Int J Public Health (2012) 57:243–246
DOI 10.1007/s00038-011-0315-z

HINTS & KINKS

Classification and regression trees

N. Speybroeck

Received: 9 February 2011 / Revised: 2 September 2011 / Accepted: 8 October 2011 / Published online: 21 October 2011

Keywords Classification trees · Regression trees · Model · Health

Public health data are often complex, unbalanced and contain missing values. The relationships between a health outcome and its determinants may not be linear and necessitate higher order interactions.

Classification and Regression Trees (CaRTs) are analytical tools that can be used to explore such relationships. They can be used to analyze either categorical (resulting in classification trees) or continuous health outcomes (resulting in regression trees). Figure 1a shows an illustrative example, of a classification tree (CT) result, using a binary health outcome, e.g. being diseased or healthy. The building of a CT begins with a parent node, containing all individuals in a data set, which is then split at a determined value along a range of values for a variable thus producing two child nodes with greater homogeneity (purity) than the parent node.

In Fig. 1a, the most homogeneous grouping is reached by first splitting the individuals—denoted by dots—into two groups depending on whether they have a number of years of education less than some specified level c1. The program iterates on all possible values of the independent variables to identify the first splitter as well as the cut-point for the split, c1, aiming, as far as possible, to result in one group with only diseased individuals and another group with healthy individuals.

The group of individuals below c1 still contain a mixture of healthy and diseased individuals. Child nodes are recursively treated as parent nodes, thereby continuously splitting the data trying to further sift out diseased from healthy individuals. In our example, the group of individuals with years of education less than c1 is split again according to income, at the level c2. The individuals to the left of the c2 line, and below c1, those with relatively low levels of literacy and income, also have the highest percentage of subjects being sick. Most often it is not possible to reach the ideal result in Fig. 1a, but the program aims to create groups which are as homogeneous as possible.

For a binary explanatory variable, only one split is possible, each level defining a group. For categorical variables with k levels, there are $2^{k-1}-1$ possible splits. For numeric explanatory variables, a split is defined by values less than, some value, and for h unique values there are $h-1$ possible splits. From all possible splits of all explanatory variables, called splitters; the program selects the one that maximizes the homogeneity (purity) of the two resulting groups.

The final result resembles an inverted tree (see Fig. 1b), structured as a sequence of simple yes/no questions. The tree building process results in a saturated tree with one case or only cases of one type in each terminal node (i.e., a nodes which is not split further) probably greatly overfitting the information contained within the data set. However, as with stepwise linear regression procedures, adding variables will continuously increase the fit of the model to the data, but at the cost of increasing the true fit to an independent data set. Breiman et al. (1993) showed that an optimal tree can be obtained through a process of using a learning data set (e.g. one part of the data) to prune the saturated tree and select among the so obtained sequence of nested trees, an optimal tree with an appropriate fit to a

N. Speybroeck (✉)
Institute of Health and Society, Université Catholique
de Louvain, Brussels, Belgium
e-mail: Niko.Speybroeck@uclouvain.be

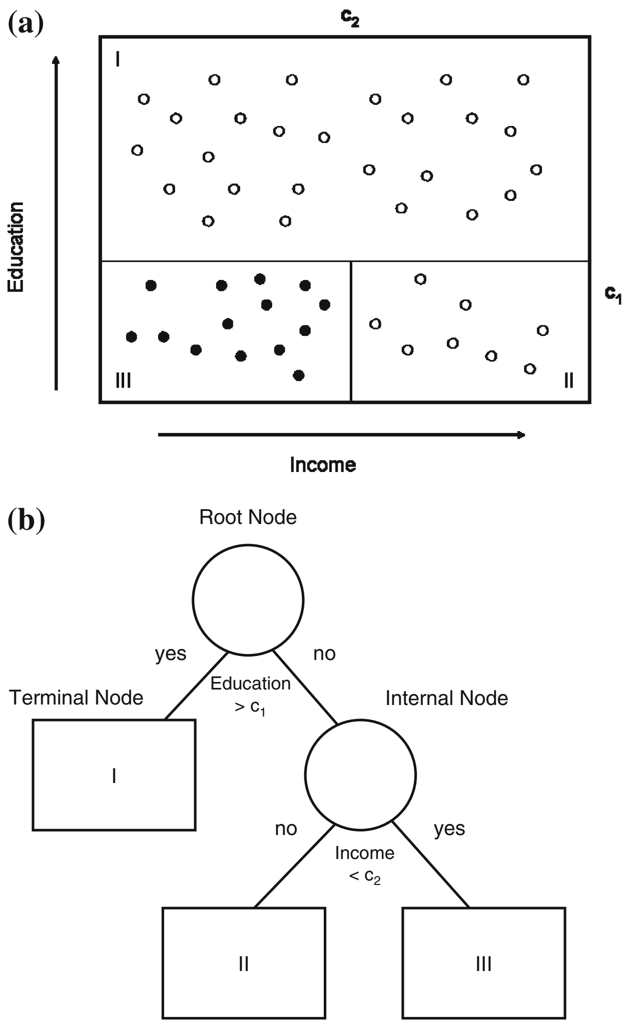

Fig. 1 **a** Illustrative classification according to individuals with bad health (represented by *filled circles*) and individuals with good health (represented by *empty circles*). **b** Corresponding classification tree, indicating how a root node, containing all individuals is divided into two groups to obtain the best separation between individuals with and without good health. The resulting subsets of cases consist of three terminal nodes and one internal node

learning data set (e.g. the part of the data not used to construct the tree). This process helps in retaining only a simple tree, guaranteeing robustness.

CaRT can be valuable for analyzing complex data, providing an informative and original way to show results namely in the form of decision trees—a significant departure from the more traditional statistical procedures, in which linear combinations are the primary method of expressing relationships between variables. Complex interactions become interpretable even by non-statisticians. In Fig. 1 for example, wealth influences health but especially in the less educated part of the population.

CaRT also has the ability to handle the missing values in both response and explanatory variables (Speybroeck et al. 2004). Missing data are common in public health studies, and for many types of models they can be problematic. CTs can treat missing responses as a special category, thereby providing information on response bias. Explanatory variable with missing values can be mimicked by explanatory variables with non-missing values for these cases. Such variables are known as surrogate variables.

CaRT can deal with a wide range of response variables. For continuous data for example, the aim will be to minimize the sums of squares within nodes. Such a regression tree (RT) was used to indicate that in Slovakian district, unemployment rate was the strongest predictor of alcohol-related mortality among males (Rosicova et al. 2011). Count data e.g. disease vector counts, can be dominated by zeros, and the splits may have to be based on explicit statistical models, e.g. a Poisson model. One of the strengths of a RT in analyzing such ecological data is the ability to deal with multicollinearity problems. From two closely related climate variables e.g. rainfall and humidity, a RT will select only one variable as the most important (primary) splitter and deal with the non-selected variables through computing importance scores, indicating their role as a surrogate to the primary splits. The important score measures a variable's ability to mimic the selected tree and to play a role as a stand-in for the primary splits.

The aforementioned flexibility of CaRTs, incurs a number of costs. With a strong linear relationship, trees will not perform as well as linear regressions. CaRT is a rather blunt instrument, with the subdivision of data into two groups being based on only one value of only one explanatory variable. The identification of distinct subgroups does not allow the estimation of net effects of independent variables, and obtaining probability levels or confidence intervals is not straightforward. CART can thus not substitute but complement the different types of regressions. Protopopoff et al. (2009) recommended to use CaRT with an a priori consideration of the independent variables to include by hypothesizing expected relationships through a conceptual model.

In the Appendix, we show how to conduct a CaRT analysis with an illustrative example from the literature, demonstrating code of one of the available software packages, R. The method is firstly illustrated with the R-command rpart which uses the "classical" Breiman algorithm. Unfortunately, this groundbreaking algorithm shows a number of flaws, which may require the use of more modern algorithm such as the one used in the R-command ctree. Probably the most important flaw concerns a selection bias toward covariates with many possible splits (White and Liu 1994) or missing values (Kim and Loh 2001). New algorithms such as ctree, employ splitting criteria that avoid variable selection bias (Hothorn et al. 2006). In addition, the tree growth with ctree is based on statistical stopping rules, so pruning should not be required.

CaRT models were applied in a large number of contexts, ranging from analyzing malaria (Yewhalaw et al. 2009; Thang et al. 2008) to mad cow disease (Saegerman et al. 2004), as well as categorising diseases according to their impact (Cardoen et al. 2009; Havelaar et al. 2010).

This paper illustrates that CaRT models can handle several types of outcomes, but in a different way than regression models. Interventions, developed from regression model results, are geared toward the average member of the population, without consideration of population subgroups as the primary target. A CaRT analysis owns the ability to efficiently segment populations into meaningful subsets. This allows Public Health professionals to identify segments of populations that are marginalized and to efficiently target and maximize the distribution of public health resources.

Acknowledgments I would like to express my thanks to the Reviewer for the constructive and interesting comments.

Appendix: R code to run the decomposition (Comments in different font)

The R software is free of charge and can be downloaded from http://www.r-project.org. An R package called rpart can handle several types of outcomes and generate classification and regression trees. As an example we will indicate how a CT can be constructed for analyzing the relation between malaria (infected/non-infected) and its determinants in Vietnam (Thang et al. 2008) can be generated through the rpart package. After installing the package rpart into R, the following code (in different font) can be copied and used into R and immediately used after having adapted the variables to the users' needs.

```
library(rpart)
# To grow a tree, use the command
rpart(Malaria ~ Age + Forrest + Education + Income + Bednet + Housetype + Ethnicity + Gender, method = class)
# with Forrest Activity, Education, Income, Bednet use, House structure, Ethnicity and Gender the
# explanatory variables [these variables were used in Thang et al. (2008)]
# method can be e.g. "class" for classification trees, "anova" for regression trees, "poisson" for count data.
# detailed summary of splits
summary(fit)
# prune the tree and select the minimal error tree (i.e., with the smallest cross-validated error)
pfit <- prune(fit, cp = fit$cptable[which.min(fit$cptable[,"xerror"]),"CP"])
# detailed summary of the pruned tree
summary(pfit)
# plot the final tree
plot(pfit)
```

A simplified version of the resulting tree in Thang et al. (2008) is shown in Fig. 2. The tree starts with a root node, containing all the 3023 individuals in the sample, with a malaria prevalence (pr) of 14%. The root node is first split into two subgroups according to the wealth status, with the malaria prevalence in the poorer subgroup being higher (pr = 16%) than in the richer subgroup (pr 9%). The richer subgroup is split again into a subgroup engaged in regular forest activity (pr = 31%) and a group not engaged in regular forest activity (pr = 8%). The latter subgroup was split according to their bednet use, with bednet users showing a lower prevalence (pr = 7%) than non-bednet users (pr = 26%).

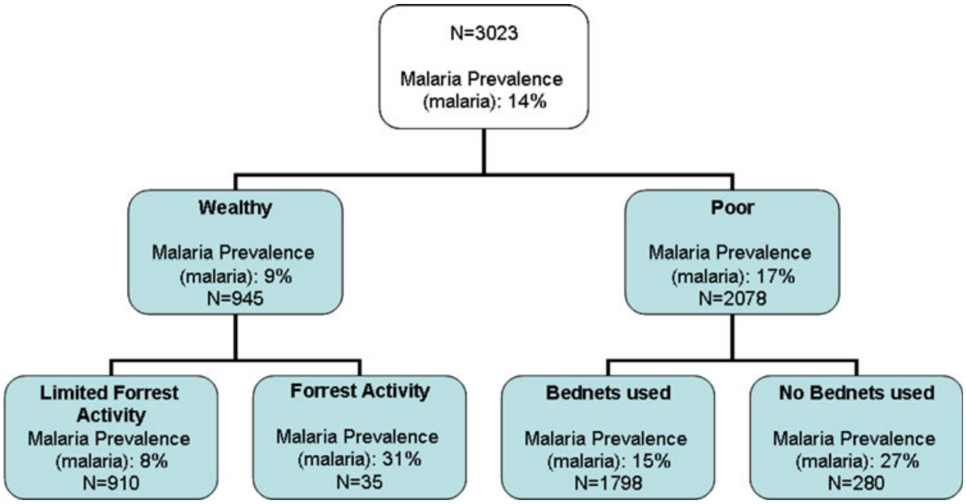
N=3023
Malaria Prevalence (malaria): 14%
Wealthy
Malaria Prevalence (malaria): 9%
N=945
Poor
Malaria Prevalence (malaria): 17%
N=2078
Limited Forrest Activity
Malaria Prevalence (malaria): 8%
N=910
Forrest Activity
Malaria Prevalence (malaria): 31%
N=35
Bednets used
Malaria Prevalence (malaria): 15%
N=1798
No Bednets used
Malaria Prevalence (malaria): 27%
N=280

Fig. 2 Illustrative classification tree for malaria in Vietnam (adapted from Thang et al. 2008)

The example simplified for the sake of brevity (see reference for more information), indicates that CaRTs can be powerful tools for the analysis of complex public health data.

Conditional inference trees can be created via the function ctree (see Hothorn et al. 2006 for additional background)

```
# The party package provides regression trees.
library(party)
# To grow a conditional inference tree, use the command.
ctree(Malaria ~ Age + Forrest + Education + Income + Bednet + Housetype + Ethnicity + Gender)
```

References

Breiman L, Friedman JH, Stone CJ, Olshen RA (1993) Classification and regression trees. Chapman and Hall, New York

Cardoen S, Van Huffel X, Berkvens D, Quoilin S, Ducoffre G, Saegerman C, Speybroeck N, Imberechts H, Herman L, Ducatelle R, Dierick K (2009) Evidence-based semi-quantitative methodology for prioritization of food-borne zoonoses. Food-borne Pathog Dis 6:1083–1096

Havelaar AH, van Rosse F, Bucura C, Toetenel MA, Haagsma JA, Kurowicka D, Heesterbeek JH, Speybroeck N, Langelaar MF, van der Giessen JW, Cooke RM, Braks MA (2010) Prioritizing emerging zoonoses in the Netherlands. PLoS One 5:e13965

Hothorn T, Hornik K, Zeileis A (2006) Unbiased recursive partitioning: a conditional inference framework. J Comput Graph Stat 15:651–674

Kim H, Loh W (2001) Classification trees with unbiased multiway splits. J Am Stat Assoc 96:589–604

Protopopoff N, Van Bortel W, Speybroeck N, D'Alessandro U, Coosemans M (2009) Ranking malaria risk factors to guide malaria control efforts in African Highlands. PLoS One 25:e8022

Rosicova K, Geckova AM, Rosic M, Speybroeck N, Groothoff JW, van Dijk JP (2011) Socioeconomic factors, ethnicity and alcohol-related mortality in regions in Slovakia. What might a tree analysis add to our understanding? Health Place 17:701–709

Saegerman C, Speybroeck N, Roels S, Vanopdenbosch E, Thiry E, Berkvens D (2004) Decision support tools in clinical diagnosis in cows with suspected bovine spongiform encephalopathy. J Clin Microbiol 42:172–178

Speybroeck N, Berkvens D, Mfoukou-Ntsakala A, Aerts M, Hens N, Van Huylenbroeck G, Thys E (2004) Classification trees versus multinomial models in the analysis of urban farming systems in Central Africa. Agric Syst 80:133–149

Thang ND, Erhart A, Speybroeck N, Hung LX, Thuan LK, Hung TK, Van Ky P, Coosemans M, D'Alessandro U (2008) Malaria in Central Vietnam: analysis of risk factors by multivariate analysis and classification tree models. Malar J 7:28

White A, Liu W (1994) Bias in information based measures in decision tree induction. Mach Learn 15:321–329

Yewhalaw D, Legesse W, Van Bortel W, Gebre-Selassie S, Kloos H, Duchateau L, Speybroeck N (2009) Malaria and water resource development: the case of Gilgel-Gibe hydroelectric dam in Ethiopia. Malar J 8:21