Ficha de Exercícios 02

Gonçalo Rodrigues Pinto - A83732
Universidade do Minho

(05 de Março de 2021)

Resumo

O presente relatório descreve o trabalho de introdução ao Weka. O presente trabalho teve como objetivo a identificação das principais características de um determinado dataset através do ecrã de pré processamento bem como correr algoritmos de classificação sobre o mesmo. De forma ao seu desenvolvimento foi colocado um conjunto de questões sobre vários datasets de exemplo disponibilizados pelo Weka. Assim foi possível através de uma ferramenta que possui uma coleção de algoritmos de machine learning retirar diversas conclusões sobre os dados presentes no datasets utilizados através da execução de tarefas *Data Mining*.

1 Introdução

No 2º semestre do 1º ano do Mestrado em Engenharia Informática da Universidade do Minho, existe uma unidade curricular enquadrada no perfil de Descoberta do Conhecimento denominada por Engenharia do Conhecimento, que tem como objetivo a introdução ao conceito de descoberta do conhecimento.

A presente ficha enquadra-se nesta unidade curricular e pretende introduzir ao Weka(Waikato Environment for Knowledge Analysis) que possui uma coleção de algoritmos de machine learning para execução de tarefas de Data Mining. É um software que permite pré-processar grandes volumes de dados, aplicar diferentes algoritmos de Machine Learning e comparar vários outputs.

Nesta ficha pretendeu-se saber instalar e configurar o Weka para a primeira utilização, carregar os datasets de exemplo no "Explorer", identificar principais características do dataset através do ecrã de pré processamento e correr algoritmos de classificação.

2 Questões

1. ***Abrir o Weka / Explorer e carregar o data set "contact-lens.arff". Com este dataset carregado responda às seguintes questões:***

   (a) ***Quantas instâncias (registos) tem este dataset?***
   Possui 24 instâncias este dataset.

   (b) ***Quantos atributos (colunas) tem este dataset?***
   Possui 5 atributos este dataset.

   (c) ***Quantos e quais os valores possíveis para o atributo "age"?***
   No atributo "age"existe 8 valores do tipo young, 8 valores do tipo pre-presbyopic e 8 do presbyopic.

   (d) ***Quais os valores possíveis para o atributo "contact-lens"?***
   Os valores possíveis para o atributo "contact-lens"são soft, hard e none.

   (e) ***Qual o atributo que tem "reduced" como um dos valores?***
   O atributo que tem "reduced" como um dos valores é tear-prod-rate.

2. ***Abrir o Weka/Explorer e carregar o data set "iris.arff". Com este dataset carregado responda às seguintes questões:***

   (a) ***Quantas instâncias registos tem este dataset?***
   Possui 150 instâncias este dataset.

   (b) ***Quantos atributos (colunas) tem este dataset?***
   Possui 5 atributos este dataset.

   (c) ***A classe "iris-setosa" tende a ter maiores ou menores valores de "sepal.length"?***
   A classe "iris-setosa" tende a ter menores valores de "sepal.length".

   (d) ***A classe "iris-viginica" tende a ter maiores ou menores valores de "petal.width"?***
   A classe "iris-viginica" tende a ter maiores valores de "petal.width".

   (e) ***Qual destes atributos, sozinho, parece dar uma melhor indicação da "class"?***
   O atributo, que sozinho, parece dar uma melhor indicação da "class" é o "petallength".

3. ***Abrir o Weka/Explorer e carregar o data set "weather.nominal.arff". Com este dataset carregado responda às seguintes questões:***

   (a) ***Identificar quais os atributos deste dataset?***
   Os atributos deste dataset são *"outlook", "temperature", "himidity", "windy"e "play"*.

(b) ***A utilização de um algoritmo de classificação poderá trazer conhecimento específico através dos dados apresentados. Indique um objetivo que possa ser atingido com a aplicação de algoritmos de classificação, quando o mesmo for executado em dados semelhantes mas previamente desconhecidos***
Neste caso é concluir se é possível jogar mediante as condições meterológicas.

4. ***Abrir o Weka/Explorer e carregar o data set "glass.arff" . Com este dataset carregado responda às seguintes questões:***

   (a) ***Abrir o separador "Classify" e escolher o algoritmo J48 ("trees")***
   Algoritmo executado com opções por *default* e com *cross-validation* a 10 *folds*.

   (b) ***Observar a "Confusion Matrix" e indicar quais as maiores falhas no processo de classificação.***

```
=== Confusion Matrix ===

  a  b  c  d  e  f  g   <-- classified as
 50 15  3  0  0  1  1 |  a = build wind float
 16 47  6  0  2  3  2 |  b = build wind non-float
  5  5  6  0  0  1  0 |  c = vehic wind float
  0  0  0  0  0  0  0 |  d = vehic wind non-float
  0  2  0  0 10  0  1 |  e = containers
  1  1  0  0  0  7  0 |  f = tableware
  3  2  0  0  0  1 23 |  g = headlamps
```

Figura 1: *"Confusion Matrix"*

As maiores falhas encontram-se na classificação de a ("build wind float") para b ("build wind non-float").

   (c) ***Qual o número de "headlamps" que foram classificadas como "build wind float"?***
   O número de "headlamps" que foram classificadas como "build wind float" foram 3.

   (d) ***Qual o número de instâncias classificadas corretamente como "vehic wind non-float"?***
   O número de instâncias classificadas corretamente como "vehic wind non-float" foram 0.

   (e) ***Qual o número de instâncias classificadas corretamente como "vehic wind float"?***
   O número de instâncias classificadas corretamente como "vehic wind float" foram 6.

(f) ***Na lista de resultados obtidos clique com o botão direito e selecione "Visualize tree". Copiar os resultados para a ficha de solução e descrever sucintamente o processo de classificação do algoritmo.***

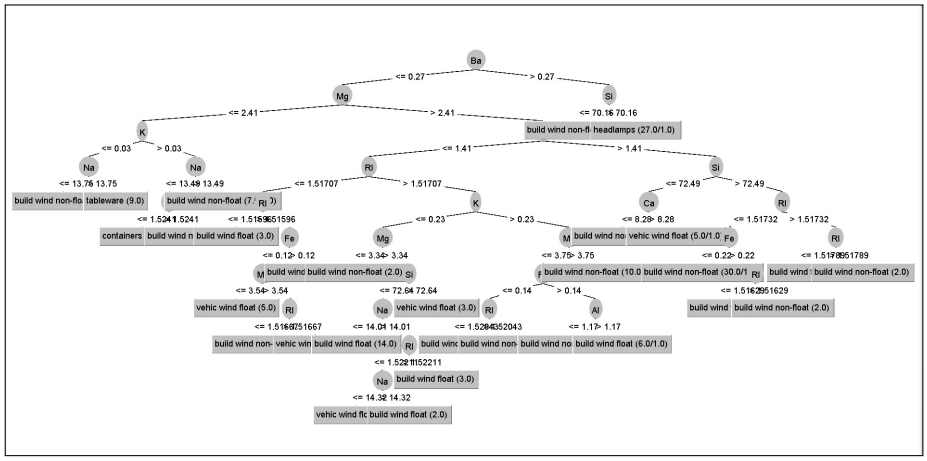

Figura 2: *Resultados obtidos da opção "Visualize tree".*

Neste caso de estudo o algoritmo classifica a informação pelos valores numéricos dos atributos, organizando na árvore abaixo apresentada. Percorrendo os ramos é possível tomar decisões quando um novo caso surgir através dos valores (atributos) desse caso. Por exemplo, foi considerado que o elemento fulcral na construção do vidro é o Bário (Ba) caso este valor seja superior a 0.27 e o Silício superior a 70.16 é aconselhável optar pela "headlamps".

5. ***Abrir o Weka / Explorer e carregar o data set "labor.arff" . Com este dataset carregado responda às seguintes questões:***

   (a) ***Correr o algoritmo de classificação J48 com os parâmetros por defeito. Indicar a percentagem de instâncias corretamente classificadas.***
   A percentagem de instâncias corretamente classificadas foi 73.6842%.

   (b) ***Utilizando somente 2 casas decimais, abra a configuração do algoritmo J48 e coloque a opção "unpruned" a "True". Corra novamente a classificação e indique a percentagem de instâncias corretamente classificadas***
   A percentagem de instâncias corretamente classificadas foi 78.9474%.

6. ***Abrir o Weka / Explorer e carregar novamente o data set "glass.arff". Com este dataset carregado responda às seguintes questões:***

   (a) ***Retirar o atributo "Fe". Qual o resultado da classificação?***
   O resultado da classificação obtido foi 144 instâncias classificadas como corretas, representado 67.2897% dos registos por outro lado foi obtido 70 registos incorretos (32.7103%).

   (b) ***Retirar todos excepto "Ri", "Mg". Qual o resultado da classificação?***
   O resultado da classificação obtido foi 147 instâncias classificadas como corretas, representado 68.6916% dos registos por outro lado foi obtido 67 registos incorretos (31.3084%).

3 Conclusão

O presente relatório descreveu, de forma sucinta, o trabalho de introdução ao Weka.

Após a realização deste trabalho, compreendi a instalação e configuração do Weka para a primeira utilização, como também foi possível aprender a carregar os datasets de exemplo no "Explorer" além de que permitiu identificar principais características do dataset através do ecrã de pré processamento bem como correr algoritmos de classificação .

Por fim, espero que os conhecimentos obtidos e consolidados sejam de enorme utilidade tendo uma perspetiva futura.