Adapting Peepholing to Regression Trees

Luis Torgo[1] and Joana Marques[2]

[1] LIACC-FEP, University of Porto, R. de Ceuta, 118, 6., 4050-190 Porto, Portugal
ltorgo@liacc.up.pt
http://www.liacc.up.pt/~ltorgo
[2] Aveiro Digital
jvcmarques@hotmail.com

Abstract. This paper presents an adaptation of the peepholing method to regression trees. Peepholing was described as a means to overcome the major computational bottleneck of growing classification trees by Catlett [3]. This method involves two major steps: shortlisting and blinkering. The former has the goal of eliminating some continuous variables from consideration when growing the tree, while the second tries to restrict the range of values of the remaining continuous variables that should be considered when searching for the best cut point split. Both are effective means of overcoming the most costly step of growing tree-based models: sorting the values of the continuous variables before selecting their best split. In this work we describe the adaptations that are necessary to use this method within regression trees. The major adaptations involve developing means to obtain biased estimates of the criterion used to select the best split of these models. We present some preliminary experiments that show the effectiveness of our proposal.

1 Introduction

Regression trees [5] handle multivariate regression problems obtaining models that have proven to be quite interpretable and with competitive predictive accuracy. Moreover, these models can be obtained with a computational efficiency that hardly has parallel in competitive approaches, turning these models into a good choice for a large variety of data mining problems where these features play a major role. Nevertheless, the growth rate of databases creates problems even for these efficient methods. As such, techniques for reducing the computational requirements of growing regression trees are of key importance for handling extremely large problems.

Many strategies exist to try to overcome the problems of using a certain modelling technique on very large data sets. These can be classified in two broad categories: techniques that are independent of the modelling approach; and techniques that involve the optimization of model construction. In the former we can distinguish two major approaches: reducing the number of features; and reducing the number of cases. Regards the second category, the approaches are generally model-specific and basically try to address the computational bottlenecks of the respective algorithm for obtaining the models. The work presented in this paper

C. Bento, A. Cardoso, and G. Dias (Eds.): EPIA 2005, LNAI 3808, pp. 293–303, 2005.

is of this type: we try to address the major computational bottleneck of growing regression trees.

One of the seminal works on handling large problems using tree-based models is the work by Catlett [3]. This author has thoroughly addressed the issue of the computational bottlenecks of growing classification trees, and presented several approaches to try to overcome them. One of the key contributions of Catlett's work was the identification of the major bottleneck of the process of growing a tree-based model: the selection of the best test on a continuous variable for any node of a tree. In this context, Catlett has presented a technique, *peepholing*, which specifically addresses the problem of reducing the computational complexity of this task. This technique consists of two major steps: *shortlisting* and *blinkering*. The first tries to eliminate some continuous variables from the set to be considered when searching for the best test of a node. The second tries to reduce the range of values to be considered when searching for the best cut point for a test in one of the variables not eliminated by the shortlisting step. Among these two steps, the former has the larger impact in terms of reducing the computational complexity of growing a tree according to Catlett.

Catlett has focused his work on classification trees. The peepholing method is strongly connected to the criterion used to select the best test of a node. The criteria used in classification trees are quite different from the usual criterion (least squares) used for growing regression trees. The goal of this work is to try to adapt the peepholing method to the growth of least squares regression trees. We identify the key issues that require modification and propose solutions in order to be able to use peepholing with regression trees.

The next section presents a brief overview of the theory behind the methods used to grow least squares regression trees. Section 3 describes the peepholing method presented by Catlett [3]. In Section 4 we describe our adaptation of this method to regression trees. The results of our preliminary experiments with this adaptation are shown in Section 5. Finally, the main conclusions of this work are given in Section 6.

2 Least Squares Regression Trees

Regression trees are usually obtained by using a least squares error criterion (e.g. [2]). A regression tree can be seen as a kind of additive regression model [4] of the form,

$$rt(x) = \sum_{i=1}^{l} k_i \times I(x \in D_i) \tag{1}$$

where $k_i's$ are constants; $I(.)$ is an indicator function returning 1 if its argument is true and 0 otherwise; and $D_i's$ are disjoint partitions of the training data D such that $\bigcup_{i=1}^{l} D_i = D$ and $\bigcap_{i=1}^{l} = \phi$.

These models are sometimes called piecewise constant regression models. Regression trees are constructed using a recursive partitioning (RP) algorithm