Técnicas de Classificação em Problemas Relacionados a Doenças Cardíacas

André Pinz Borges

apborges@utfpr.edu.br

Douglas Guisi, Jonatas Almeida Jr., Marcelo Teixeira, Richardson Ribeiro, Gabriel Souza e Hudson Lapa

UTFPR – Pato Branco

Roteiro

- Introdução
- Sistemas de Aprendizagem
- Metodologia
- Resultados
- Conclusões
- Trabalhos futuros

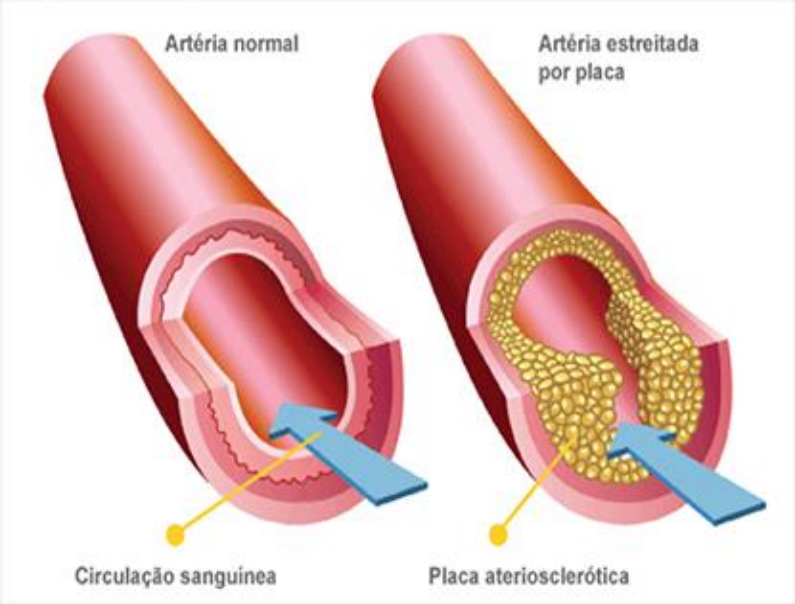
Introdução

- Objetivo: aplicar técnicas de classificação em problemas relacionados à doenças cardíacas

- Doenças cardiovasculares: distúrbios do coração e dos vasos sanguíneos
  - Ex: aterosclerose (acúmulo de gordura)

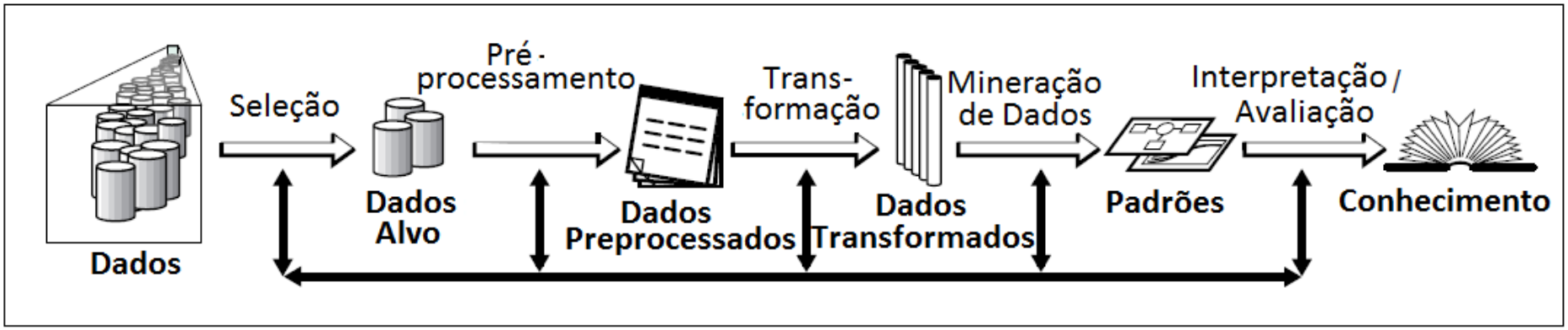
- Representam 30% dos óbitos no Brasil [1]

- Hábitos como sedentarismo, tabagismo e diabetes [2, 3]



Introdução

- Consultas, exames, prescrições médicas, diagnósticos, vacinações são tarefas dos profissionais de saúde
- Resultado: conjunto de dados na ordem dos milhares
- Como obter conhecimento a partir destes dados?
  - Descoberta de conhecimento em bases de dados [4]
    - *Knowledge Discovery in Databases* - KDD



Aprendizagem com Mineração de Dados

- Auxílio na tomada de decisões [5]
  - Prescrições, exames, encaminhamentos, causas de doenças, sintomas, tratamentos....
- Sistemas formados por técnicas e procedimentos que se baseiam em **algoritmos** aplicados em grupos de dados para **extrair conhecimento** [4]
- Tipos de algoritmos:
  - Agrupamento:
  - Classificação
  - Associação
- Permitem a criação de regras sobre os dados

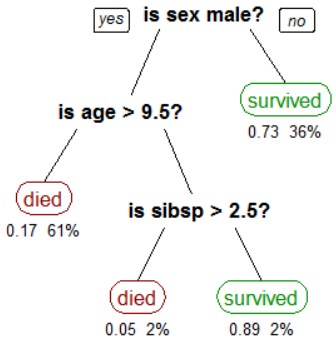
Aprendizagem com Mineração de Dados

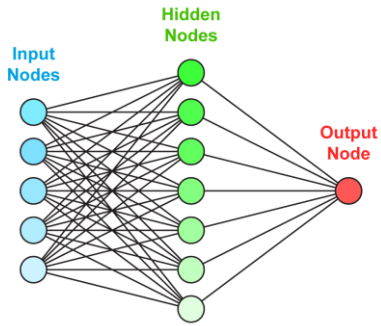
- Uso de algoritmos de classificação para encontrar modelos (funções) usados para predizer objetos ainda não classificados

  - **Árvores de decisão**: divide um problema em partes menores
    - Geração de uma árvore usada para categorizar novas instâncias
    - Custos de vacinas, diagnósticos..

  - **Naive Bayes**: classificador probabilístico usado para calcular independência de variáveis

  - **Redes Neurais**: previsor de eventos inspirado na estrutura dos neurônios humanos





Metodologia

Etapas:

1. Seleção de escopo de doenças
2. Descoberta de relacionamentos
3. Procura por bases de dados em repositórios
4. União das bases
5. Aplicação de técnicas de mineração de dados

Metodologia

1. Seleção de escopo de doenças
2. Descoberta de relacionamentos

- Estudou-se as principais causas de cada doença, com foco para aquelas que se repetiam em diferentes doenças.

- Principais resultados:
  - Histórico familiar, níveis de colesterol HDL e LDL,
  - Hábitos (sedentarismo, alcoolismo, stress, ...)

Metodologia

3. Procura por bases de dados em repositórios

- Dados obtidos de repositórios públicos (UCI)
  - Heart-statlog: 270 instâncias e 13 atributos
  - Diabetes: 9 atributos 768 instâncias
  - Arrhytmia: 452 instâncias e 279 atributos
- Dados numéricos (atributos com valores inteiros ou reais)
- Remoção dos atributos iguais

Metodologia

4. União das bases

- Bases selecionadas: diabetes e parada cardíaca
- A união foi feita eliminando apenas atributos repetidos em ambas (pré-processamento).
  - sexo e pressão sanguínea
- Para popular a nova base concatenaram-se os dados já existentes nas bases anteriores
  - Pois esses dados já respeitavam a faixa de valores permitida para cada atributo
- A base final é constituída por um conjunto de 263 instâncias, as quais foram submetidas aos algoritmos de classificação.

Metodologia

5. Aplicação de técnicas de mineração de dados

- Software *Weka* para descoberta do conhecimento
- Linguagem JAVA para implementação dos algoritmos

```
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
```

Exemplo de JAVA + Weka para Mineração de Dados

Resultados

- Avaliação: divisão em dados de treinamento e testes
  - Validação cruzada *k-fold* para avaliar a capacidade de generalização e aprendizagem (*10-fold*)
  - Conjunto de dados é dividido em *k* subconjuntos
    - *k-1* são usados para treinamento e 1 para testes
    - Processo repetido *k* vezes até que todos os subconjuntos de dados sejam usados no conjunto de testes.
    - Diferentes classificadores são obtidos e a precisão das classificações pode ser avaliada.

Resultados

- Algoritmos utilizados para testes:
  - J48
  - Rede Neural Perceptron Multicamadas
  - Naive Bayes
  - Classificação via Regressão
  - Logística Bayesiana com Regressão

Resultados

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

Tabela 1. Percentual de acerto dos algoritmos na detecção de diabetes.

Resultados

- Base numérica
  - Tornou propícia a aplicação de um algoritmo de regressão para a classificação dos dados
  - Resultou em melhor classificação deste tipo de algoritmo em relação aos demais

Conclusões

- Falta de sistemas comerciais da área da saúde que usam técnicas de MD
  - Soluções para prontuários eletrônicos, suporte à decisão...
- Não é possível fazer MD com dados de um país inteiro
- Dados não são disponibilizados
- Possíveis frentes:
  - Uso de novas bases de dados para descoberta de doenças
  - Estender a metodologia para novas doenças
  - Interpretação dos resultados com profissionais de saúde
  - Firmar parcerias para novas pesquisas
  - ...

Referências

[1] 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.

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

[3] 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.

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

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

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

Técnicas de Classificação em Problemas Relacionados a Doenças Cardíacas

André Pinz Borges

apborges@utfpr.edu.br

Douglas Guisi, Jonatas Almeida Jr., Marcelo Teixeira, Richardson Ribeiro, Gabriel Souza e Hudson Lapa

UTFPR – Pato Branco