Estudo Sobre o Uso de Árvores de Decisão na Área da Saúde

Cynthia Moreira Maia[1], Julio Cartier Maia Gomes[1], Luana Dantas Chagas[1]

[1]Universidade Federal Rural do Semi-Árido (UFERSA)
CEP 59515-000 – Angicos – RN – Brazil

Departamento de Ciências Exatas, Tecnológicas e Humanas – DCETH
Universidade Federal Rural do Semi-Árido (UFERSA) – Angicos, RN – Brazil

cynthia-norte@hotmail.com, juliocartier@gmail.com,
luana.dantas@ufersa.edu.br

Abstract. *This article presents a research of papers that use the decision tree technique applied in databases in the health area. The main databases of academic publications were consulted, where the most relevant results were found. Such research demonstrates the several advantages provided by the use of the decision tree technique in health.*

Resumo. *Neste artigo é apresentada uma pesquisa de trabalhos que utilizam a técnica de árvore de decisão aplicada em bases de dados na área da saúde. Foram consultadas as principais bases de publicações acadêmicas, onde se constatou-se as pesquisas com resultados mais relevantes. Tais pesquisas demonstram as diversas vantagens proporcionadas pelo uso da técnica de árvores de decisões na saúde.*

1. Introdução

Na década de 80, chegou ao Brasil a Internet. Uma rede mundial de computadores que permitiu a comunicação global, possibilitando ao usuário o acesso a informações globalizadas de forma rápida e precisa. Desde 2013, a população global da Internet cresceu cerca de 20% - de 2,4 bilhões para 3,2 bilhões de pessoas. A todo instante há pessoas compartilhando textos, fotos, áudios, vídeos, postando tweets, realizando compras na web, os dados vão sendo gerados o tempo todo sem que se perceba. A cada minuto do dia os usuários do Facebook postam 4.166.667 mensagens, no Twitter os usuários enviam 347.222 tweets, na Netflix os assinantes transmitem 77,160 horas de vídeo, já á Amazon recebe 4,310 visitantes únicos (DOMO 2015).

Em 2011, foi realizado o trabalho de Witten et al. (2011), explanando que a cada dia uma enorme quantidade de dados é gerada. Uma estimativa de que a cada 20 meses dobra a quantidade de dados armazenada nos banco de dados do mundo. Diante dessa grande quantidade de dados, é possível realizar análises que permitam descobrir relacionamentos entre os dados que estão de forma oculta, podendo gerar informações valiosas que visam auxiliar na solução de diversos problemas. Isso acontece porque muitas informações, possivelmente úteis, podem estar sendo perdidas, ficando ocultas nos banco de dados.

Na área de finanças tem-se problemas como previsão de falências, detecção de fraudes e análise de risco de crédito. Na agropecuária, dentre os problemas

que podem ser tratados, podem ser citados: definição de logísticas de armazenamento e transporte, redução de perdas e desperdícios, barateamento e expansão da produção agrícola. Na Bioinformática, tem-se a análise da forma de proteínas, localização de proteínas no meio celular, identificação de genes em sequências de DNA, entre outros (Faceli et al. 2011)**.** Assim, muitas organizações analisam constantemente seus dados para obtenção de sucesso na tomada de decisão. Para tal, existem técnicas e algoritmos que contribuem nesse processo. A área que estuda essas técnicas e algoritmos denomina-se mineração de dados.

A mineração de dados (do inglês, *Data Mining*) consiste na aplicação de técnicas inteligentes e algoritmos para exposição de conhecimentos que estão implícitos (Han & Kamber 2006). Isso significa que é possível extrair informações que não estão perceptíveis a nível de processamento humano. Por exemplo, em uma base de dados de um hospital, ao analisar os dados, seria possível descobrir que "homens com idade de 17 anos possui maior probabilidade ter anemia". Os principais algoritmos existentes para a mineração de dados podem ser categorizados em: regras de classificação, regressão, regras de associação (Witten & Frank 2005).

Nesse trabalho será abordado a regra de classificação em Árvore de Decisão. A classificação é o processo que localiza um determinado grupo de funções que descrevem e diferenciam classes ou conceitos, com o intuito de utilizar o modelo localizado para prever a classe de objetos que ainda não foi classificado. Por exemplo, suponha que o gerente de um supermercado está interessado em descobrir que tipo de características de seus clientes os classificam em bom comprador ou mau comprador (De Amo 2003). As Árvores de Decisão são representações simples do conhecimento e um meio eficiente de construir classificadores que predizem classes baseadas nos valores de atributos de um conjunto de dados. Uma Árvore de Decisão usa a estratégia dividir para conquistar para resolver um problema de decisão. Dessa forma, um problema complexo é dividido em problemas mais simples, aos quais recursivamente é aplicada a mesma estratégia. As soluções dos subproblemas podem ser combinadas, na forma de uma árvore, para produzir uma solução do problema complexo (Faceli et al. 2011).

A Mineração de dados tem sido amplamente utilizada em diferentes áreas, principalmente na medicina, indústria, marketing, agronegócios, educação, entre outras (Berry & Linoff 2004; Dunstone & Yager 2008; Stone et al. 2008; Silva et al. 2008; Romero & Ventura 2010). Na área da saúde, essa estratégia tem ganhado bastante receptividade, sendo utilizada para controle de epidemias, auxílio a exames clínicos, monitoramento do estado de pacientes, dosagem de medicamentos e auxílio para o diagnostico médico. Nesse contexto, é apresentado um estudo sobre Árvore de Decisão aplicada na área da saúde, promovendo um embasamento para os benefícios desta nessa área.

2. Mineração de dados

Os dados são informações que não foram tratadas e organizadas de forma compreensível para sua utilização, ou seja, está na forma sua bruta. A informação, por sua vez, são os dados que foram processados para um formato útil. E o conhecimento, é a capacidade de agir com as determinadas informações que foram geradas. Dessa forma, mineração de dados permite por meio da aplicação de técnicas a "descoberta de informações" presentes nos bancos de dados. Uma área que envolve banco de dados, métodos probabilísticos de estatística e aprendizado de máquina na descoberta de

informações úteis. A mineração de dados é parte principal de um processo que tem como entrada uma base de dados e como saída uma informação (Fayyad et al. 1996).

2.1 Etapas para Descoberta de Conhecimento

Existem algumas etapas para a extração de conhecimentos a partir de um conjunto de dados, como mostrado na Figura 1. De acordo com Rezende (2005), a primeira etapa consiste no conhecimento do domínio, que é a identificação do problema, bem como os objetivos esperados. Na etapa de pré-processamento, será realizado uma limpeza nos dados, evitando dados inconsistentes, com ruídos, desbalanceados, incompletos e/ou redundantes. A etapa de extração de padrões compreende a escolha da tarefa de mineração de dados a ser empregada, a escolha do algoritmo e a extração dos padrões propriamente dita. A etapa de Pós-processamento, consiste na análise, avaliação e na apresentação de forma compreensível aos humanos. Ao final, tem-se a utilização do conhecimento descoberto.

Figura 1- Etapas do Processo de Extração do Conhecimento (Rezende, Pugliesi, Melanda & Paula 2003).

2.2 Tarefa de Classificação

As regras de classificação baseiam-se na construção de um modelo com caráter preditivo dos dados.

De acordo com Tan & Steinbach & Kumar (2006), um modelo de classificação é útil para as finalidades:

- Modelagem Descritiva: Funciona como uma ferramenta explicativa dos dados analisados.
- Modelagem Preditiva: É utilizado para prever o rótulo de classes de registros não conhecidos.

2.2.1 Árvore de Decisão

Uma Árvore de Decisão é uma árvore onde cada nó interno (não terminal) representa um teste ou decisão sobre o item de dado considerado (Goebel & Gruenwald 1999). Um exemplo pode ser mostrado na figura a seguir.

Uma Árvore de Decisão é uma árvore onde cada nó interno (não terminal) representa um teste ou decisão sobre o item de dado considerado (Goebel & Gruenwald 1999). Um exemplo pode ser mostrado na figura a seguir.

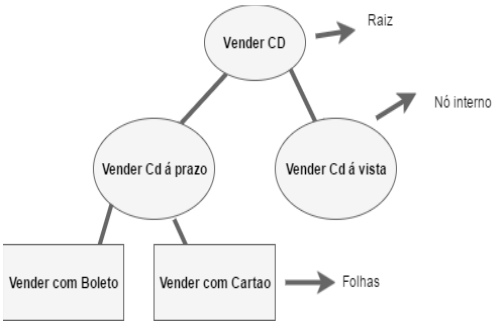

Figura 2- Modelo Árvore de Decisão Simples

Assim a Árvore de Decisão pode ser representada como um conjunto de regras que serão direcionadas de acordo com as condições impostas na árvore. A árvore é construída de forma recursiva, partindo da raiz e dependendo do resultado do teste usado pelo nó a árvore se direciona para os nós filhos, repetindo o método até que seja concluído quando o nó interno é alcançado. A Figura 2 representa uma situação de venda de Cds. A raiz seria Vender Cd, dividindo esse problema em subproblemas, tem-se que o Cd será vendido a prazo ou a cartão, significando os nós internos. Se for vendido a prazo, será direcionado para duas opções, vender com boleto ou vender com cartão, correspondendo as folhas. Assim de acordo com as decisões é possível traçar uma solução do problema.

Para construir a Árvore de Decisão, foram desenvolvidos alguns algoritmos, como: ID3 (Quinlan 1979), CART (Breiman et al. 1984), C4.5 (Quinlan 1993), abrangendo outros. Os principais passos do algoritmo para construção de uma Árvore de Decisão podem ser descritos no algoritmo 1, como mostrado a seguir. Obtendo como entrada para a função GeraÁrvore um conjunto de dados D. No passo 3, o algoritmo avalia o critério de parada. Se mais divisões do conjunto de dados são necessárias, é escolhido o atributo que maximiza alguma medida de impureza, descrito no passo 5. No passo 7, a função GeraÁrvore é recursivamente aplicada a cada partição do conjunto de dados D (Faceli et al. 2011).

1. Algoritmo Básico para construção da Árvore de Decisão (Faceli et al. 2011)

1. Algoritmo para construção de uma Árvore de Decisão
Entrada: Um conjunto de treinamento **D** = {(xi,yi), i=1,…,n} **Saída:** Árvore de Decisão **1** /* **Função GeraÁrvore (D)** * /; **2** **se** critério de parada **(D)** = Verdadeiro **então** **3** **Retorna:** um nó folha rotulado com a constante que minimiza a função perda; **4** **fim** **5** Escolha o atributo que maximiza o critério de divisão em **D**; **6** **para cada** partição dos exemplos **D**, baseado nos valores do atributo escolhido **faça** **7** Induz uma subárvore Árvorei = GeraÁrvore **(Di)**; **8** **fim** **9** **Retorna:** Árvore contendo um nó de decisão baseado no atributo escolhido, e descendentes Árvorei;

Observa-se de maneira simples a utilização de um algoritmo básico na construção da Árvore de Decisão a partir de um conjunto de dados. Na próxima seção será descrito brevemente alguns trabalhos utilizando algoritmos na construção de Árvores de Decisão voltados para a área da saúde.

4. Árvore de Decisão na Área da Saúde

Vários trabalhos foram realizados utilizando a mineração de dados na área da saúde. Essa seção visa mostrar os trabalhos em que se constatou resultados relevantes na utilização da técnica de Árvore de Decisão na saúde.

4.1 Diagnóstico de Asma

O trabalho de Morais *et al* (2012) propôs o sistema InteliMED. Este trata-se da construção de um sistema móvel de suporte remoto para diagnósticos médicos iniciais, desenvolvendo uma solução para apoiar o diagnóstico clínico de asma. Técnicas de mineração de dados foram aplicadas, produzindo uma Árvore de Decisão capaz de sugerir um diagnóstico com uma taxa de acerto de 91,61% para asma. O algoritmo utilizado para indução da árvore foi o J48. O trabalho não disponibilizou as informações da matriz de confusão.

4.2 Análise dos Algoritmos de Mineração J48 e *Apriori* Aplicados na Detecção de Indicadores da Qualidade de Vida e Saúde

Em Librelotto & Mozzaquatro (2013), é apresentado um estudo comparativo entre duas técnicas de mineração de dados, a técnica de classificação e associação: J48 e Apriori, aplicadas na identificação e classificação de indicadores de saúde na geração de perfis de usuários. De acordo com os resultados, o algoritmo de classificação apresentou grau de certeza de 91,61% e o grau de incerteza de 8,37% em comparação com o de associação que apresentou grau de certeza de 80% e o grau de incerteza de 20%. Ocorreu um diferencial de 11,61% apontando o algoritmo J48 como o mais eficaz na identificação e geração de perfis dos usuários. O trabalho não disponibilizou as informações da matriz de confusão.

4.3 Modelo de Suporte à Decisão Aplicado à Identificação de Indivíduos Não Aderentes ao Tratamento Anti-Hipertensivo

O trabalho de Medeiros et al. (2014), teve como objetivo desenvolver um modelo de apoio à tomada de decisão para identificar indivíduos não aderentes ao tratamento anti-hipertensivo. Propondo a utilização de Árvore de Decisão sobre um banco de dados de adesão ao tratamento envolvendo 118 usuários hipertensos de uma Unidade Básica. A matriz de classificação detalha os acertos e erros encontrados na Árvore de Decisão. Com 21 acertos entre os 31 pacientes classificados como aderentes e 82 acertos entre os 87 pacientes classificados como não aderentes. Ou seja, positivos verdadeiros que correspondem a 21, negativos verdadeiros que correspondem a 82, falsos negativos que correspondem 10 e falsos positivos que correspondem 5. Obteve-se como resultado uma árvore capaz de classificar corretamente 87,28% dos indivíduos. O modelo de Árvore de Decisão proposto auxilia na identificação de usuários não aderentes, de modo a contribuir com as equipes de saúde na abordagem a esses indivíduos.

4.4 Uso de Técnicas de Mineração de Dados Para a Identificação Automática de Beneficiários Propensos ao Diabetes Mellitus Tipo 2

O trabalho de Carvalho & Dallagassa & Silva (2015), propôs um modelo baseado em

técnicas de mineração de dados para a identificação automática de beneficiários com propensão a doenças crônicas. Utilizou uma base de dados de uma operadora de plano de saúde do estado do Paraná. O algoritmo utilizado foi o J48. Para o processo de mineração de dados foram selecionadas 12 variáveis, considerando um conjunto de 43.375 beneficiários, sendo descobertas 843 regras, com uma taxa de acerto de 88,9%. A matriz de classificação gerou 12.296 registros classificados corretamente entre os 12.771 classificados sem indicativo da doença e 822 acertos entre os 1.976 classificados com indicativo da doença. O trabalho apresentou a matriz de confusão, verdadeiro positivo que correspondem a 12.296, falso positivo que correspondem a 475, falso negativo que correspondem a 1.154, verdadeiro negativo que correspondem a 822. O modelo mostrou-se eficiente podendo ser aplicado para a seleção de beneficiários para programas de prevenção de doenças crônicas..

4.5 Integração Entre Serviços de Saúde no Cuidado as Pessoas Vivendo com AIDS

No trabalho de Medeiros et al. (2015) objetivou-se construir um modelo de suporte à decisão para auxiliar profissionais do serviço de saúde especializado a identificar os padrões de comportamento no uso dos serviços da Estratégia Saúde da Família (ESF) das pessoas vivendo com HIV/AIDS. O modelo foi criado a partir de um banco com dados de 141 pessoas portadoras de AIDS, de um serviço ambulatorial especializado referência no estado da Paraíba (PB). Foi utilizado o Weka para gerar a Árvore de Decisão, empregando o J48. A matriz de classificação detalha os acertos e erros encontrados na Árvore de Decisão. Com 23 indivíduos classificados corretamente entre os 40 classificados de maneira satisfatória e 90 acertos entre os 101 classificados insatisfatoriamente. O trabalho apresentou a matriz de confusão, verdadeiro positivo que correspondem a 23, falso positivo que correspondem a 17, falso negativo que correspondem a 11, verdadeiro negativo que correspondem a 90. O modelo possibilitou o estabelecimento de 23 regras, com um percentual de acerto de 80,1%, as quais poderão dar suporte a tomada de decisão dos profissionais na identificação de situações onde se apresenta a necessidade de estimular a utilização da Estratégia Saúde da Família pelos usuários.

5. Conclusão

Este trabalho conduz um estudo sobre a utilização da técnica de Árvore de Decisão na área da saúde, permitindo encontrar conhecimentos mediantes a construção de Árvores de Decisão a partir de bases de dados associadas à área da saúde. Com o estudo realizado, foi possível concluir que a aplicação do modelo de Árvore de Decisão facilita o entendimento dos analistas na identificação dos possíveis caminhos alternativos a serem seguidos para que se possa realizar determinados propósitos definidos. Apresentando análises interativas dos resultados de forma simples e clara, por meio da observação da árvore, facilitando o entendimento do problema em questão, com alto grau de interpretabilidade.

Os resultados do estudo foram valiosos como observado, podendo descobrir informações interessantes, que possam auxiliar os gestores da saúde no diagnóstico de doenças, podendo dar suporte a tomada de decisão desses profissionais. A grande aplicabilidade de Árvores de Decisão na saúde se dá devido a sua flexibilidade, robustez e interpretabilidade. As inúmeras vantagens intensificam a importância da utilização deste método na extração de conhecimento em bases de dados na área da saúde, ajudando a validar a eficiência que seria difícil de ser verificado somente a partir de

análises visuais e consultas simples a essas bases de dados. Como proposta de trabalhos futuros, o trabalho despertou o interesse de realizar testes em bases de dados da saúde por meio da técnica de classificação, Árvore de Decisão, para validar sua eficiência, como constatado nos trabalhos desse estudo.

6. Referências

Berry, M. J. Linoff, G. S. (2004). Data Mining Techniques: For Marketing, Sales, And Customer Relationship Management. New York: John Wiley e Sons.

Breiman, L. Friedman, J. H. Olshen, R. A. Stone, C. I. (1984). Classification and regression trees. Belmont, Calif.: Wadsworth.

Carvalho, D. R. Dallagassa, M. R. Da Silva, S. H. (2016). Uso de Técnicas de Mineração de Dados para a Identificação Automática de Beneficiários Propensos ao Diabetes Mellitus Tipo 2. Informação & Informação,[s.l.], v. 20, n. 3, p.274-296. Universidade Estadual de Londrina.

De Amo, S. (2003). Curso de Data Mining, disponível em <www.deamo.prof.ufu. br/CursoDM.html>, último acesso em 26 de fevereiro de 2017.

De Morais, D. C. S. et al. (2012). Sistema Móvel de Apoio a Decisão Médica Aplicado ao Diagnóstico de Asma – InteliMED.

De Medeiros, L. B. et al. (2015). Integration of health services in the care of people living with aids: an approach using a decision tree.

Dunstone, T. Yager, N. (2008) Biometric System And Data Analysis: Design, evaluation, and data mining. New York: Springer.

Faceli, K. et al. (2011) Inteligência Artificial: Uma Abordagem de Aprendizagem de Máquina. Rio de Janeiro: LTC.

Fayyad, U. Piatetsky-Shapiro, G. Smyth, P. (1996) The KDD Process For Extracting Useful Knowledge From Volumes Of Data. In: Communications of the ACM, 39(11), 27-34.

Goebel, M. Gruenwald, L. (1999). A Survey of Data Mining and Knowledge Discovery Software Tools. ACM SIGKDD Explorations, New York, v. 1, no. 1, p. 20-33.

Han, J. Kamber, M. (2006) Data Mining. Concepts and Techniques. Second edition. The Morgan Kaufmann Series in Data Management Systems. Elsevier Inc.

James, J. (2015) Data Never Sleeps 3.0.

Librelotto, S. R. Mozzaquatro, P. M. (2013). Análise dos Algoritmos de Mineração J48 e Apriori Aplicados na Detecção de Indicadores da Qualidade de Vida e Saúde. Revista Interdisciplinar de Ensino, Pesquisa e Extensão, v. 1, n. 1, p.1-12, 2013.

Medeiros, A. R. C. et al. (2014) Modelo de suporte à decisão aplicado à identificação de indivíduos não aderentes ao tratamento anti-hipertensivo. Saúde Debate**,** Rio de Janeiro, v. 38.

Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc., San Mateo, CA,USA.

Quinlan, J. R. (1979). Discovering rules by induction from large collections of examples. In D. Michie (Ed.), Expert systems in the microelectronic Age. Edinburgh University Press.

Rezende, S. O. Pugliesi, J. B. Melanda, E. A. Paula, M. F. (2003). Mineração de dados. In S. O. Rezende (Ed.), Sistemas Inteligentes – Fundamentos e Aplicações, pp. 307–335. Editora Manole.

Rezende, S. O. (2005). Mineração de Dados.

Romero, C. Ventura, S. (2010) Educational Data Mining: A Review Of The State Of The Art. In: Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 40(6), 601-618.

Silva, L. A. Rangayyan, R. M. Hernandez, E. D. M. (2008). Classification of breast masses using a committee machine of artificial neural networks. Journal of Electronic Imaging, 17(1), 013017-013017.

Stone, Z. Zickler, T. Darrell, T. (2008) Autotagging Facebook: Social Network Context Improves Photo Annotation. In: Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference, pp. 1-8.

Tan, P. Steinbach, M. Kumar, V. (2006) Introduction to Data Mining. 1a edição. Michigan State University, University of Minnesota, Army High Performance Computing Research Center,USA.

Weka 3 - Data Maning Software in Java The University of Waikato, disponível em: <http://www.cs.waikato.ac.nz/ml/weka/> último acesso em: dezembro,2016.

Witten, I. H. Frank, E. (2005) Data mining: practical machine learning tools. 2. ed. San Francisco, CA: Morgan Kaufmann.

Witten, I. H. Frank, E. Hall, M. A. (2001) . Data Mining: Practical machine learning tools and techniques. 3ª. ed. San Francisco: Morgan Kaufmann.