

Introdução a Modelagem e Aprendizado

ELTDI - DAELN UTFPR-CT

Prof. Heitor S. Lopes (2023) - Exercício #1

Objetivo: explorar a classificação de dados com árvores e regras de decisão

Parte 1:

- Faça *download* do *dataset hepatitis* disponível no *UCI Machine Learning Repository*. Este *dataset* tem a finalidade de diagnóstico de hepatite. Adeque o *dataset* para ser lido no software Orange.
- Utilize as ferramentas gráficas do Orange para obter *insights* acerca dos dados.
- Com o Orange, obtenha os valores da acurácia com os métodos *baseline* ZeroR e OneR, e também da árvore de decisão sobre o conjunto de treinamento (utilize os parâmetros *default*)
- Explique por quê os métodos *baselines* obtêm uma porcentagem tão alta de instâncias corretamente classificadas.
- Com base no resultado do algoritmo OneR informe qual, entre todos os atributos do *dataset*, é o mais relevante para o diagnóstico da *hepatite*.
- Com base na árvore de decisão gerada, enumere quais atributos do dataset são irrelevantes para o diagnóstico.
- Analise os resultados obtidos pelos métodos utilizados sob os pontos de vista de compreensibilidade da árvore gerada e das métricas de qualidade da classificação. Considerando a importância relativa das classes DIE e LIVE para o diagnóstico do paciente, qual o método mais adequado para este dataset?

Parte 2:

- O *dataset soybean*, disponível no site da disciplina, se refere ao diagnóstico de 19 doenças comuns da soja. Ele tem 35 atributos e 683 instâncias. Faça o *upload* deste *dataset* no Orange.
- Utilizando a ferramenta de visualização *Distributions*, o que é possível preliminarmente inferir sobre o dataset?
- Selecione a coluna “*class*” como o atributo-alvo (classe) e todos os demais como atributos precursores. Use validação cruzada estratificada de *5-folds* para o treinamento de uma Árvore de Decisão com todos os parâmetros *default*. Informe o tamanho da árvore obtida (número total de nós e número de nós-folhas) e as medidas de qualidade

(acurácia, *precision*, *recall* e *F1 score*). Justifique qual a medida de qualidade adequada para este caso.

- d. Mostre a matriz de confusão gerada pelo treinamento/teste da árvore de decisão. Observando a árvore, identifique quais foram as classes que tiveram 100% e 0% de acerto, respectivamente. Justifique este comportamento (em especial para as classes com 0% de acerto).
- e. Repita o item (c) agora utilizando Regras de Decisão. Analise as regras obtidas de acordo com a sua qualidade (informe a métrica utilizada) e compreensibilidade.
- f. Compare qualitativamente os resultados obtidos com os dois métodos (árvore de decisão e regras de classificação)