COMSTA: 6177 | Model 3Gsc | pp. 1–12 (col. fig: NIL)

ARTICLE IN PRESS

Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

HHCART: An oblique decision tree

D.C. Wickramarachchi [a,*], B.L. Robertson [b], M. Reale [b], C.J. Price [b], J. Brown [b]

[a] Department of Statistics, Faculty of Applied Sciences, University of Sri Jayewardenepura, Gangodawila, Nugegoda, Sri Lanka

[b] School of Mathematics and Statistics, University of Canterbury, Private Bag 4800, Christchurch, 8042, New Zealand

ARTICLE INFO

Article history:
Received 28 August 2014
Received in revised form 9 November 2015
Accepted 9 November 2015
Available online xxxx

Keywords:
Oblique decision tree
Data classification
Statistical learning
Householder reflection
Machine learning

ABSTRACT

Decision trees are a popular technique in statistical data classification. They recursively partition the feature space into disjoint sub-regions until each sub-region becomes homogeneous with respect to a particular class. The basic Classification and Regression Tree (CART) algorithm partitions the feature space using axis parallel splits. When the true decision boundaries are not aligned with the feature axes, this approach can produce a complicated boundary structure. Oblique decision trees use oblique decision boundaries to potentially simplify the boundary structure. The major limitation of this approach is that the tree induction algorithm is computationally expensive. Hence, as an alternative, a new decision tree algorithm called HHCART is presented. The method uses a series of Householder matrices to reflect the training data at each non-terminal node during tree construction. Each reflection is based on the directions of the eigenvectors from each class' covariance matrix. Considering of axis parallel splits in the reflected training data provides an efficient way of finding oblique splits in the unreflected training data. Experimental results show that the accuracy and size of HHCART trees are comparable with some benchmark methods. The appealing feature of HHCART is that it can handle both qualitative and quantitative features in the same oblique split.