, foram simulados processos de redução de redundâncias, bem como identificados os pares de regras (Regra Geral e suas respectivas Regras de Exceção) sobre o conjunto total descoberto.

Este estudo demonstra que a aplicação de tecnologias alternativas para melhor aproveitar o potencial dos dados disponíveis demanda grande esforço, na busca de um conjunto de dados que configure uma fonte confiável de pesquisa. As maiores dificuldades podem residir na construção de bancos de dados baseados em prontuários incompletos ou ilegíveis, nas informações contraditórias ou na ausência de preenchimento de campos importantes.

A falta de cuidado no preenchimento dos prontuários pode ser creditada ao contexto de trabalho do fisioterapeuta, o qual o obriga, muitas vezes, a buscar o resultado terapêutico com base em seu conhecimento empírico. Isso pode resultar em remissão dos sintomas, mas não contribui com o registro e o armazenamento adequado das informações do exame clínico.

Outro fator explicativo é o fato de que, mesmo com o avanço da informática em saúde, o registro das informações na área da fisioterapia encontra-se, em sua maioria, em prontuários de papel (18). As variadas condições que podem estar associadas à avaliação e ao diagnóstico fisioterapêutico tornam a utilização dos registros em papel ineficazes no auxílio ao profissional, em virtude da falta de praticidade na coleta dos dados e na visualização de seus resultados, dificultando, ainda mais, o tratamento dos dados e escondendo todo seu potencial revelador.

Os devidos registro e armazenamento podem contribuir não apenas com a composição de valioso banco de dados, mas, principalmente, com o uso desses dados, visando a minerar o que está subjacente aos dados brutos e trazer à tona informações que levarão a descobertas importantes para o redirecionamento de práticas e a tomada de decisões.

Neste estudo, não foram considerados todos os prontuários disponíveis, por se tratar de uma demonstração do potencial das técnicas. Assim, sugere-se que

novos experimentos sejam realizados, considerando todo o conjunto de dados, bem como, que se avalie a oportunidade de incorporar os padrões descobertos num sistema que apoie o fisioterapeuta no momento do estabelecimento dos diagnósticos.

Vale destacar que, em outras áreas da saúde, a utilização do processo KDD permitiu agregar significativo poder de decisão aos gestores, como relatado no estudo de classificação de beneficiários para o programa de gerenciamento de casos (6). Em um primeiro momento, o método foi adotado para o gerenciamento de casos de diabetes, mas, atualmente, esse modelo já está sendo aplicado para diversas outras doenças crônicas.

Referências

1. Hand DJ. Introduction. In: Berthold M, Hand DJ, editor. Intelligent data analysis. Berkeley: Springer-Verlag; 1999. p. 1-14. doi:10.1007/3-540-48412-4.

2. Kobus LSG. Aplicação da descoberta de conhecimento em base de dados para identificação de usuários com doenças cardiovasculares elegíveis para programas de gerenciamento de caso [dissertação]. Curitiba: Pontifícia Universidade Católica do Paraná; 2006.

3. Lopes L. Aprendizagem de máquina baseada na combinação de classificadores em bases de dados da área de saúde [dissertação]. Curitiba: Pontifícia Universidade Católica do Paraná; 2007.

4. Vianna RCXF, Moro CMCB, Moises SJ, Carvalho D, Nievola JC. Mineração de dados e características da mortalidade infantil. Cad Saúde Pública. 2010;26(3):535-42. doi:10.1590/S0102-311X2010000300011.

5. Von Stein Jr. A, Malucelli A, Bastos LC, Carvalho DR, Cubas MR, Paraiso EC. Classificação de microareas homogêneas de risco com uso de mineração de dados. Rev Saúde Pública. 2010;44(2):292-300. PMid:20339628.

6. Dallagassa MR. Concepção de uma metodologia para identificação de beneficiários com indicativos de diabetes mellitus tipo 2 [dissertação]. Curitiba: Pontifícia Universidade Católica do Paraná; 2009.

7. Kuretzki CH. Técnicas de mineração de dados aplicadas em bases de dados para saúde a partir de protocolos eletrônicos [dissertação]. Curitiba: Universidade Federal do Paraná; 2009.

8. Fancying N, Jieying B, Xuebiao Z. Study on China's food security status. Agriculture and Agricultural Science Procedia. 2010;1:301-10. doi:10.1016/j.aaspro.2010.09.038.

9. Machado DZ. Estudo do desenvolvimento da imagem corporal interna usando processo de descoberta do conhecimento [dissertação]. Curitiba: Pontifícia Universidade Católica do Paraná; 2011.

10. Fayyad U, Piatetsky-Shapiro G, Smyth P, Uthurusamy R. Advances in knowledge discovery and data mining. American Association for Artificial Intelligence. Menlo Park: MIT Press; 1996.

11. Freitas AA, Lavington SH. Mining very large databases with parallel processing, Norwell: Kluwer Academic Publishers; 1998. doi:10.1007/978-1-4615-5521-6.

12. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Monterey: Wadsworth and Brooks; 1984.

13. Kubat M, Bratko I, Michalski RS. A review of machine learning methods. In: Michalski RS, Bratko I, Kubat M, editor. Machine learning and data mining: methods and applications. London: John; 1998. p. 3-66.

14. Hussain F, Liu H, Lu H. Exception rule mining with a relative interestingness measure. Lecture notes in Artificial Intelligence2000;1805:86-97.

15. Quinlan JR. C4.5 Programs for machine learning. San Diego: Morgan Kaufmann; 1993.

16. Borgelt C. Apriori – Association Rule Induction. 2004 [citado 17 maio 2011]. Disponível em: http://www.borgelt.net/apriori.html.

17. Weka. Waikato Environment for Knowledge Analysis [citado 17 maio 2011]. Disponível em: www.cs.waikato.ac.nz/ml/weka.

18. 18. Camargo RS, Moser ADL, Bastos LC. Abordagem dos métodos avaliativos em fibromialgia e dor crônica aplicada à tecnologia da informação: revisão da literatura em periódicos, entre 1998 e 2008. Rev Bras Reumatol. 2009;49(4):431-46. doi:10.1590/S0482-50042009000400009.

Recebido: 03/07/2011
Received: 07/03/2011

Aprovado: 18/01/2012
Approved: 01/18/2012