Técnicas de Classificação em Problemas Relacionados a Doenças Cardíacas

Douglas Guisi
UTFPR PB
Av. do Conhecimento
Pato Branco – Pr
guisiagudos@gmail.com

Jonatas Loureiro de Almeida Jr.
UTFPR PB
Av. do Conhecimento
Pato Branco – Pr
jonatas.adiemla@gmail.com

André Pinz Borges
UTFPR PG
Av Monteiro Lobato, s/n, Km 04
Ponta Grossa – Pr
apborges@utfpr.edu.br

Marcelo Teixeira
UTFPR PB
Av. do Conhecimento
Pato Branco – Pr
marceloteixeira@utfpr.edu.br

Richardson Ribeiro
UTFPR PB
Av. do Conhecimento
Pato Branco – Pr
richardsonr@utfpr.edu.br

Gabriel Gomes de Sousa,
Hudson dos Santos Lapa
UTFPR PB
sousa@alunos.utfpr.edu.br
hudsonslpa@gmail.com

RESUMO

Este artigo apresenta a aplicação de técnicas de classificação em problemas relacionados a doenças cardíacas. As doenças cardiovasculares são os distúrbios ligados ao coração e aos vasos sanguíneos, com uma vasta gama de síndromes clínicas, sendo a aterosclerose a mais frequente em diagnóstico. Neste artigo, realizou-se um estudo com base nos repositórios de dados destas doenças, a fim de encontrar os parâmetros presentes em exames laboratoriais e prontuários médicos que melhor se conectam para a descoberta de pacientes com possíveis riscos de doenças cardiovasculares. Os resultados atingidos por esta pesquisa mostram que a existência de diabetes, hábitos e concentrações de substâncias no corpo, aumentam a possibilidade de contração de novas doenças relacionadas.

Palavras Chaves

Mineração de Dados; Classificação; Doenças Cardíacas.

ABSTRACT

This article presents the application of classification techniques in problems related to heart-disease. Cardiovascular diseases are disorders related to heart and blood vessels, with a wide range of clinical syndromes, the most common diagnosis of atherosclerosis. In this article, we carried out a study based on data repositories of these diseases to find the parameters present in laboratory tests and medical records to better connect to the discovery of patients with possible risk of cardiovascular disease. The results obtained in this study show that the existence of diabetes, habits and concentrations of substances in the body, increase the possibility of further contraction of diseases.

Keywords

Data Mining; Classification, Heart Diseases.

1. INTRODUÇÃO

As doenças cardiovasculares possuem consideráveis síndromes clínicas, porém as relacionadas à aterosclerose (acúmulo de gordura nos vasos sanguíneos) são as mais frequentes em diagnósticos [6] [29]. Segundo o Ministério da Saúde do Brasil, as doenças cardiovasculares são as responsáveis por aproximadamente 30% dos óbitos, colocando o Brasil entre os 10 países mais afetados do mundo [24].

Os fatores que contribuem para o surgimento destas síndromes podem ser genéticos, mas os principais dependem dos hábitos do paciente, como sedentarismo e tabagismo [6]. A diabete, doença não relacionada ao sistema circulatório também é objeto de estudo como um fator das doenças cardiovasculares. Pacientes com diabetes *mellitus* tipo 2, por exemplo, tem um risco de duas a quatro vezes maior de vir a desenvolver doenças coronárias [19]. Esta doença evolui sem apresentar sintomas notórios ou conhecimento do paciente, fatores que dificultam o diagnóstico. Estima-se que metade das pessoas portadoras não tem conhecimento de estarem com a diabetes *millitus* tipo 2 [22].

Atividades clínicas, como consultas, exames laboratoriais, prescrições médicas, diagnósticos, vacinações, entre outras, são realizadas diariamente por diferentes profissionais da área da saúde. Os dados dessas atividades quando agrupados, emergem um conjunto de dados geralmente na ordem de milhares. Uma possibilidade para obter o conhecimento implícito inerente a um relacionamento específico desses dados é aplicar as etapas do processo de descoberta de conhecimento (*Knowledge Discovery in Databases* - KDD), explorando principalmente os métodos de Mineração de Dados [9][10][28]. Segundo [13], as principais aplicações da Mineração de Dados na saúde estão em efetividade de tratamentos médicos, gerenciamento de sistemas, e detecção de uso indevido e/ou fraudes de recursos destinados à saúde.

Na efetividade de tratamentos médicos, as aplicações de Mineração de Dados podem gerar informações que auxiliem o profissional da área da saúde nas tomadas de decisões como prescrições, exames e encaminhamentos. Ou ainda por meio da análise dos dados, pode-se chegar a conclusões sobre causas de uma doença, sintomas e tratamentos.

Neste artigo integramos o WEKA[31], um software com uma coleção de algoritmos de aprendizagem de máquina para executar tarefas de mineração de dados, à bases de dados *online* de pacientes, obtidos a partir de repositórios de dados (*benchmarks*) relacionados às doenças cardíacas, para a detecção de diabetes e doenças cardiovasculares.

Foram testados algoritmos de classificação usando diferentes métodos, como árvores de decisão, redes neurais artificiais, regressão e redes bayesianas, na intenção de encontrar os melhores modelos para os conjuntos de dados utilizados.

2. SISTEMAS DE APRENDIZAGEM COM MINERAÇÃO DE DADOS

Sistemas de aprendizagem baseados em mineração de dados são geralmente formados por técnicas e procedimentos, que se baseiam na aplicação de algoritmos sobre grupos de dados para a extração de algum conhecimento implícito, que esteja inerente a um relacionamento específico desses dados [28].

Os algoritmos utilizados em sistemas de aprendizagem são normalmente baseados em diferentes áreas do conhecimento (*e.g.*, matemática e estatística), e trabalham por meio de agrupamentos, classificação ou associação, possibilitando a criação de regras sobre os dados. O conhecimento é gerado a partir da interpretação dessas regras [28].

Neste artigo é utilizado o paradigma de classificação. Classificação é o processo de encontrar um conjunto de modelos (funções) que descrevem e distinguem classes ou conceitos, com o propósito de utilizar o modelo para predizer a classe de objetos que ainda não foram classificadas [21].

Um método bastante conhecido são as árvores de decisão [25], aplicado na inferência indutiva. Uma árvore de decisão usa a estratégia dividir para conquistar para resolver um problema de decisão. Nas árvores de decisão, um problema complexo é dividido em problemas mais simples. Nesses subproblemas é aplicada recursivamente a mesma estratégia. As soluções dos subproblemas podem ser combinadas na forma de uma árvore, para produzir a solução de um problema complexo [21]. Essa é a ideia básica por trás de algoritmos baseados em árvores de decisão. Em domínios da área médica, árvores de decisão foram usadas para a análise de custo-efetividade de vacinas [8], problemas de fratura ortopédica [30], tomadas de decisões clínicas [5] [2] [3], tratamento de feridas crónicas [17], etc.

Outras técnicas para classificação foram inspiradas em paradigmas probabilísticos e também foram testadas com dados da área médica. Por exemplo, o *Naive Bayes* é um classificador probabilístico baseado na hipótese no qual as variáveis fornecidas são independentes [18]. Suas primeiras aplicações na medicina surgiram na década de 90, sendo elas na predição do prognóstico de pacientes e seleção de tratamentos. Um exemplo de sua utilização está na classificação de radiografias de tórax para o diagnóstico de pneumonia infantil [27].

Outra técnica são as Redes Neurais Artificiais, inspiradas na estrutura dos neurônios humanos [4]. A principal função delas é a predição de eventos. Uma das dificuldades com esse método, dependendo do problema, é sua complexidade e compreensão das predições realizadas. Vários estudos e aplicações das redes Neurais Artificias em problemas na área médica são encontrados, por exemplo, em [1] [7] [11] [12] [15].

Há ainda diversas outras aplicações que se beneficiaram das técnicas de classificação, como o desenvolvimento de antidepressivos [16], doença da artéria coronária [20], análise de fatores genéticos para a predisposição ao câncer cervical [26], e processamento de sinais EEG [14]. Esses são apenas alguns dentre os vários trabalhados dedicados ao uso de sistemas de aprendizagem na área da saúde.

3. METODOLOGIA

Para gerar a descoberta do conhecimento foi utilizado o Weka. Esse software oferece recursos para a execução de tarefas relacionadas ao pré-processamento de dados como, por exemplo, a seleção e a transformação de atributos (e.g., remoção de dados errôneos, criação de escalas para dos dados). O Algoritmo 1 demonstra um exemplo de como utilizar uma técnica de mineração de dados com a linguagem Java.

```
Instances data = new Instances(new BufferedReader(new
FileReader(
"C:\\weather.nominal.arff")));
data.setClassIndex(data.numAttributes()-1); //Definir o
índice dos dados
String[] options = new String[1];
options[0] = "-U"; //Árvore sem podas
J48 tree = new J48();//Nova instância da árvore
tree.setOptions(options); //Configurar o classificador
tree.buildClassifier(data); //Construir o classificador
```

Algoritmo 1. Exemplo em Java para invocar uma técnica de mineração de dados.

Uma prática utilizada na mineração de dados da saúde é o uso de repositórios de dados de acesso geral e/ou restrito. Existem vários repositórios *on-line* gratuitos e irrestritos, que contém milhares de bases de dados reais, as quais podem ser utilizadas pelos pesquisadores em seus algoritmos e técnicas de mineração de dados. Os dados clínicos estão ligados intrinsecamente a regras éticas e de sigilo, de modo que não é possível encontrá-los com facilidade e muito menos utilizá-los para análise em pesquisas com Mineração de Dados. Logo, o uso de dados de repositórios permite aos pesquisadores experimentar algoritmos na busca por informações.

Este trabalho utilizou as seguintes bases: Heart-statlog[1], Diabetes[2] e Arrhytmia[3], com um objetivo mais específico: elencar os atributos de cada base de dados relacionada a uma doença cardiovascular, que podem ocasionar a ocorrência de uma segunda doença do gênero. A motivação deste trabalho parte de uma pesquisa, desenvolvida por [23], que revelou que doenças cardiovasculares como infarto, hipertensão e angina, possuem relação com a diabetes. Segundo [29] os principais riscos para a ocorrência de doença cardiovascular são: elevado LDL (conhecido como o colesterol ruim), baixo HDL (conhecido como o colesterol bom), alta pressão sanguínea, elevada glicose no sangue, sedentarismo, obesidade, consumo de tabaco. Dessa forma, é possível identificar relações entre a existência de diabetes, hábitos e concentrações de substâncias no corpo e a possibilidade de contração de novas doenças relacionadas.

Após a verificação das correlações existentes entre diferentes atributos e o risco de contrair determinada doença, foram realizados testes computacionais utilizando técnicas de mineração de dados a fim de mostrar resultados com as correlações existentes. A abordagem utilizada para a mineração dos dados foi a classificação. Essa funcionalidade tem por objetivo segmentar um conjunto de dados em subconjuntos específicos. Os dados pertencentes a um subconjunto gerado devem possuir características em comum. Dessa forma, ao final da execução de uma tarefa de classificação, serão criados *n* subconjuntos de dados, cada qual agrupando dados com características específicas. Uma vez aplicada, essa funcionalidade pode atuar sobre novos conjuntos, prevendo em qual subconjunto eles se enquadram.

Para a realização dos testes de classificação foram utilizadas as bases de dados de detecção de diabetes e de doenças cardiovasculares disponíveis nos repositórios públicos indicados. Da base diabetes foi usado 9 atributos, dentre eles: número de vezes que a mulher engravidou, pressão sanguínea (em *mm/Hg*), grossura da camada de pele do tríceps (em *mm*), IMC e idade. Nas bases relacionadas a doença cardiovascular foi utilizado 13 atributos, dentre eles: idade, sexo, escala numérica de dor no peito

[1] http://tunedit.org/repo/UCI/heart-statlog.arff

[2] http://storm.cis.fordham.edu/~gweiss/data-mining/weka-data/diabetes.arff

[3] http://tunedit.org/repo/UCI/arrhythmia.arff

(0-10), nível de açúcar no sangue (em *mg/dl*), taxa máxima de batimentos cardíacos alcançada no exame, entre outros.

O objetivo da aplicação de Mineração de Dados foi detectar a presença ou ausência de diabetes, dada a informação prévia de doença cardíaca. Para a criação da nova base de dados realizou-se uma junção das duas anteriores, eliminando apenas atributos repetidos em ambas: sexo e pressão sanguínea. Portanto, a nova base de dados foi composta por 20 atributos, 1490 instâncias (registros) e duas classes que indicam (positivo) ou não (negativo) a presença de diabetes. Após isso, foram aplicadas técnicas de classificação.

4. RESULTADOS EXPERIMENTAIS

Os resultados experimentais foram obtidos com a ferramenta WEKA e são apresentados na Tabela 1. Para dividir o conjunto de dados em treinamento e teste, foi empregado o método de validação cruzada *k-fold* [28] como forma de avaliar a capacidade de generalização das técnicas de aprendizagem a partir dos conjuntos de dados de treinamento e teste. No procedimento de validação cruzada *k-fold*, o conjunto de dados é dividido aleatoriamente em *k* subconjuntos. Destes, *k*-1 são utilizados para o treinamento e um é utilizado para teste. Esse processo é repetido *k* vezes até que todos os subconjuntos de dados sejam utilizados no conjunto de teste. Dessa forma, diferentes classificadores são obtidos, e a precisão das classificações dos conjuntos de treinamento e de teste pode ser avaliada. Todos os testes foram realizados utilizando validação cruzada com 10-*fold*.

Os resultados da Tabela 1 mostram que as técnicas de classificação por Regressão e *Naive Bayes* obtiveram os melhores resultados, detectando a presença de diabetes tipo 2 com um percentual de acerto médio de 65,7%, sendo que a técnica de regressão foi a que melhor identificou a diabetes nos pacientes. Estas técnicas buscam encontrar funções, geralmente de primeira ordem, capazes de mapear as instâncias do conjunto de dados em valores reais. Exemplos da utilização dessas técnicas se encontram na estimativa da probabilidade de um paciente vir a possuir a doença, dado o resultado de um conjunto de diagnósticos de exames [20] e [14].

Table 1. Percentual de acertos dos algoritmos.

Algoritmos	Acertos (%)
Classificação por Regressão	72,6
Naive Bayes	71,1
Rede Neural Perceptron Multicamadas	65,0
J48 (árvore de decisão)	61,9
Logística Bayesiana com Regressão	61,2

A base de dados utilizada foi numérica, isto é, todas as instâncias possuem em seus atributos domínios com valores inteiros ou reais, tornando propício a aplicação de técnicas de Regressão e Probabilísticas na classificação dos dados. Isto também explica a melhora na taxa de classificação obtida por estas técnicas em relação às demais, uma vez que as demais aceitam outros tipos de dados e, portanto, não são especificamente criados para trabalhar com números.

5. CONCLUSÕES E DISCUSSÕES

Apesar do processo de descoberta do conhecimento apresentar aplicabilidade à área da saúde, nota-se há falta de sistemas comerciais integrados. Existem diversas empresas que promovem serviços de prontuário eletrônico de pacientes, sistemas de suporte a decisão e controle de gestão, porém ainda não é possível fazer uma mineração de dados utilizando os dados clínicos de um país inteiro, por exemplo. Isso se deve ao isolamento entre esses sistemas, que ocorre por estratégias comerciais ou mesmo para respeitar as regras de sigilo dos dados. No entanto a ideia de utilizar escopos de dados abrangentes como deste trabalho pode representar a possibilidade de algumas descobertas importantes, como por exemplo, quais fatores climáticos influenciam na transmissão de determinadas doenças, entre outras informações de nível regional e nacional. Essas hipóteses serão objetos de estudos.

Neste trabalho foi realizado um estudo em repositórios de dados de doenças cardiovasculares, na intenção de encontrar quais os parâmetros presentes em exames laboratoriais ou prontuários médicos que melhor se relacionam na descoberta de pacientes com possíveis riscos de doenças cardiovasculares.

Nos repositórios usados para este trabalho, foram selecionados atributos na intenção de identificar a ocorrência de uma segunda doença do gênero. Outras frentes estão sendo investigadas, como por exemplo, usar outras bases que possam fazer relação a outras doenças cardiovasculares; estender a metodologia proposta a outros problemas da saúde como doenças cancerígenas e respiratórias. Há ainda questões de interação dos resultados do projeto com instituições, como por exemplo, firmar parcerias com instituições públicas e/ou privadas para o benefício de usar dados oriundos de sistemas de prontuários eletrônicos do Brasil. Algumas dessas hipóteses já estão sobre investigação para trabalhos futuros.

6. AGRADECIMENTOS

Está pesquisa é apoiada por Decit/SCTIE/MS, por intermédio do CNPq, com apoio da Fundação Araucária e da SESA-PR. Programa Pesquisa para o Sistema Único de Saúde: Gestão Compartilhada em Saúde - PPSUS.

7. REFERÊNCIAS

[1] Akay, M. and Welkowitz, W. (1993) “Acoustical detection of coronary occlusions using neural networks”. Journal of Biomedical Engineering, 15(6), pp. 469-73.

[2] Aleem, I. S, Jalal, H., Sheikh, A. A. (2009) "Clinical decision analysis: Incorporating the evidence with patient preferences". In: Patient Preference and Adherence. Vol. 3, pp. 21-24.

[3] Bae,J-M. (2014) “The clinical decision analysis using decision tree”. In: Epidemiology and Health. Vol. 36, pg 1-7.

[4] Bishop, C. M. (1995) "Neural Networks for Pattern Recognition". Oxford: Oxford University Press.

[5] Bonner, G. (2001) “Decision making for health care professionals: use of decision trees within the community mental health setting”. In: Journal of Advanced Nursing, vol. 35, pp. 349-356.

[6] Brasil (2006) “Prevenção clínica de doenças cardiovasculares, cerebrovasculares e renais”. Brasília: Ministério da Saúde.

[7] Costa, F. O. de, Motta, L. C. S. and Nogueira, J. L. T. (2010) “Uma abordagem baseada em Redes Neurais Artificiais para o auxílio ao diagnóstico de doenças meningocócicas”. In: Revista Brasileira de Computação Aplicada, v. 2, n. 1, pp. 79-88.

[8] De Soarez, P. C. (2009) “Use of decision analysis in the programs of vaccination against varicella”. São Paulo: “Faculdade de Medicina da Universidade de São Paulo”.

[9] Fayyad, U. M. (1996) "Data mining and knowledge discovery: making sense out of data". IEEE Expert: IEEE

Expert: Intelligent Systems and Their Applications. Vol. 11, Issue 5, pp. 20-25.

[10] Hann, J. and Kamber, M. (2006) "Data Mining: Concepts and Techniques". Second Edition. San Franciso, CA : Morgan Kaufmann.

[11] Iyer, A., Jeyalatha, S. and Sumbaly, R. (2015) “Diagnosis of Diabetes Using Classification Mining Techniques”. International Journal of Data Mining & Knowledge Management Process (IJDKP). Vol.5, No.1, pp. 1-14.

[12] Junga, S-K. and Kim, T-W. (2016) “New approach for the diagnosis of extractions with neural network machine learning”. American Journal of Orthodontics and Dentofacial Orthopedics, Vol 149, Issue 1, pp. 127-133.

[13] Koh, H. and Tan, G. (2005) “Data Mining Applications in Healthcare”. In: Journal of Healthcare Information, v. 19, pp.64-72.

[14] Kutlu, Y., Isler, Y., Kuntalp, D. and Kuntalp, M. (2006) “Detection of Spikes with Multiple Layer Perceptron Network Structures” In: Signal Processing and Communications Applications, 2006.

[15] Lahner, E., Intraligi, M., Buscema, M. (2008) “Artificial neural networks in the recognition of the presence of thyroid disease in patients with atrophic body gastritis”, World Journal of Gastroenterology. Vol. 14, pp. 563-8.

[16] Lesch, K. P. (2004) “Serotonergic gene expression and depression: implications for developing novel antidepressants”. In: Journal of Affective Disorders, vol. 62, p.57-76.

[17] Letourneau, S., Jensen, L. (2008) "Impact of a decision tree on chronic wound care". Journal Wound Ostomy Continence Nurs, vol. 25, pp. 240-247.

[18] Lewis, D. D. (2005) “Naive (Bayes) at forty: The independence assumption in information retrieval”. In: Lecture Notes in Computer Science, v. 1398, pp. 4-15.

[19] Lima, B. P., Morais, C. A., Contrera, D., Casale, G., Pereira, M., Gronner, M., Diogo, T., Torquarto, M., Oishi, J. and Leal, A. (2003) “Prevalence of diabetes mellitus and impaired glucose tolerance in the urban population aged 30-69 years in Ribeirão Preto (São Paulo), Brazil”. In: São Paulo Med. J., v.121, pp.224-230.

[20] Liping, A. and Lingyun, T. (2005) “A rough neural expert system for medical diagnosis, Services Systems and Services Management”, vol. 2, pp. 1130-1135.

[21] Mitchell, T. (1997) "Machine Learning". New York: McGraw-Hill.

[22] Moreira, T. (2013) “Narrativas de pessoas com diabetes atendidas na rede básica: determinantes da hospitalização”. Dissertação de Mestrado. UFB.

[23] Nesto, R. (2004) “Correlation between cardiovascular disease and diabetes mellitus: current concepts”. In: The American Journal of Medicine, v. 116, Issue 5.

[24] Portal, B. (2016) “Doenças cardiovasculares causam quase 30% das mortes no País”. http://www.brasil.gov.br/saude. Acesso em: 05 de junho de 2016.

[25] Quinlan, C. (1993) “C4.5: Programs for machine learning”. Morgan Kaufmann.

[26] Saraee, M. and Ritchings, T. (2004) “Medical data mining: case of cervical cancer screening”. In: METMBS 04, 21-24 June 2004.

[27] Souza, R. T. (2013) “Avaliação de classificadores na classificação de radiografias de tórax para o diagnóstico de pneumonia infantil”. Dissertação de Mestrado. Programa de Pós-graduação em Ciência da Computação (INF), Universidade Federal de Goiás, pp. 63.

[28] Witten, I. and Frank, E. (2005) “Data Mining: Practical machine learning tools and techniques”. In Morgan Kaufmann.

[29] Yusuf, S., Reddys, S., Ôunpuu, S. and Anand S. (2001) “Global burden of cardiovascular diseases - part I: general considerations, the epidemiologic transition, risk factors, and impact of urbanization”, In: Clinical Cardiology: New Frontiers, v. 104, pp. 2746–2753

[30] Zorman, M., Podgorelec, V., Kokol, P., Peterson and M., Lane, J. (2000) “Decision tree's induction strategies evaluated on a hard real world problem”. In: Proceedings of the 13th IEEE Symposium on Computer-Based Medical Systems CBMS'2000, pp. 19-24.

[31] Hall, M., Frank, E., Holmes, G. Pfahringer, B., Reutemann, P., Witten, I. H. (2009) “The WEKA Data Mining Software: An Update”. SIGKDD Explorations, Volume 11, Issue 1.