APLICAÇÃO DO PROCESSO COMPUTACIONAL DE MINERAÇÃO DE DADOS PARA IDENTIFICAR REGRAS DE DECISÃO EM DIABETES

Silvani Weber da Silva Borges (PIBIC/ UNIOESTE)[1], Feng Chung Wu (Orientador)[12], Newton Spolaôr[1], Huei Diana Lee[12]

[1]Universidade Estadual do Oeste do Paraná (UNIOESTE)
[2]Universidade Estadual de Campinas (UNICAMP)

{silvani.borges2, wufengchung, newtonspolaor, hueidianalee}@gmail.com

Objetivos

Aplicar a Mineração de Dados (MD) para obter árvores com regras de decisão que podem apoiar o diagnóstico de Diabetes Mellitus (DM).

Métodos e Procedimentos

Neste trabalho foi adotado o conjunto de dados *Pima Indians* (https://goo.gl/cSHsA0). *Pima* possui 768 registros de pacientes, descritos por 8 sinais e características (atributos), e rotulados como positivo ou negativo para o teste de DM.

Pima foi submetido à MD, a qual inclui recursos matemáticos e de computação para extrair padrões e anomalias dos dados que podem auxiliar especialistas do domínio em tomadas de decisão custosas ou subjetivas. Para tanto, foram conduzidas 3 fases: pré-processamento, extração de padrões e pós-processamento [1].

Na fase (1), o algoritmo ReliefF foi usado para ranquear sinais e características que melhor diferenciam registros positivos de negativos.

Na fase (2), o algoritmo de classificação J48 foi aplicado para aprender árvores de decisão que testam um grupo de atributos de pacientes para apoiar o diagnóstico de DM.

Na fase (3), árvores geradas usando 50% dos Atributos Ranqueados (AR) e 100% dos atributos (*Baseline*) foram avaliadas conforme a validação cruzada 10 *folds*. A coerência das regras dessas árvores foi também analisada.

Resultados

Na Tabela 1 são apresentados a média (e o desvio padrão) da taxa de erro de classificação.

Tabela 1. Erro das árvores de decisão geradas.

AR	26,58 (6,48)
Baseline	25,65 (3,78)

Ao não encontrar diferença estatisticamente significativa (teste *t* não emparelhado, α=0,05), ReliefF se destaca por reduzir 50% dos atributos candidatos a regras de decisão e por manter um desempenho preditivo competitivo.

Uma regra obtida de AR sobressai ao rotular com sucesso 129 registros (16,80% do total): "*se nível glicêmico <= 127 mg/dl e índice de massa corporal <= 26,4 kg/m², então negativo para diabetes*". Essa e outras regras envolvem condições de risco que aumentam a resistência à insulina e diminuem a função pancreática [2].

Conclusões

A MD gera regras potencialmente úteis em DM, motivando o uso de algoritmos mais precisos e a sua aplicação em outros domínios.

Referências Bibliográficas

[1] LEE, H. D.; Seleção de atributos importantes para extração de conhecimento de bases de dados. Tese de doutorado. USP, Brasil, 2005.

[2] ROBBINS,S.L.; COTRAN, R.S.; KUMAR, V. (2016). *Bases patológicas das doenças.* Rio de Janeiro: Elsevier.

APPLICATION OF THE DATA MINING PROCESS TO IDENTIFY DECISION RULES IN DIABETES

Silvani Weber da Silva Borges (PIBIC/ UNIOESTE)[1], Feng Chung Wu (Orientador)[12], Newton Spolaôr[1], Huei Diana Lee[12]

[1]Western Paraná State University (UNIOESTE)
[2]University of Campinas (UNICAMP)

{silvani.borges2, wufengchung, newtonspolaor, hueidianalee}@gmail.com

Objective

To apply the Data Mining (DM) computational process to build trees with decision rules that can support Diabetes mellitus (DI) diagnosis.

Materials and Methods

In this work, the *Pima Indians* dataset (https://goo.gl/cSHsA0) was studied. *Pima* has 768 records of patients data, described by 8 signals and characteristics (features), and labeled as positive or negative for the DI test.

Pima was submitted to DM, which includes mathematical and computational resources to extract patterns and anomalies from data in order to support domain experts in costly or subjective decision taking processes. To do so, 3 DM phases were carried out: pre-processing, pattern extraction and post-processing [1].

Phase 1: the ReliefF algorithm was used to rank signals and characteristics that better differentiate positive from negative records.

Phase 2: the J48 classification algorithm was applied to learn decision trees that test a group of patients' features to support DI diagnosis.

Phase 3: trees built using 50% of the Ranked Features (RF) and 100% of the features (*Baseline*) were evaluated according to the 10-*fold cross validation strategy*. The coherence of the rules from these trees was also analyzed.

Results

Table 1 presents the average (and the standard deviation) classification error rate.

Table 1. Error rate of the decision trees built.

RF	26,58 (6,48)
Baseline	25,65 (3,78)

By not finding statistically significant difference (unpaired *t* test, α=0,05), ReliefF is highlighted by reducing 50% of the features candidates for decision rules and by keeping a competitive predictive performance.

A rule inherent to the RF-based tree stands out by correctly labelling 129 records (16,80% of the total): "*if blood sugar level <= 127 mg/dl and body mass index <= 26,4 kg/m², then negative for diabetes*". This rule and other ones include risk factors that enlarge insulin resistance and reduce the pancreatic function [2].

Conclusions

DM builds rules potentially useful for DI test, motivating the use of more precise algorithms and the DM application in other domains.

References

[1] LEE, H. D.; Selecting important features for knowledge discovery in databases (in Portuguese). PhD Thesis. USP, Brazil, 2005.

[2] ROBBINS,S.L.; COTRAN, R.S.; KUMAR, V. (2014). *Pathologic basis of disease.* Rio de Janeiro: Elsevier.