Chapter IX
Generating and Verifying Risk Prediction Models Using Data Mining

Darryl N. Davis
University of Hull, UK

Thuy T.T. Nguyen
University of Hull, UK

ABSTRACT

Risk prediction models are of great interest to clinicians. They offer an explicit and repeatable means to aide the selection, from a general medical population, those patients that require a referral to medical consultants and specialists. In many medical domains, including cardiovascular medicine, no gold standard exists for selecting referral patients. Where evidential selection is required using patient data, heuristics backed up by poorly adapted more general risk prediction models are pressed into action, with less than perfect results. In this study, existing clinical risk prediction models are examined and matched to the patient data to which they may be applied using classification and data mining techniques, such as neural nets. Novel risk prediction models are derived using unsupervised cluster analysis algorithms. All existing and derived models are verified as to their usefulness in medical decision support on the basis of their effectiveness on patient data from two UK sites.

INTRODUCTION

Risk prediction models are of great interest to clinicians. They offer the means to aide the selection of those patients that need referral, to medical consultants and specialists, from a general medical population. In many medical domains, including cardiovascular medicine, no gold standard exists for selecting

referral patients. Existing practice relies on clinical heuristics backed up by poorly adapted generic risk prediction models. In this study existing clinical risk prediction models are examined and matched to the patient data to which they may be applied using classification and data mining techniques, such as neural nets. The evidence from earlier research suggests that there are benefits to be gained in the utilization of neural nets for medical diagnosis (Janet, 1997; Lisboa, 2002).

In this chapter, the cardiovascular domain is used as an exemplar. The problems associated with identifying high risk patients, (i.e. patients at risk of a stroke, cardiac arrest or similar life threatening event), are symptomatic of other clinical domains where no gold standard exists for such purposes. In routine clinical practice, where domain specific clinical experts are unavailable to all patients, the patient's clinical record is used to identify which patient's are most likely to benefit from referral to a consulting clinician. The clinical record typically, although not always, contains generic patient data (for example age, gender etc.), a patient history of events related to the disease (for example, past strokes, cardiovascular related medical operations), and a profile of measurements, and observations from medical examinations, that characterize the nature of the patient's cardiovascular system. The general practitioner may use a risk prediction model, together with observations from medical examinations, as an aid in determining whether to refer the patient to a consultant. Currently, any such risk prediction model will be based on a general clinical risk prediction system, such as APACHE (Rowan et al., 1994) or POSSUM (Copeland, 2002; Yii & Ng, 2002), which generate a score for patients. Clinicians expert in the disease may well use further risk prediction models, based on their own research and expertise. Such risk prediction models are described in more detail in the second section. The strengths and flaws of the available models for the current clinical domain are explored in the third section, where they are used in conjunction with supervised neural nets. It should be noted, that although this chapter predominantly reports on the use of supervised neural nets and unsupervised clustering in predicting risk in patients, a wide range of other techniques, including decision trees, logistic regression, Bayesian classifiers (Bishop, 2006; Witten & Eibe, 2005) have been tried. The results from applying these other techniques are not given, but typically are similar to or worse than the results presented here. The fourth and fifth sections present an alternative to the coercion of outcome labels, arising from current risk prediction models, with the use of unsupervised clustering techniques. The results from these sections are discussed in the sixth section. The problems associated with making available to clinicians, risk prediction models that arise from the application of data mining techniques, are discussed in that and the concluding section.

RISK PREDICTION MODELS

In this section, two forms of risk prediction model, as used in routine clinical practice, are introduced. The first, POSSUM, typifies the application of generic models to specific medical disciplines. The second set reflect the clinical heuristics regularly used in medicine. The data used throughout this case study is from two UK clinical sites. The attributes are a mixture of real number, integer, Boolean and categorical values. The data records typically contain many default and missing values. For both sites there is typically too high a data value space (i.e. the space of all possible values for all attributes in the raw data) for the data volume (i.e. the number of records) to perform naïve data mining, and some form of data preprocessing is required before using any classifier if meaningful results are to be obtained.