IR2 – Data Mining 2002–2003
Exercises II - WEKA

Peter Lucas
Institute for Computing and Information Sciences
University of Nijmegen

1 Introduction

The purpose of this practical is to let you build up experience in the practical issues involved in the use of data-mining tools for data analysis. Use will be made of WEKA (Waikato Environment for Knowledge Analysis)[1], which is a typical example of a suite of data-mining tools, in this case available in the public domain, and very similar to commercially available packages. It has the advantage of being completely OS independent, as it is based on standard Java. (It runs very well under Linux, currently a very popular OS in the data-mining community.)

Before continuing, you first have to down load and install Java from the SUN Java site, as well as the WEKA package.

2 Datasets and data analysis

Recall that the analysis of any dataset is only possible if the person carrying out the analysis is sufficiently versed in the problem domain, and has access to domain experts who are able (and willing!) to assist in the interpretation of achieved results.

During the practical, there are therefore teams consisting of:

- yourself, the data-mining expert, possibly with a colleague;
- Peter Lucas, MD: the medical expert.

Each team will tackle a number of problems:

1. prediction of life expectancy for patients with hepatitis, and
2. prediction of life expectancy for patients with non-Hodgkin lymphoma (NHL) of the stomach.

using symbolic *classification* techniques from the area of machine learning. The two mentioned problems seem very similar in the sense that both are medical and have to do with life expectancy. However, there are also differences: hepatitis is usually due to a viral infection, whereas non-Hodgkin lymphoma is a form of cancer. Thus, the analyses will also give insight into differences between these two types of disease.

The datasets can be down loaded at:

[1] `http://www.cs.waikato.ac.nz/ml/weka`

- `http://www.cs.kun.nl/~peterl/teaching/DM/hepatitis.arff`
- `http://www.cs.kun.nl/~peterl/teaching/DM/nhl.arff`

The analytical techniques we will focus on for now are the rule-learning algorithms zeroR, oneR, and PRISM, and the two tree-learning schemes ID3 and C4.5 (called J4.8 in WEKA). These have been covered during the second lecture.[2]

3 The data-mining process

3.1 Visualisation

Inspect the relationships between variables in each of the datasets using the data visualiser of WEKA. Such simple inspections can yield much insight.

▶ *Which two variables in the* `hepatitis` *dataset are clearly related to each other? Try to answer the same question for the* `nhl` *dataset.*

3.2 Preprocessing

Even if a dataset is delivered in the right format, it may still need preprocessing in order to be able to apply a learning method. Preprocessing involves:

- selection of relevant variables or attributes;
- discretisation of the domain of continuous variables if the learning method is only able to deal with discrete variables;
- dealing with missing values if the learning method is unable to do so itself.

Preprocessing in WEKA is done both manually (e.g. by ticking relevant attributes) and automatically. Automatic preprocessing is done by *filters*. These have to be selected from the `Filters` menu. Discretisation of domains is done by the `DiscretizeFilter`. Handling of missing values is done by the `ReplaceMissingValues` filter. Filters are added, and resulting new datasets can replace the input dataset in an analysis. The idea is similar to the concept of pipes in Linux/Unix.

▶ *Investigate which filters are required for the analysis of each of the datasets (*`hepatitis` *and* `nhl`*). Apply these filters and replace the datasets by the generated datasets.*

3.3 Learning classification models

We will only use the classification part of WEKA in the following experiments. Use the `Percentage split` to obtain *training* and *test sets*

▶ *Experiment with the percentage of split.*

In this set of machine-learning experiments, the results obtained by particular machine-learning parameters, such as attributes selected and machine-learning schemes, are studied. The aim is to obtain insight into how particular choices affect the results.

[2]Details can be found at `http://www.cs.kun.nl/~peterl/teaching/DM/classifiers4.ps.gz`

3.3.1 Analysis of hepatitis data

- *After selecting appropriate preprocessing steps, apply the zeroR, oneR and PRISM algorithms to the hepatitis dataset. Compare the results, and save the results you think are significant. Which variables are most significant in predicting outcome of hepatitis, and which learning method yields the best results?*
- *Do the same as mentioned above, but now using ID3 and J4.8. Again, explain your preferences.*

3.3.2 Analysis of NHL data

- *Again, after selecting appropriate preprocessing steps, apply the zeroR, oneR and PRISM algorithms, but now to the NHL dataset. Compare the results, and save those results you think are significant. Which variables are most significant in predicting outcome of treatment of NHL of the stomach, and which learning method yields the best results?*
- *Answer the same questions as above, but now using ID3 and J4.8. Again, explain your answers.*

3.3.3 Comparison

Compare the results obtained for the two datasets by the five learning methods. You can do this as follows

- *Construct a matrix of the following form:*

	ML method				
Dataset	zeroR	oneR	PRISM	ID3	J4.8
Hepatitis					
NHL					

for each of the preprocessing choices, and fill in details about performance.