O Algoritmo de Classificação C4.5 e suas Aplicações na Área Médica

Maicon Bastos Palhano[1], Gabriel Felippe[1], Ruano Marques Pereira[1], Priscyla Waleska T. de Azevedo Simões[1], Merisandra Cortês de Mattos Garcia[1]

[1]Grupo de Pesquisa em Inteligência Computacional Aplicada/Curso Ciência da Computação/UNACET/UNESC – Criciúma – Santa Catarina - Brasil

{gabrielheavy, ruanopereira}@hotmail.com, {maiconpalhano,pri, mem}@unes.net

Abstract. *Technological developments in the area of data storage and the consequent accumulation of these brought the need to acquire useful knowledge from these repositories of data. Hence came the data mining technique with several useful tasks for obtaining knowledge, such as classification. Among their methods has been the decision trees and algorithms that implement different. In this article we will address the C4.5 algorithm and some examples of their applications in databases in the medical field in order to assist in decision making.*

Keywords: *Data Mining, Classification Task, Decision Tree, C4.5 Algorithm.*

Resumo. *A evolução tecnológica na área de armazenamento de dados e o conseqüente acúmulo destes, trouxe a necessidade de adquirir conhecimento útil a partir desses repositórios de dados. Sendo assim, surgiu a técnica de data mining com diversas tarefas úteis para obtenção de conhecimento, como por exemplo, a classificação. Dentre os seus métodos tem-se o de árvores de decisão e diferentes algoritmos que o implementam. Neste artigo será abordado o algoritmo C4.5 e exemplos de algumas aplicações suas em bases de dados da área médica, a fim de auxiliar na tomada de decisão.*

Palavras-chave*: Data Mining, Tarefa de Classificação, Árvores de Decisão, Algoritmo C4.5.*

1. Introdução

A exploração de dados pode ocorrer por meio do *Data Mining* (DM) que serve para extrair conhecimentos dos dados e descobrir relações significativas previamente desconhecidas [1].

O DM é um processo que envolve diversas tarefas, tendo-se a classificação que consiste no processo de busca de modelos ou funções que possam descrever e distinguir relações ou conceitos entre os dados, com o propósito de prever características entre informações ainda não armazenadas na base. (HAN; KAMBER, 2001, tradução nossa).

A publicação do algoritmo C4.5 foi realizada em 1987, tendo como desenvolvedor John Ross Quinlan [2].

O algoritmo tem como objetivo gerar um modelo classificador na forma de uma árvore de decisão, apresentando dois estados durante o processo, os quais são: folha que indica um ponto no final da classificação, sendo atribuída a uma classe e nó de decisão, onde baseando-se no atributo em análise, poderá conter uma ramificação seguida de uma folha ou uma sub-árvore para cada possível valor encontrado na base [2].

2. Algumas aplicações do algoritmo C4.5 na área médica

Neste artigo foi realizada uma pesquisa em dois periódicos internacionais recentes, relacionada à aplicabilidade de *data mining* na área médica, em especial da tarefa de classificação utilizando o algoritmo C4.5, o critério utilizado para a escolha dos periódicos foi o qualis do periódico.

O uso de técnicas para classificação de sangue e hemoglobina pelo algoritmo C4.5 para a triagem de talassemia (doença hereditária autossômica recessiva que afeta o sangue) tem se mostrado uma técnica adequada, mostrando-se precisa no processo de teste clínico, onde envolveram 1000 amostras e 13 classes de anormalidade de talassemia. Com dois parâmetros importantes, hemograma completo e tipo de hemoglobina, pode-se representar diferentes aspectos do sangue e cada um desses atributos contribuiu com informações para certos tipos de diagnósticos [3].

Outra aplicação envolve o uso do método de árvore de decisão com diversos algoritmos, sendo que o C4.5 foi um dos aplicados, com o intuito de criar novos índices de prognósticos para a diferenciação de subgrupos de pessoas com câncer de mama. Uma análise retrospectiva foi feita em 381 pacientes com este câncer e diversos atributos foram avaliados, como idade, histórico de câncer familiar, tamanho do tumor, entre outros. Considerando-se o fator de prognóstico, novos índices foram identificados, sendo que o algoritmo C4.5 obteve um resultado melhor em relação a outros, como CHAID, QUEST e ID3, mostrando que a aplicabilidade do algoritmo traz uma nova possibilidade de melhoria nos prognósticos [4].

3. Considerações finais

O *data mining* pode ser aplicado em diversas áreas, como por exemplo na medicina, onde percebe-se a utilidade da sua aplicação, pois como visto pelos artigos apresentados o algoritmo C4.5 pode tanto classificar pacientes doentes e não doentes, como classificar uma doença em vários tipos, prever o prognóstico da doença, possibilitando assim se afirmar que o algoritmo pode auxiliar na tomada de decisão médica, tanto no diagnóstico como prognóstico de doenças, sendo uma forma de auxílio a prática e pesquisa médica, auxiliando na tomada de decisão dos especialistas.

Referências

[1] WITTEN, I. H.; FRANK, Eibe. **Data Mining:** Pratical Machine Learning Tools and Techniques. San Francisco: Morgan Kaufmann, 2005. 525 p.

[2] QUINLAN, J. R. ***C4.5:*** *Programs for Machine Learning.* Morgan Kaufmann Publishers, 1993.

[3] DAMRONGRIT S. et al. Classification of complete blood count and haemoglobin typing data by a C4.5 decision tree, a naïve Bayes classifier and a multilayer perceptron for thalassaemia screening, Biomedical Signal Processing and Control, Volume 7, Issue 2, March 2012, Pages 202-212.

[4] MEVLUT T.; FUSUN T.; IMRAN K. O. The comparisons of prognostic indexes using data mining techniques and Cox regression analysis in the breast cancer data, Expert Systems with Applications, Volume 36, Issue 4, May 2009, Pages 8247-8254.