Introdução a Modelagem e Aprendizado
ELTD1 - DAELN UTFPR-CT

Prof. Heitor S. Lopes (2024) - Exercício #1

Objetivo: explorar a classificação de dados com árvores de decisão

Parte 1:

a. Faça *download* do *dataset hepatitis* disponível no UCI *Machine Learning Repository*. Este *dataset* tem a finalidade de diagnóstico de hepatite. Faça as adequações necessárias do *dataset* para ser lido no software Orange.

b. Utilize as ferramentas gráficas do Orange para obter *insights* acerca dos dados.

c. Separe os dados em conjunto de treinamento e teste, na proporção de 80/20 dos dados.

d. Com o Orange, utilize o conjunto de treinamento e uma árvore de decisão com parâmetros padrão para a classificação dos dados. Observe a árvore de decisão gerada e informe: qual o atributo (variável) que tem maior valor preditivo, bem como os atributos que são irrelevantes para o diagnóstico.

e. Com a árvore treinada no item anterior, calcule os valores da acurácia (com os parâmetros *default*) no conjunto de treinamento. Primeiramente, utilize apenas o atributo previsor mais importante (isto é chamado de OneRule), depois com a árvore de decisão completa.

f. Comparando os resultados do item anterior, explique por quê utilizando apenas o OneRule se obtém uma alta porcentagem de instâncias corretamente classificadas.

g. 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:

a. 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.

b. Utilizando a ferramenta de visualização *Distributions*, o que é possível preliminarmente inferir sobre o dataset?

c. Selecione a coluna "*class*" como o atributo-alvo (classe) e todos os demais como atributos previsores. 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. Utilizando o Orange, gere um conjunto de regras com a melhor acurácia e cobertura possível para este conjunto de dados. Discuta, qualitativamente, estes resultados quando comparados a árvores de decisão anteriormente criadas, no que se refere a compreensibilidade e qualidade.