IR2 – Data Mining 2002–2003

Assignment

Peter Lucas
Institute for Computing and Information Sciences
University of Nijmegen

1 What are you requested to do?

In this assignment you are expected to use various data-mining techniques in analysing a single dataset. 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 best under Linux, though, currently a very popular OS in the data-mining community.)

You are expected to develop a prognostic model of the disease non-Hodgkin lymphoma of the stomach using each of the following techniques:

1. base-line techniques: zeroR and oneR
2. symbolic techniques: PRISM, ID3 and J4.8 (i.e. C4.5)
3. statistical techniques: naive Bayesian classifiers and logistic regression

For each of these techniques, you are expected to determine the achieved performance for the test set:

1. when the test set is assumed to be equal to the training set;
2. using a split of the dataset into a a 66% training set and 34% test set;
3. using 10-fold cross validation.

The results achieved should be summarised, compared for the various techniques and methods of evaluation, and discussed in a report, approximately 2 pages in size. In the report try to explain the significance of your results for helping doctors in the selection of appropriate treatment for patients with non-Hodgkin lymphoma of the stomach.

2 The problem: prognosis of NHL of the stomach

2.1 The dataset

The dataset you will use in this assignment can be down loaded at:

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

The dataset is about primary non-Hodgkin lymphoma of the stomach.

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

2.2 Non-Hodgkin lymphoma of the stomach

Primary non-Hodgkin lymphoma (NHL) of the stomach is a relatively rare malignant disorder, accounting for about 5% of gastric tumours. Until recently, the aetiology of gastric NHL was unknown; it is now generally believed that the main factor in the pathogenesis of this disease is a chronic infection with the bacterium *Helicobacter pylori*. *H. pylori* has been shown to be an important causative factor in the development of mucosa-associated lymphoid tissue (MALT) in the stomach, which, by largely unknown mechanisms, may undergo malignant change.

Various treatment modalities are in use for this disease, varying from chemotherapy, radiotherapy, surgery and, more recently, *H. pylori* eradication, i.e. elimination of the bacterium from the stomach by means of antibiotic drugs, to particular combinations of these therapies. Due to the rare nature of the condition, reports on clinical experience with specific therapeutic regimes usually concern small numbers of patients. When the numbers of patients are larger, studies have been carried out over a long period of time, during which variation in diagnostic workup and treatment occurs. Furthermore, most studies are retrospective in nature, without a predefined treatment regime, thus precluding a comparison between various therapeutic strategies. As a result, the prognostic impacts of particular patient features are still far from clear, even when only considering past experience at a single institution.

Several researchers have recognised the need for decision support in the clinical management of patients with NHL, NHL of the stomach included. Therapy selection for gastric NHL is a complicated process, because only part of the patient findings necessary for therapy selection may be known at a particular stage of the disease, and knowledge of adverse reactions to particular treatments in patient groups may influence treatment selection significantly. Moreover, knowledge of outcome-specific clinical profiles, such as the typical clinical picture of a patient with microscopic evidence of tumour cell elimination after treatment, may help in clinical management and in the choice of appropriate therapy. Much work has been done in the identification of both pretreatment and treatment prognostic factors that help identifying patients at risk. For example, the histological classification of NHL of the stomach in low-grade versus high-grade malignancy using the MALT concept has been shown to be such a prognostic factor. Recently, a number of centres have developed a prognostic model for aggressive nodal NHL, called the International Prognostic Index (IPI), which includes five clinical features for predicting prognosis in patients. This prognostic model has also been applied to NHL patients in general. However, it is unclear whether it is sufficient to include only five features in such a prognostic model. Moreover, which techniques is most suitable for constructing a prognostic model of gastric NHL is also unclear at the moment.

2.3 Relevant variables

The information used in the clinical management of primary gastric NHL is subdivided in pretreatment information, i.e. information that is required for treatment selection, treatment information, i.e. the various treatment alternatives, and posttreatment information, i.e. side effects, and early and long-term treatment results for the disease. The variables are presented in Table 1. It is said that a patient has 'bulky disease' if tumour size exceeds 10 cm in maximal diameter as observed endoscopically, or if there is invasive growth into surrounding tissues or organs. The variable 'clinical stage' is according to the Ann Arbor classification with Musshoff's modification for NHL, with as possible values I, II_1, II_2, III, and IV. This

Pretreatment Variables	**Treatment Variables**	**Posttreatment Variables**
age	chemotherapy (CT)	5-year result
bulky disease	radiotherapy (RT)	
clinical stage	surgery	
clinical presentation	combination therapy (CT-next-RT)	
general health status		
histological classification		

Table 1: Selected variables (abbreviated names).

variable expresses severity of the disease, where the prognosis for patients in stage I is usually favourable and for patients in stage IV is rather grim. The variable 'clinical presentation' summarises the presence of particular gastric complications due to NHL at the time of presentation. Possible values are: 'hemorrhage', 'perforation', 'obstruction', and a variety of non-acute symptoms and signs, such as dyspepsia and weight loss, referred to by the value 'none'. Hemorrhage must be acute and be of sufficient significance to warrant blood transfusion. The variable 'general health status' represents the general health condition of the patient; possible values are: 'good' (WHO index 0 and 1), 'average' (WHO index 2) and 'poor' (WHO index 3 and 4). Histological classification is assumed to be based on the recent MALT concept of gastrointestinal NHL, with subdivision into low-grade and high-grade malignancy.

The variables 'chemotherapy', 'radiotherapy', and 'combination therapy' were combined into one variable with name 'CT&RT-SCHEDULE' (abbreviated to CR) with possible values: chemotherapy (CT), radiotherapy (RT), chemotherapy followed by iceberg radiotherapy (CT-next-RT), and neither chemotherapy nor radiotherapy (none). Possible values for the variable 'surgery' are: 'curative', 'palliative' or 'none', where curative surgery means total or partial resection of the stomach with the complete removal of locoregional tumour mass.

The variable '5-year result' represents the patient either or not surviving five years following treatment.