Decision Tree Induction

Roberta Siciliano
University of Naples, Federico II, Italy

Claudio Conversano
University of Cagliari, Italy

INTRODUCTION

Decision Tree Induction (DTI) is a tool to induce a classification or regression model from (usually large) datasets characterized by n objects (records), each one containing a set **x** of numerical or nominal attributes, and a special feature *y* designed as its outcome. Statisticians use the terms "predictors" to identify attributes and "response variable" for the outcome. DTI builds a model that summarizes the underlying relationships between **x** and *y*. Actually, two kinds of model can be estimated using decision trees: *classification trees* if *y* is nominal, and *regression trees* if *y* is numerical. Hereinafter we refer to classification trees to show the main features of DTI. For a detailed insight into the characteristics of regression trees see Hastie et al. (2001).

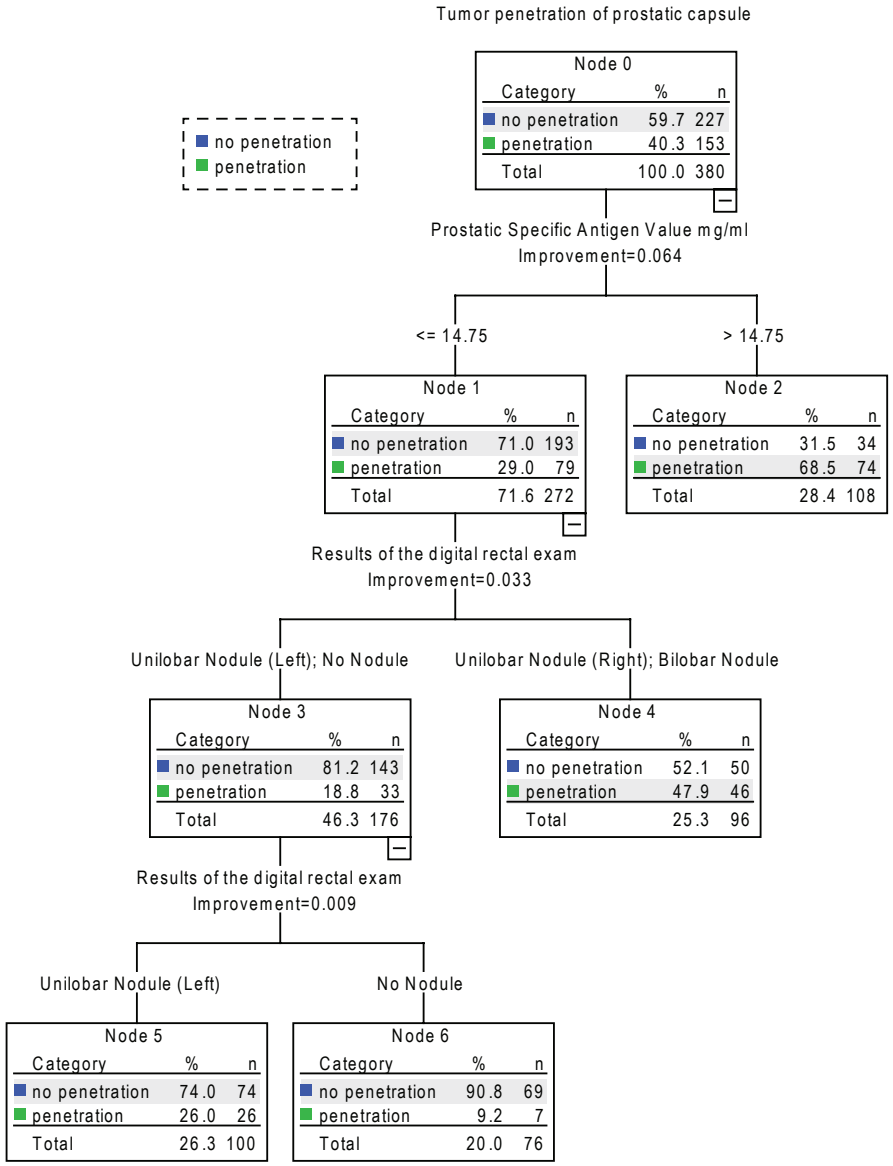
As an example of classification tree, let us consider a sample of patients with prostate cancer on which data such as those summarized in Figure 1 have been collected. Suppose a new patient is observed and we want to determine if the tumor has penetrated the prostatic capsule on the basis of the other available information. Posing a series of questions about the characteristic of the patient can help to predict the tumor's penetration. DTI proceeds in such a way, inducing a series of follow-up (usually binary) questions about the attributes of an unknown instance until a conclusion about what is its most likely class label is reached. Questions and their alternative answers can be represented hierarchically in the form of a decision tree, such as the one depicted in Figure 2.

The decision tree contains a root node and some internal and terminal nodes. The root node and the internal ones are used to partition instances of the dataset into smaller subsets of relatively homogeneous classes. To classify a previously unlabelled instance, say $i^* \left(i^* = 1, \ldots, n\right)$, we start from the test condition in

Figure 1. The prostate cancer dataset

Age in years	Result of the digital rectal exam	Result of he detection of capsular involvment in rectal exam	Prostatic specific antigen value mg/ml	Tumor penetration of prostatic capsule
65	Unilobar Nodule (Left)	No	1.400	*no penetration*
70	No Nodule	Yes	4.900	*no penetration*
71	Unilobar Nodule (Right)	Yes	3.300	*penetration*
68	Bilobar Nodule	Yes	31.900	*no penetration*
69	No Nodule	No	3.900	*no penetration*
68	No Nodule	Yes	13.000	*no penetration*
68	Bilobar Nodule	Yes	4.000	*penetration*
72	Unilobar Nodule (Left)	Yes	21.200	*penetration*
72	Bilobar Nodule	Yes	22.700	*penetration*
…	……..	…..	…….	…….....

Figure 2. An illustrative example of a decision tree for the prostate cancer classification



the root node and follow the appropriate pattern based on the outcome of the test. When an internal node is reached a new test condition is applied, and so on down to a terminal node. Encountering a terminal node, the modal class of the instances of that node is the class label of i^*. Going back to the prostate cancer classification problem, a new subject presenting a prostatic specific antigen value lower than 4.75, and an unilobar nodule on the left side will be classified as "no penetration". It is evident that decision trees can easily be converted into IF-THEN rules and used for decision making purposes.

BACKGROUND

DTI is useful for data mining applications because of the possibility to represent functions of numerical and