

Curso: Mestrado em Engenharia Biomédica**U.C.:** Aprendizagem e Extração do Conhecimento

Ficha de Exercícios 05	
Docente:	Hugo Peixoto José Machado
Tema:	Explorar o WEKA
Ano Letivo:	2021-2022 – 1º Semestre
Duração da aula:	2 horas

1. Parte I

O dataset usado neste exercício é o dataset de doenças cardíacas disponível no ficheiro `heart-c.arff`, obtido no repositório da UCI. Este dataset descreve fatores de risco para doenças cardíacas. O atributo num representa o atributo da classe (binária):

class <50 - nenhuma doença

class > 50_1 - aumento do nível de doença cardíaca.

O principal objetivo deste exercício é prever doenças cardíacas a partir de outros atributos no dataset. Obviamente, trata-se de um problema de classificação. O software a ser usado é o Weka. A descrição deste exercício é gradual. Portanto, espera-se que possa entender melhor os vários aspetos e questões envolvidos no processo de KDD.

1. Data Understanding

O primeiro passo para abordar o problema é familiarizar-se com os dados.

Responder às seguintes perguntas ajudará a entender melhor os dados. O dataset `heart-c.arff` contém algumas informações sobre os dados que armazena. Pode abri-lo num editor de texto.

De seguida carregue o dataset no Weka.

[1] Para cada atributo, encontre as seguintes informações:

[a] O tipo de atributo (p.e. nominal, ordinal, numérico.).

[b] Percentagem de valores ausentes nos dados.

[c] Máx, min, média e desvio padrão (se aplicável).

[d] Existem instâncias que tenham um valor para um determinado atributo que nenhuma outra instância tem, i.e. registos únicos?

[e] Estude o histograma no canto inferior direito e descreva informalmente como o atributo parece influenciar o risco de doença cardíaca. Ao passar o rato pelos gráficos consegue observar umas labels, o que significam?

[2] Mude para o separador *Visualize*, na parte superior da janela, para visualizar gráficos de dispersão 2D para cada par de atributos.

[a] O atributo `thalach` parece estar mais/menos associado a doenças cardíacas? E o atributo `ca`?

[b] Observando os gráficos encontre qual o atributo que parece estar mais correlacionado com o atributo `thalach`. Essa correlação é positiva ou negativa?

[3] Observando o gráfico $\text{thalach}(X) / \text{oldpeak}(y)$ Investigue uma possível associação desses atributos com o atributo `class`, ou seja, tente identificar possíveis áreas “densas” de doenças cardíacas (se existirem).

2. Data Processing

A segunda etapa diz respeito ao processamento dos dados de modo que os dados transformados estejam numa forma mais adequada para os algoritmos de data mining. Todos as alíneas devem partir do ficheiro original.

[1] Seleção de atributos.

Investigue a possibilidade de usar o filtro Weka `AttributeSelection` para selecionar um subconjunto de atributos com boa capacidade de previsão. Em seguida, compare os resultados obtidos com as conclusões obtidas na seção anterior. Guarde o conjunto de dados com os atributos selecionados no ficheiro [heart-c1.arff](#).

[2] Lidar com valores ausentes.

Os registos não deverão ser eliminados e é aconselhável atribuir valores onde faltam dados, usando um método adequado. Considere os seguintes métodos para lidar com valores ausentes e investigue cada possibilidade no Weka.

[a] Substitua os valores ausentes pela média do atributo, se o atributo for numérico. Caso contrário, substitua os valores ausentes pela moda do atributo (se o atributo for nominal). Guarde o conjunto de dados que obteve sem valores ausentes no ficheiro [heart-c2.arff](#).

[b] Investigue a possibilidade de usar regressão (linear) para estimar os valores ausentes para cada atributo. Guarde o conjunto de dados que obteve sem valores ausentes no ficheiro [heart-c3.arff](#).

[3] Eliminar outliers.

Investigue a possibilidade de usar o filtro Weka: *Unsupervised -> Attribute -> InterquartileRange* para detectar outliers. (Neste caso para a deteção de outliers deverá apenas selecionar atributos numéricos).

Ao efetuar a aplicação do filtro encontrará instâncias classificadas como outliers. Reabra o ficheiro `heart-c.arff` e elimine as linhas correspondentes e guarde o conjunto de dados obtido sem outliers no ficheiro [heart-c4.arff](#).

3. Modeling

O terceiro passo é usar algoritmos de classificação disponíveis no Weka para descobrir padrões ocultos nos dados. Deve repetir as etapas descritas abaixo para cada um dos conjuntos de dados criados durante o pré-processamento, além de usar também o dataset original.

[1] Comece com o classificador `OneR`.

[a] O que pode concluir? Compare as suas conclusões com as conclusões que obteve na seção 1.1.

[b] Compare a precisão do classificador obtida no conjunto de treino (training set) com a estimativa de precisão obtida através do método 10 fold-cross validation. Como explica esta diferença (se existir)?

[2] Use o classificador JRip, ou seja, a versão Weka do classificador de regras RIPPER.

[a] Crie um classificador com e sem rule pruning. Qual se comporta melhor? Justifique a sua resposta.

[3] Use o classificador J48, ou seja, a versão Weka do classificador C4.5 da árvore de decisão.

[a] Explore o uso de diferentes parâmetros no algoritmo J48, como pruning("unpruned") e número mínimo de registos nas folhas("minNumObj").

[b] Descreva os padrões que obteve e compare com as conclusões obtidas nas questões anteriores.

[4] Explore outros algoritmos de classificação e vá guardando os resultados.

4. Evaluation

Na etapa [3] construiu vários modelos. Por fim, é necessário comparar os diferentes modelos e apresentar as suas conclusões.

[1] O Weka oferece várias medidas de avaliação de desempenho. Escolha algumas medidas de desempenho e justifique a sua escolha.

[2] Resuma numa tabela as medidas de desempenho para cada classificador e cada dataset.

[3] O que pode concluir?