Luís Fernando Raínho Alves Torgo

INDUCTIVE LEARNING OF TREE-BASED REGRESSION MODELS

Tese submetida para obtenção
do grau de Doutor em Ciência de Computadores

Departamento de Ciência de Computadores
Faculdade de Ciências da Universidade do Porto
Setembro de 1999

Dedication

Para o meu avô.

Acknowledgements

I would like to thank my supervisor, Pavel Brazdil, for all his support since I started my research on Machine Learning. He has managed to provide me with the ideal mixture of research guidance and freedom to follow my own intuitions. Thanks also for all his useful comments and suggestions that certainly improved this thesis. Finally, I would like also to thank him for all his effort in management tasks that provided me and my colleagues with such excellent working conditions.

Thanks to all my colleagues at LIACC for such pleasant working atmosphere. Thanks in particular to all those in NIAAD (machine learning group) for useful comments and discussions. A special word to my two colleagues with whom I share the office: Alípio Jorge and João Gama. Finally, I would like to thank my colleague Joaquim Pinto da Costa for his support with questions concerning statistics.

I would like to thank the Faculty of Economics of the University of Porto and the Portuguese Government (Programa Ciência and PRAXIS XXI) for all financial support that made this thesis possible. Thanks also are due to the European Community for its financial support to several research projects and networks on which I was involved.

Thanks to all my good friends for all nice moments that allowed me to escape from my research preoccupations. Their friendship made things much easier.

Special thanks to Catarina. Her love and joy are an unlimited source of happiness.

Finally, a very special thanks to my family and in particular my parents. Without their love and unquestionable support nothing of this would ever be possible.

Abstract

This thesis explores different aspects of the induction of tree-based regression models from data. The main goal of this study is to improve the predictive accuracy of regression trees, while retaining as much as possible their comprehensibility and computational efficiency. Our study is divided in three main parts.

In the first part we describe in detail two different methods of growing a regression tree: minimising the mean squared error and minimising the mean absolute deviation. Our study is particularly focussed on the computational efficiency of these tasks. We present several new algorithms that lead to significant computational speed-ups. We also describe an experimental comparison of both methods of growing a regression tree highlighting their different application goals.

Pruning is a standard procedure within tree-based models whose goal is to provide a good compromise for achieving simple and comprehensible models with good predictive accuracy. In the second part of our study we describe a series of new techniques for pruning by selection from a series of alternative pruned trees. We carry out an extensive set of experiments comparing different methods of pruning, which show that our proposed techniques are able to significantly outperform the predictive accuracy of current state of the art pruning algorithms in a large set of regression domains.

In the final part of our study we present a new type of tree-based models that we refer to as *local regression trees*. These hybrid models integrate tree-based regression with local modelling techniques. We describe different types of local regression trees and show that these models are able to significantly outperform standard regression trees in terms of predictive accuracy. Through a large set of experiments we prove the competitiveness of local regression trees when compared to existing regression techniques.

Contents

List of Tables

List of Figures

List of Symbols

X_i is an input (predictor) variable (or attribute).

A is the set of attributes of a problem.

a is the number of attributes (or predictor variables) of a problem, that is $a = \#A$.

χ_i is the domain of the variable X_i .

Y is a goal variable.

Υ is the domain of variable Y.

$\mathbf{x}_i$ is a vector of attribute values having as domain $\chi = \chi_1 \times \ldots \times \chi_a$.

$\langle \mathbf{x}_i, y_i \rangle$ is a case or example with domain $\chi \times \Upsilon$.

$D = \{\langle \mathbf{x}_i, y_i \rangle\}_{i=1}^{n}$ is a data set or sample of cases.

n is the number of training cases of a problem.

$\mathbf{D = X \mid Y}$ is a matrix representation of a data set, where $\mathbf{X}$ is a $n \times a$ matrix containing the n vectors of a attributes values, and $\mathbf{Y}$ is a column vector containing the n goal variable values.

$r(\beta, \mathbf{x})$ is a regression model with parameters β, which maps an attribute vector into a goal variable value, *i.e.* $r : \chi \rightarrow \Upsilon$.

$r(\beta, \mathbf{x}_i)$ is the predicted value of the model $r(\beta, \mathbf{x})$ for a case $\langle \mathbf{x}_i, y_i \rangle$.

T is a tree-based model.

t is a node of a tree T.

T_t is a sub-tree of T rooted at node t.

t_L is the left child node of node t.

t_R is the right child node of node t.

l is a leaf node of a tree.

$\tilde{T}$ is a set containing all leaf nodes of a tree T.

$\#\tilde{T}$ is the number of leaves of a tree, *i.e.* the cardinality of set $\tilde{T}$.

P_i is a path from the root node t_0 to a node t_i. We represent this path by the conjunction of logical tests of the nodes in the path.

D_t is the sub-set of the training sample D falling in node t, that is

$D_t = \{\langle \mathbf{x}_i, y_i \rangle \in D : \mathbf{x} \to P_t\}$.

Chapter 1

Introduction

This dissertation addresses a particular methodology to deal with a data analysis problem usually known as multivariate regression. A *multivariate regression problem* can be loosely defined as the study of the *dependency relation* between a *goal* (*dependent*, *response*) variable and a set of *independent* (*predictor*, *input*) variables. The goal of multivariate regression analysis is to *obtain a model of this relationship*. This model can be used either for *understanding* the interactions between the variables of the domain under study, or to *predict* the value of the goal variable of future instances of the same problem. Regression models are usually obtained through an *inductive process* that uses examples (*the training sample*) of an unknown regression relationship. Consider for instance a collection of observations regarding some financial indicators of thousands of companies and their respective stock values. We could be interested in obtaining a regression model that relates the stock value with some of these indicators. This model could then be used to predict the stock value of a new company for which we know the value of the indicators. Moreover, if our model is somehow intelligible it could reveal some yet unknown relationships in this domain.

The existence of a functional relationship between a target variable and a set of predictor variables is common in real world data analysis (Drapper & Smith, 1981).

Finding a model of such relationship is very important because it allows us to predict and comprehend the behaviour of our data. These and other arguments justify the existence of so many research works carried out about this data analysis problem. Still, the majority of studies about regression analysis concern methods of generating a *multivariate linear model* between the response variable and the predictors. Although being one of the most successful data analysis techniques, linear regression has obvious limitations due to the linear structure that is imposed on the data. To deal with data with more complex structure one needs more sophisticated tools. *Non-parametric approaches to regression* are able to deal with a wide range of regression domains by not imposing any pre-defined global form to the regression surface (Hardle, 1990). These methods assume some functional form only at a *local level*, meaning that they do not try to fit one single model to all given sample. This methodology leads to a regression surface that is globally non-parametric (*i.e.* an approximation that does not have a fixed global functional form). This is an important feature if one wishes to develop a tool that is applicable to a wide range of regression problems. That is one of the goals of the work carried out in this dissertation.

The approach to regression analysis followed in this thesis is usually known as *tree-based regression*. The regression models obtained with this methodology can be represented in the form of a tree. This tree consists of a hierarchy of nodes, starting with a top node known as the *root node*. Each node of the tree contains *logical tests* on the predictor variables, with the exception of the bottom nodes of the hierarchy. These latter are usually known as the leaves of the tree. The leaves contain the predictions of the tree-based model. Each path from the root node to a leaf can be seen as a conjunction of logical tests on the predictor variables. These conjunctions are logical representations of “sub-areas” of the overall regression surface being approximated. For each of these local areas different values of the goal variable can be predicted. Although the form of the logical tests imposes some restrictions on the type of local areas that can be represented, the number of local areas and their “size” is not pre-defined. Because of this regression trees are able to represent a wide range of regression surfaces through a flexible composition of smaller surfaces. The main features of tree-based regression are its *wide range of applicability*, *low computation time* and its *comprehensibility* due to the use of a symbolic representation of the regression surface. These models are studied in several research fields like social sciences, statistics, engineering and artificial intelligence.

Most existing methods of growing a regression tree try to minimise the *mean squared error* of the resulting model. This method has important computational advantages stemming from the mathematics behind this criterion. However, mean squared error is known to be quite sensitive to *outliers* (*i.e.* data points that diverge a lot from the bulk of data). These data items may distort the mean squared error statistic possibly leading to wrong decisions in the process of growing a tree. We see this as an opportunity for considering other growth criteria that do not suffer from this effect.

As we have mentioned, comprehensibility is one of the main advantages of regression trees. However, as these models do not assume any fixed form of the regression surface, they can easily *overfit the given training sample*, by generating too many local areas to approximate the unknown surface. This leads to models that are unnecessarily large with consequences in terms of comprehensibility. Moreover, overfitting may also cause *lower predictive accuracy*, particularly with unreliable data (*i.e. noisy* data). *Pruning* is the usual method of avoiding overfitting in regression trees. Pruning is usually regarded as a search through the space of all possible pruned trees of an overly large tree that overfits the training data. Current approaches to this search problem use methods relying either on *estimates of the true prediction error* of the trees (*e.g.* Breiman *et al.*, 1984) or other criteria like *minimising the binary description length* of the tree model (Robnik-Sikonja & Kononenko, 1998).

The non-parametric features of regression trees usually lead to competitive predictive accuracy in a large set of domains. However, regression trees can be seen as fitting constants in a set of partitions of the input space (the multidimensional space defined by the predictor variables). As we have mentioned, a regression tree is a logic representation of this set of partitions. In each of these regions of the input space regression trees assume that the unknown regression function takes a constant value. This leads to a *highly non-smooth set of local regions*, that all together form the regression surface approximation provided by a regression tree. In this thesis we explore the possibility of introducing a further level of detail within these partitions by *using smother models* within each leaf of the trees.

1.1 Objectives

The main objective of this thesis is to study different aspects of tree-based regression with the aim of improving the accuracy of these models, while trying to retain their main features concerning applicability, processing time and comprehensibility of the models. Our study is focused on deriving trees from large sets of data and thus we are particularly concerned with methods that are sufficiently efficient. We start our study by exploring computationally efficient methods of growing regression trees using two different error criteria. We improve the computational efficiency of the method of growing least squares regression trees, and we consider the alternative of growing trees that minimise the mean absolute deviation. This latter methodology aims at overcoming difficulties with data outliers. We then address the important question of avoiding overfitting of the training data with the objective of providing a better compromise between accuracy and size of the models. Finally, we describe a new type of models, called local regression trees, that permit to attain higher accuracy than existing approaches to tree-based regression, through the use of smoother models in their leaves. Throughout this thesis we have extensively compared our proposals with existing methods. These comparisons were carried out using several data sets and different training sample sizes.

1.2 Main Contributions

This dissertation presents a thorough study of tree-based regression models. The study is divided in three main parts. The first addresses the generation of tree models. We review several existing techniques of growing least squares (LS) and least absolute deviation (LAD) regression trees. We describe new algorithms for inducing LAD trees that increase the efficiency of this task. We present a theoretical study concerning the possibility of extending a theorem described by Breiman *et al.* (1984), to LAD trees. This extension is relevant because the results of the theorem significantly improve the computational efficiency of finding the best discrete split for a tree node. We also describe an experimental comparison between LS and LAD regression trees. This comparison shows that these two methodologies have different application goals. While LS trees try to minimise the square error of the tree resulting in a model with lower probability of committing large prediction errors, LAD trees minimise the average absolute deviation, which leads to models that on average tend to be more accurate but may occasionally commit larger prediction errors.

The second part of this thesis addresses overfitting avoidance within regression trees. We review the major existing pruning techniques and then focus our study on pruning by tree selection. We describe two new methods of exploring the large space of pruned trees that entail accuracy advantages when compared to existing approaches. We present a new method of using Cross Validation estimates, which extends their applicability in the context of pruning by tree selection. We extend the use of *m* estimators (Cestnik, 1990) within regression trees. We describe an *m* estimator of the Mean Absolute Deviation that allows using this method with LAD trees. We also derive an expression of the standard error of *m* estimates, which allows using the 1-SE rule (Breiman *et al.*,1984) during pruning, with large advantages in terms of tree size. Finally, we present a new error estimator (*ChiEst*) for LS trees based on the sampling distribution properties of the mean squared error, that provides very competitive accuracy with low computation costs.

We have carried out an extensive empirical comparison of different methods of pruning by tree selection. We have concluded that selection based on error estimates obtained either with Cross Validation or our *ChiEst* method are the best ways to achieve higher accuracy in a large set of benchmark domains. However, we have also observed that these methods entail some costs in terms of average tree size. With respect to computation time our *ChiEst* method achieved significant advantages when compared to other existing pruning algorithms.

We have also compared our pruning proposals with three state-of-the-art pruning algorithms. We have observed significant accuracy advantages of our methods at the cost of some tree size, particularly when compared to CART (Breiman *et al.*,1984) and CORE (Robnik-Sikonja & Kononenko,1998) pruning algorithms.

Regarding local regression trees we describe three variants of these models: kernel trees; partial linear trees; and local linear trees. We have confirmed our hypothesis concerning the increased predictive accuracy of these models when compared to standard regression trees. However, we have also observed that this advantage comes at the cost of larger computational requirements and loss of some of comprehensibility of the models. We also show that local regression trees can be seen as a means to overcome some of the limitations of the local models we have integrated in their leaves. In effect, the focusing effect provided by local regression trees is largely beneficial in terms of computation time when compared to the use of local models per see. Finally, in an empirical comparison with three of the most well-known existing regression methodologies, we have observed that partial linear trees are among the most competitive in terms of predictive accuracy.

1.3 Organisation of the Thesis

This dissertation is organised as follows. Chapter 2 presents an overview of the existing work on learning by examples within the propositional framework. This description is particularly focused on regression methods. We provide an overview of the major regression techniques used in the research fields of statistics, neural networks and machine learning.

Chapter 3 presents two methods of growing regression trees. The first is the method of growing a tree by minimising the mean squared error of the resulting model. We discuss several computational issues of this methodology and present some algorithms that improve its efficiency. We then address the question of growing regression trees by trying to minimise the mean absolute deviation. We present several algorithms for growing this type of trees and discuss its computational difficulties. We also provide several techniques that allow a large computational speed-up of the task of growing these trees.

Chapter 4 addresses one of the most important aspects of tree-based regression: overfitting avoidance. We follow the method of pruning by selecting from a set of previously generated pruned trees. We describe two new algorithms for generating these pruned trees. We study several alternative ways of evaluating tree models and apply them in the context of pruning by tree selection. We show the results of an empirical comparison between different methods of pruning by tree selection. Finally, we compare our pruning proposals with three of the most successful existing pruning algorithms.

In chapter 5 we describe a new type of tree-based regression models, named local regression trees. These models integrate regression trees with local modelling techniques. The integration is carried out using local models in the leaves of regression trees. We describe three variants of local regression trees: kernel trees; partial linear trees; and local linear trees. We carry out a series of experiments designed with the goal of finding out the advantages and disadvantages of local regression trees.

We also include a series of Annexes that provide further details on the work carried out in the thesis. Annex A describes the material and methods used in the experimental parts of the thesis. This annex also includes details concerning the form of presenting experimental results used throughout the thesis. Annex B presents the full tables of results of several experiments carried out in thesis that were not included in the main text for clarity reasons.