Universidade do Minho

Mestrado Integrado em Engenharia Biomédica
Sistemas de Aprendizagem e Extração de Conhecimento
4º Ano, 2º Semestre
Ano letivo 2012/2013

Ficha prática – Introdução ao WEKA
Março, 2013

Tema

Análise e Extração de Conhecimento.

Objectivos de aprendizagem

Com a realização desta ficha prática pretende-se que os alunos:

- Adquiram noções básicas sobre análise e extração de conhecimento de dados;
- Identifiquem os objetivos a alcançar no processo de extração de conhecimento (classificação, associação, segmentação, e outros).

Enunciado

Usando o ficheiro de dados com informação sobre os clientes de uma instituição bancária que aderiram a um produto financeiro publicitado por meios eletrónicos (Tabela 1), realize os passos indicados ao longo desta ficha prática, com vista à familiarização com o ambiente de exploração de dados WEKA (Waikato Environment for Knowledge Analysis – www.cs.waikato.ac.nz/~ml/weka).

Tabela 1

Excerto do conjunto de dados («data set») a analisar. O conjunto de dados completo está disponível eletronicamente.

id	age	sex	region	income	married	children	car	save_act	current_act	mortgage	pep
ID12101	48	FEMALE	INNER_CITY	17546.0	NO	1	NO	NO	NO	NO	YES
ID12102	40	MALE	TOWN	30085.1	YES	3	YES	NO	YES	YES	NO
ID12103	51	FEMALE	INNER_CITY	16575.4	YES	0	YES	YES	YES	NO	NO
ID12104	23	FEMALE	TOWN	20375.4	YES	3	NO	NO	YES	NO	NO
ID12105	57	FEMALE	RURAL	50576.3	YES	0	NO	YES	NO	NO	NO
ID12106	57	FEMALE	TOWN	37869.6	YES	2	NO	YES	YES	NO	YES
ID12107	22	MALE	RURAL	8877.07	NO	0	NO	NO	YES	NO	YES
ID12108	58	MALE	TOWN	24946.6	YES	0	YES	YES	YES	NO	NO
ID12109	37	FEMALE	SUBURBAN	25304.3	YES	2	YES	NO	NO	NO	NO
ID12110	54	MALE	TOWN	24212.1	YES	2	YES	YES	YES	NO	NO
ID12111	66	FEMALE	TOWN	59803.9	YES	0	NO	YES	YES	NO	NO
...	...	...	...	...	...	...	...	...	...	...	...

O objetivo deste cenário é o de induzir um perfil para caracterizar os clientes, potenciais alvos desta campanha.

Sequência de passos a realizar:

1. Obter o ficheiro "bank-data.csv";

2. Analisar o conteúdo do ficheiro com um editor de texto (Notepad/Bloco de Notas), verificando:
   i) atributos;
   ii) tipos de dados;
   iii) valores;
   iv) número de casos.

3. Analisar o conteúdo do ficheiro com uma folha de cálculo e produzir algum tipo de análise sobre os dados, como por exemplo:
   i) contador de ocorrências sobre alguns atributos (sexo, idade, região);
   ii) contador de ocorrências por classes (idade entre 0-10, 11-20, ...)
   iii) valores médios globais (idade média, vencimento médio);
   iv) valores médios por atributo (média da idade dos homens e das mulheres);
   v) gráficos ilustrativos dos dados ou dos resultados calculados.

4. Analisar o conteúdo do ficheiro com o ambiente de análise de conhecimento WEKA:
   i) na janela de preparação dos dados ("Preprocess"):
      - verificar que o atributo "id" não apresenta qualquer utilidade em termos analíticos;
      - a secção "Current relation" apresenta um resumo da informação do conjunto de dados;
      - a secção "Attributes" permite visualizar e manipular os atributos presentes no conjunto de dados;
      - a secção "Selected attribute" resume os dados do atributo seleccionado e visualiza graficamente esse conteúdo.
   ii) ainda na janela "Preprocess", remover o atributo "id";
   iii) gravar o conjunto de dados resultante em formato .ARFF (formato nativo do WEKA), por exemplo, "bank-data.arff";
   iv) usar um editor de texto para consultar o conteúdo deste ficheiro.

5. Preparação de dados:
   i) editar o ficheiro "bank-data.arff" com um editor de texto;
   ii) para o atributo "children", substituir o termo "numeric" pela lista de valores {0, 1, 2, 3};
   iii) gravar o ficheiro com esta alteração;
   iv) utilizar de novo o ambiente WEKA para consultar o ficheiro "bank-data.arff";
   v) enumerar («discretize») os atributos numéricos "age" e "income":
      - na janela "Preprocess", na secção "Filter", carregar no botão "Choose" e seleccionar "weka/filters/unsupervised/attribute/discretize";
      - clicar no nome do filtro agora seleccionado para aceder às suas propriedades;
      - clicar no botão "More" (explorar);
      - na propriedade "bins", alterar o seu valor para 3;
      - aceitar, carregando "OK";
      - na secção "Filter", clicar "Apply", para submeter o conjunto de dados ao filtro programado;
      - gravar o resultado num novo ficheiro, por exemplo, "bank-data-discrete.arff";
      - explorar este novo ficheiro com um editor de texto.

6. Regras de Associação:
   i) utilizar o ambiente WEKA para consultar o ficheiro "bank-data-discrete.arff";
   ii) seleccionar a janela de pesquisa de associações "Associate";
   iii) na secção "Associator" está indicado o algoritmo de associação "Apriori" com os respectivos parâmetros pré-definidos;
   iv) editar opções dos parâmetros, clicando sobre o nome do algoritmo de associação;
   v) explorar através do botão "More";
   vi) procurar obter diversas associações, para um número máximo de regras (por exemplo 100);
   vii) experimentar com diversas métricas;

7. Segmentação («clustering»)
   i) utilizar o ambiente WEKA para consultar o ficheiro “bank-data.arff”;
   ii) seleccionar a janela de segmentação “Cluster”;
   iii) na secção “Clusterer”, clicar em “Choose” para seleccionar o processo segmentador;
   iv) escolher o algoritmo de segmentação “weka/clustererers/SimpleKMeans”;
   v) clicar nos parâmetros identificados por defeito e alterar o número de segmentos (“clusters”) para 6;
   vi) iniciar a execução do algoritmo (“Start”);
   vii) o resultado do algoritmo apresenta o «centroid» de cada segmento;
   viii) na secção “Result list”, clicando com o botão direito, surge um conjunto de opções que permitem visualizar os resultados graficamente.

8. Classificação
   i) utilizar o ambiente WEKA para consultar o ficheiro “bank-data-training.arff”;
   ii) aceder à janela de classificação “Classify”;
   iii) na secção “Classifier”, clicar em “Choose” e seleccionar o algoritmo de classificação “weka/classifiers/trees/J48”, e aceitar os parâmetros identificados por defeito;
   iv) iniciar a execução do algoritmo de classificação (“Start”);
   v) na secção “Result list”, clicando com o botão direito, surge um conjunto de opções que permitem visualizar os resultados graficamente;
   vi) nas opções que surgem, escolher “visualize tree”;
   vii) experimentação com ficheiro de teste “bank-data-test.arff”:
      - na secção “Test options”, marcar “Supplied test set“ e carregar o ficheiro de testes;
      - para ver os exemplos classificados, na secção “Result list” clique com o botão direito e seleccione “Visualize classifier errors”;
      - clicar no botão “Save” para gravar os resultados em “bank-data-test-errors.arff”;
      - utilizar um editor de texto para visualizar este ficheiro.