Chapter 13
Learning Cost-Sensitive Decision Trees to Support Medical Diagnosis

Alberto Freitas
CINTESIS – Center for Research in Health Information Systems and Technologies, Portugal
University of Porto, Portugal

Altamiro Costa-Pereira
CINTESIS – Center for Research in Health Information Systems and Technologies, Portugal
University of Porto, Portugal

Pavel Brazdil
LIAAD INESC Porto L.A. – Laboratory of Artificial Intelligence and Decision Support, Portugal
University of Porto, Portugal

ABSTRACT
Classification plays an important role in medicine, especially for medical diagnosis. Real-world medical applications often require classifiers that minimize the total cost, including costs for wrong diagnosis (misclassifications costs) and diagnostic test costs (attribute costs). There are indeed many reasons for considering costs in medicine, as diagnostic tests are not free and health budgets are limited. In this chapter, the authors have defined strategies for cost-sensitive learning. They have developed an algorithm for decision tree induction that considers various types of costs, including test costs, delayed costs and costs associated with risk. Then they have applied their strategy to train and to evaluate cost-sensitive decision trees in medical data. Generated trees can be tested following some strategies, including group costs, common costs, and individual costs. Using the factor of “risk” it is possible to penalize invasive or delayed tests and obtain patient-friendly decision trees.

INTRODUCTION
In medical care, as in other areas, knowledge is crucial for decision making support, biomedical research and health management (Cios, 2001). Data mining and machine learning can help in the process of knowledge discovery. Data mining is the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable
patterns in data (Fayyad et al., 1996). Machine learning is concerned with the development of techniques which allow computers to “learn” (Tom Mitchell, 1997).

Classification methods can be used to generate models that describe classes or predict future data trends. It generic aim is to build models that allow predicting the value of one categorical variable from the known values of other variables. Classification is a common, pragmatic method in clinical medicine. It is the basis for determining a diagnosis and, therefore, for the definition of distinct strategies of therapy. In addition, classification plays an important role in evidence-based medicine. Machine learning systems can be used to enhance the knowledge bases used by expert systems as they can produce a systematic description of clinical features that uniquely characterize clinical conditions. This knowledge can be expressed in the form of simple rule or decision trees (Coiera, 2003).

A large number of methods have been developed in machine learning and in statistics for predictive modelling, including classification. It is possible to find, for instance, algorithms using Bayesian methods (naïve Bayes, Bayesian networks), inductive decision trees (C4.5, C5, CART), rule learners (Ripper, PART, decision tables, Prism), hiperplanes approaches (support vector machines, logistic regression, perceiveptron, Winnow), and lazy learning methods (IB1, IBk, lazy Bayesian networks, KStar) (Witten and Frank, 2005). Besides these base learner algorithms there are also algorithms (meta-learners) that allow the combination of base algorithms in several ways, using for instance bagging, boosting and stacking. There are a few examples that consider costs, using these techniques.

In fact, the majority of existing classification methods was designed to minimize the number of errors. Nevertheless, real-world applications often require classifiers that minimize the total cost, including misclassifications costs (each error has an associated cost) and diagnostic test costs representing the costs of obtaining the value of given attributes. In medicine a false negative prediction, for instance failing to detect a disease, can have fatal consequences, while a false positive prediction can be, in many situations, less serious (e.g. giving a drug to a patient that does not have a certain disease). Each diagnostic test has also a cost and so, to decide whether it is worthwhile to pay the costs of tests, it is necessary to consider both misclassification and tests costs. There are many reasons for considering costs in medicine. Diagnostic tests, as other health interventions, are not free and budgets are limited.

Misclassification and test costs are the most important costs, but there are also other types of costs (Turney, 2000). Cost-sensitive learning (also known as cost-sensitive classification) is the area of machine learning that deals with costs in inductive learning.

The process of knowledge discovery in medicine can be organized into six phases (Shearer, 2000), namely the perception of the medical domain (business understanding), data understanding, data preparation, application of data mining algorithms (modeling), evaluation, and the use of the discovered knowledge (deployment). The data preparation (selection, pre-processing) is normally the most time consuming step of this process (Feelders et al., 2000). The work presented in this chapter is mostly related with the fourth phase, the application of data mining algorithms, particularly classification.

With this chapter we aim to enhance the understand of cost-sensitive learning problems in medicine and present a strategy for learning and testing cost-sensitive decision trees, while considering several types of costs associated with problems in medicine.

The rest of this chapter is organized as follows. In the next section we discuss the main types of costs. Then we review related work. After that, we discuss the evaluation of classifiers. Next, we explain our cost-sensitive decision tree strategy and, subsequently, we present some experimental