Information Sciences 288 (2014) 153–173

Contents lists available at ScienceDirect

Information Sciences

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

Evolutionary induction of global model trees with specialized operators and memetic extensions

Marcin Czajkowski, Marek Kretowski [*]

Faculty of Computer Science, Bialystok University of Technology, Wiejska 45a, 15-351 Bialystok, Poland

ARTICLE INFO

Article history:
Received 18 July 2013
Received in revised form 11 June 2014
Accepted 30 July 2014
Available online 7 August 2014

Keywords:
Evolutionary algorithm
Model tree
Multiple linear regression
Machine learning

ABSTRACT

Metaheuristics, such as evolutionary algorithms (*EAs*), have been successfully applied to the problem of decision tree induction. Recently, an EA was proposed to evolve model trees, which are a particular type of decision tree that is employed to solve regression problems. However, there is a need to specialize the *EAs* in order to exploit the full potential of evolutionary induction. The main contribution of this paper is a set of solutions and techniques that incorporates knowledge about the inducing problem for the global model tree into the evolutionary search. The objective of this paper is to demonstrate that specialized *EA* can find more accurate and less complex solutions to the traditional greedy-induced counterparts and the straightforward application of *EA*.

This paper proposes a novel solution for each step of the evolutionary process and presents a new specialized *EA* for model tree induction called the Global Model Tree (*GMT*). An empirical investigation shows that trees induced by the *GMT* are one order of magnitude less complex than trees induced by popular greedy algorithms, and they are equivalent in terms of predictive accuracy with output models from straightforward implementations of evolutionary induction and state-of-the-art methods.

1. Introduction

The most common predictive tasks in data mining [17] are classification and regression. Decision trees [29,36] are one of the most popular prediction techniques. The success of tree-based approaches can be explained by their ease of application, speed of operation, and effectiveness. Furthermore, the hierarchical tree structure, where appropriate tests from consecutive nodes are sequentially applied, closely resembles a human method of decision making, which makes decision trees natural and easy to understand even for inexperienced analysts. Regression and model trees [22] are variants of decision trees, and they have been designed to approximate real-valued functions instead of being used for classification tasks. The main difference between a regression tree and a model tree is that, in the latter, a constant value in the terminal node is replaced by a regression plane.

Inducing an optimal model tree, as with the problem of learning an optimal decision tree, is known to be NP-complete [24]. Consequently, practical decision-tree learning algorithms are based on heuristics such as greedy algorithms, where locally optimal decisions are made in each tree node. Such algorithms cannot guarantee to return the globally optimal decision tree. The purpose of this paper is to illustrate the application of a specialized evolutionary algorithm (*EA*) [27] to the problem of model tree induction. The objectives are to show that evolutionary induction may result in finding globally

[*] Corresponding author.

http://dx.doi.org/10.1016/j.ins.2014.07.051
0020-0255/© 2014 Elsevier Inc. All rights reserved.

optimal solutions that are more accurate and less complex than the traditional greedy-induced counterparts and straightforward application of *EA*. This research shows the impact of the application of specialized *EA*s on the tree structure, tests in internal nodes, and models in the leaves. By incorporating the knowledge about global model tree induction, the full potential of *EA*s is exploited. Local optimizations are also proposed for *EA*s problem search, which is known as a memetic algorithm [28,7].

Our previous research showed that global inducers are capable of efficiently evolving accurate and compact univariate regression trees [25], called Global Regression Trees (*GRT*), and model trees with simple linear regression in the leaves [8,10]. In our previous papers, we proposed model trees with multiple linear regression in the leaves [9] and considered how memetic extensions improve the global induction of regression and model trees [11]. This paper reviews and significantly extends our previous work on model trees in almost every step of evolutionary induction. We introduce new specialized operators and local search components that improve pure evolutionary methods and propose a smoothing process to increase the prediction accuracy of the model tree. A new multi-objective optimization strategy (lexicographic analysis) is verified as an alternative fitness function to a weight formula. Additional data sets and new experiments illustrate the advantage of the global search solutions for popular model tree algorithms.

This paper is organized as follows. The following section provides a brief background on model trees, reviews related work, and describes some of the advantages with regard to using EAs for model tree induction. Section 3 describes the approach and demonstrates how each step of the *EA* can be improved. Section 4 presents a validation of the proposed solutions in three sets of experiments. In the last section, the paper is concluded and possible future works are sketched.

The presented experiments demonstrate how each step of the *EA* can be improved.

2. Global vs local induction

Decision trees are often built through a process that is known as a recursive partitioning. The most popular tree-induction is based on the top-down approach [35]. It starts from the root node, where the locally optimal split (test) is searched according to the given optimality measure (e.g., Gini, Twoing, or the entropy rule for classification trees and the least squared or least absolute deviation error criterion for regression trees). Next, the training data is redirected to newly created nodes, and this process is repeated for each node until some stopping-rule is violated. Finally, post-pruning [15] is applied to improve the generalization power of the predictive model. Inducing the decision tree through a greedy strategy is fast and generally efficient in many practical problems, but it usually produces locally optimal solutions.

One of the first and most well-known top-down regression tree solutions is the Classification and Regression Tree (CART) [5]. The method searches for a locally optimal split that minimizes the sum of squared residuals of the model and builds a piecewise constant model with each terminal node fitted by the training sample mean. The following solutions managed to improve the prediction accuracy by replacing single values in the leaves with more advanced models. The M5 system [39] induces a model tree that contains at leaves multiple linear models analogous to piecewise linear functions. The *HTL* [41] is even more advanced and evaluates linear and nonlinear models in terminal nodes.

Multiple authors have proposed methods to limit the negative effects of inducing the decision tree with the greedy strategy. In *SECRET* [13], authors suggest that changing a regression problem into a classification one may help in finding more globally optimal partitions. A different solution was proposed in *SMOTI* [26], where regression models exist not only in the leaves but also in the upper parts of the tree. The authors suggested that this technique allows individual predictors to have both global and local effects on the model tree. A more recent innovation for finding optimal splits in nodes was presented in *LLRT* [42]. The *LLRT* solution can do a near-exhaustive evaluation of all possible splits in a node based on the quality of fit of the linear regression models in the resulting branches.

In the literature, there have been some attempts to apply an evolutionary approach for the induction of decision trees, including regression and model trees. For an extensive review, please refer to [3]. In *TARGET* [16], the authors proposed to evolve a *CART*-like regression tree with simple genetic operators. The Bayesian information criterion (*BIC*) [37] was used as a fitness function, which penalizes the tree for over-parameterization. A more advanced system called *E-Motion* was proposed in [2]. The authors evolved univariate trees with linear models in the leaves and optimized their prediction errors and the tree size. *E-Motion* implements standard 1-point crossover and two different mutation strategies (shrinking and expanding) to variate individuals. The *GPMCC* [32] approach proposed to evolve model trees with non-linear models in the leaves. In most of the papers, performing such a global search in the space of candidate solutions successfully competes with popular greedy methods. However, almost all algorithms from [3] apply only the basic variants of *EA*, which do not incorporate knowledge of the decision tree's induction.

In this paper, we would like to fill this gap by proposing specialized operators and memetic extensions for the evolutionary induction of model trees.

To illustrate the simple scenario where evolutionary induced model trees are beneficial, we prepared two artificially generated datasets, with analytically defined decision borders (1) and (2) illustrated in Fig. 1. Both datasets contain an attribute that is linearly dependent with one or two independent attributes.

The data set on the left (denoted as *split plane3*) can be perfectly predictable with regression lines on subsets of the data resulting from a single partition at threshold $x_1 = -2$, and it is described by Eq. (1). Most of the popular greedy top-down inducers that minimize the residual sum of squares (like *CART*) or standard deviation (like *M5*) will not find the best