Journal of Biomedical Informatics 43 (2010) 701–708

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

PRIM versus CART in subgroup discovery: When patience is harmful

Ameen Abu-Hanna [a,*], Barry Nannings [a], Dave Dongelmans [b], Arie Hasman [a]

[a] Department of Medical Informatics, Academic Medical Center, University of Amsterdam, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands
[b] Department of Intensive Care, Academic Medical Center, University of Amsterdam, Meibergdreef 9, 1105 AZ Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 13 August 2009
Available online 28 May 2010

Keywords:
CART (Classification and Regression Trees)
PRIM (Patient Rule Induction Method)
Subgroup discovery
Coverage
Patience
High-dimensionality
Clinical databases
Ordinal scores
Bootstrap

ABSTRACT

We systematically compare the established algorithms CART (Classification and Regression Trees) and PRIM (Patient Rule Induction Method) in a subgroup discovery task on a large real-world high-dimensional clinical database.

Contrary to current conjectures, PRIM's performance was generally inferior to CART's. PRIM often considered "peeling of" a large chunk of data at a value of a relevant discrete ordinal variable unattractive, ultimately missing an important subgroup. This finding has considerable significance in clinical medicine where ordinal scores are ubiquitous. PRIM's utility in clinical databases would increase when global information about (ordinal) variables is better put to use and when the search algorithm keeps track of alternative solutions.

1. Introduction

Many data-analytic problems in biomedical research necessitate finding a function $f(y|\mathbf{x})$ that approximates the value of an output variable y, with some unknown probability density $p(y|\mathbf{x})$, for any value of $\mathbf{x}$ in input space. For example, one may want to predict the probability of survival of a patient based on patient and treatment variables. Various models, such as logistic regression and regression trees, and associated procedures have been described in the literature to induce such functions. Often, however, the interest is not in the approximating function itself but in minima or maxima of y. Instead of seeking a global model to predict the output variable for any subject in the population, one may be interested in regions in input space with a very high (or low) value of y. For example, one might want to identify a subgroup of patients who do not respond well to therapy, or a subgroup of genes that exhibit markedly different expression patterns. To identify these regions and/or the maximum or minimum values of y in these regions one can first induce $f(y|\mathbf{x})$ and then optimize this function. An alternative approach to determine such regions bypasses finding an approximating function (which may be a formidable task itself) and directly seeks these regions. A well-established representative of this latter approach is PRIM (Patient Rule Induction Method), which has been gaining more ground [1–13] since its introduction in [14]. PRIM is a patient bump-hunting (or subgroup discovery) algorithm. PRIM initially starts with all given data and iteratively discards observations of seemingly unpromising regions. In this manner it gradually zooms into regions with high values of y (bumps). In contrast to greedy or semi-greedy algorithms, PRIM is patient in the sense that in its heuristic search it attempts at each step to exclude only a small portion of the data. This is an attempt to guard against hasty initial decisions. By keeping enough observations for subsequent decisions, initial suboptimal choices may be recuperated from.

It is only natural to compare PRIM to approaches that induce an approximating function first, such as CART. Because CART and PRIM share the same symbolic IF–THEN representation it is important to compare their performances and understand their strengths and limitations. Indeed, in [14], where PRIM was introduced, a provisional comparison with CART was also provided in two domains: geology and marketing. From this comparison it appeared that PRIM performed better than CART in subgroup discovery tasks. This superior performance was attributed to PRIM's patience. No other studies were dedicated to comparing them on real world datasets. We are only aware of work that compared the two algorithms in the field of scenario discovery (for supporting decision analysis) on simulated data [15]. Both algorithms were found to perform the required task. Other subsequent publications on PRIM, and indeed the papers discussing the original paper of Friedman and Fisher, which appeared in the same issue, often referred to this evidence of superiority of PRIM over CART.

The objective of this paper is to systematically compare PRIM with CART on a large clinical database and inspect whether there

* Corresponding author. Fax: +31 20 6919840.
E-mail address: a.abu-hanna@amc.uva.nl (A. Abu-Hanna).

1532-0464/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2010.05.009

are circumstances common to real-world clinical databases in which PRIM is less effective than CART in a subgroup discovery task. Our findings show that PRIM's performance was often inferior to CART because it failed to find a relatively large contiguous subgroup. We show that this happens when a "peel" attempts to remove a very large group of observations based on a discrete ordinal variable. Due to PRIM's patience this peel seems unattractive but in fact it can yield better subgroups.

2. Materials and methods

2.1. PRIM and CART

CART [16] has been extensively described and investigated in the literature; tree induction has indeed become a mainstream topic in machine learning. PRIM has been well described in [14] but it is less likely to be known to researchers than CART. Our intention here is to provide an intuitive explanation and illustration of the subgroup discovery problem and the procedure that PRIM follows.

2.2. Patient Rule Induction Method

The optimization problem can be stated as follows. A sample is given of N observations $\{y_i, \mathbf{x}_i\}_1^N$ from some joint distribution with unknown probability density $p(y|\mathbf{x})$ where y denotes the output variable and $\mathbf{x}$ a vector consisting of p input variables, $\mathbf{x} = (x_1, x_2, \ldots, x_p)$. We seek a region B (called a box) in input space, in which the mean of the output variable, denoted as $\bar{f}_B$, is much larger than the population's mean $\bar{f}$. Note that when y is binary then the mean of y is equivalent to the probability of y, that is, $f(\mathbf{x}) = \text{Prob}(y = 1|\mathbf{x})$. A box is described by a conjunction of conditions on input variables. For real and discrete ordinal input variables the conditions relate to contiguous intervals. An example of a box's description is "80 < blood-pressure < 120 $\wedge$ reason-for-admission $\in$ {elective-surgery, planned-surgery}".

PRIM can return a set of boxes (this whole set is called a rule in PRIM): once a box is found its observations are removed and the search for a new box is started again. Definitions of subsequent boxes may overlap with earlier discovered ones. The probability estimates of y in a box are calculated – in the training and test sets – only after the observations of earlier boxes are removed. For example two boxes B_1 and B_2, discovered in this order, correspond to the observations in the sets $\{\mathbf{x}|\mathbf{x} \in B_1\}$ and $\{\mathbf{x}|\mathbf{x} \notin B_1 \wedge \mathbf{x} \in B_2\}$, respectively.

An important property of a box B is its "support" $\beta_B = \int_{\mathbf{x}\in B} p(\mathbf{x})dx$ which is estimated as the number of observations in B. One prefers high support subgroups with high $\bar{f}_B$, but higher support usually causes lower $\bar{f}_B$, hence one should strike a balance between support and target mean ("target" refers to the output variable).

To find these boxes PRIM applies a search procedure, which is first explained for continuous variables. PRIM includes the entire sample in an initial box, which is a rectangle in two dimensions and a hypercube in general. It then considers each face of the hypercube for shrinking by considering removing a user-specified percentage (α) of the observations for the variable at that face. It selects the "peel" that results in the remaining box having the maximum mean of the target. That is, at each step it considers two options for a variable: removing the data below the α quantile or above the $1 - \alpha$ quantile of the variable's distribution in the current box. Peeling follows essentially a hill-climbing search strategy in which each variable is considered in isolation. This peeling process continues by removing the proportion α of the remaining observations until a user-specified minimum proportion (β_0) of the initial sample is reached in the box. The meta-parameters α (peeling fraction) and β_0 (support) control the induction process.

Fig. 1(a) illustrates the steps taken by PRIM to discover a subgroup with a high density of some outcome, such as mortality, in a two-dimensional space for continuous variables. The variable x_1 may denote for example the maximal creatinine value in μmol/l and x_2 the urine production in liters, both within the first 24 h of admission to an Intensive Care Unit (ICU). In the first step, removing the proportion α of observations with the highest values of variable x_2 yield the box with the greatest target mean. In the second step removing the proportion α of the remaining observations with the highest value of variable x_1 yield the successive box with the greatest target mean. The final subgroup was obtained after 10 steps and is shown as a gray rectangle. Such rectangle may for example define observations with "$120 < x_1 < 650 \wedge 0.5 < x_2 < 1.5$". The number of observations in the subgroup should be at least β_0 of the total sample. Fig. 1(b) shows how CART recursively partitions the two-dimensional space. The first split (the line marked by '1') is defined in terms of x_1. The space at the left and right of this line is further split based on x_2 by the lines marked by '1.1' and '1.2', respectively. The final partition includes 9 subgroups, two of which with a sufficiently high outcome mean (e.g. defined as twice the global mean).

For discrete ordinal variables the PRIM algorithm has no absolute control on the number of removed observations as all observations with identical values are considered together. The number of peeled observations is chosen as the one closest to (but may exceed) α observations. For a categorical variable, PRIM inspects the removal of observations belonging to each one of the possible categories separately. For example, if the reason-for-admission variable in the current box has the domain {elective-surgery, planned-surgery, emergency} then only the sub-boxes corresponding to {planned-surgery, emergency}, {elective-surgery, emergency}, and

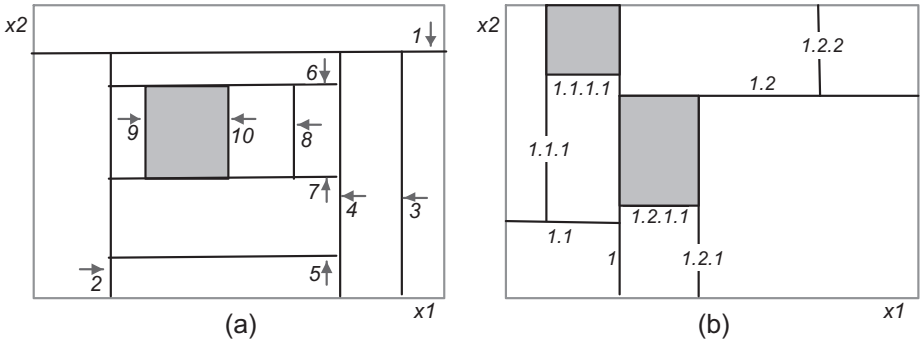

Fig. 1. (a) An illustration of how PRIM finds a first subgroup in 10 steps in a two-dimensional space, and (b) an illustration of how CART finds a partition, here shown including two subgroups with high mean outcome.