Evolutionary Algorithms for Global Decision Tree Induction

Marek Kretowski
Bialystok University of Technology, Poland

Marcin Czajkowski
Bialystok University of Technology, Poland

INTRODUCTION

Data mining (Fayyad, Piatetsky-Shapiro, Smyth & Uthurusamy, 1996) can reveal important and insightful information hidden in data. However, appropriate tools and algorithms are required to effectively identify correlations and patterns within the data. Decision trees (Kotsiantis, 2013) represent one of the main techniques for discriminant analysis prediction in knowledge discovery. The success of tree-based approaches can be explained by their ease of application, fast operation, and effectiveness. Furthermore, the hierarchical tree structure, in which appropriate tests from consecutive nodes are sequentially applied, closely resembles a human way of decision making. All this makes decision trees easy to understand, even for inexperienced analysts.

Despite 50 years of research on decision trees, many problems still remain (Loh, 2014) such as searching only for a locally optimal split in the internal nodes; appropriate pruning criterion, efficient analysis of cost-sensitive data or performing multi-objective optimization. To help resolve some of these problems, evolutionary algorithms (EAs) (Michalewicz, 1996) has been applied to decision tree induction (Barros et al., 2012). The strength of this approach lies in the global search for splits and predictions. It results in higher accuracy and smaller output trees compared to popular greedy decision tree inducers.

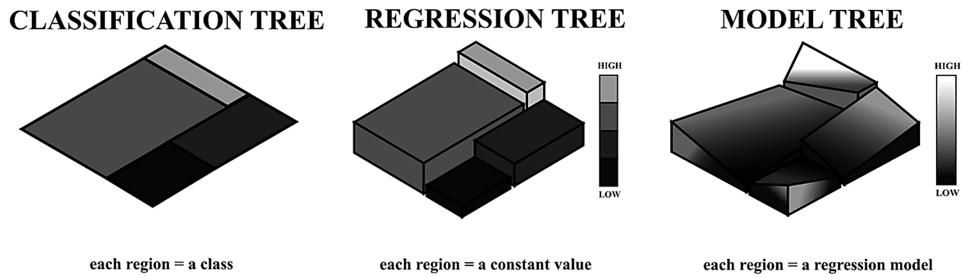
The purpose of this chapter is to illustrate the application of EAs to the problem of decision tree induction. The objectives are to show that evolutionary optimization, compared to the greedy search, may result in finding globally optimal solutions, whose complexity is significantly smaller and the prediction is highly competitive. We will cover the global induction of classification, regression, and model trees.

BACKGROUND

We may find different variants of decision trees in the literature (Loh, 2014). They can be grouped according to the type of problem they are applied to, the way they are induced, or the type of structure. Tree predictors can be used to classify existing data (classification trees) or to approximate real-valued functions (regression trees) (see Figure 1). In each leaf, classification tree assigns a class label (usually the majority class of all instances that reach that particular leaf), while the regression tree holds a constant value (usually an average value for the target attribute). A model tree can be seen as an extension of the typical regression tree (Quinlan, 1992). The constant value in each leaf of the regression tree is replaced in the model tree by a linear (or nonlinear) regression function. To predict the target value, the new tested instance is followed down the tree from a root node to a leaf using its attribute values to make routing decisions at each internal node. Next, the predicted value for the new instance is evaluated based on a regression model in the leaf.

DOI: 10.4018/978-1-5225-2255-3.ch185

Figure 1. An illustration of predicted values of the classification, regression, and model trees



Examples of predicted values of classification, regression, and model trees are given in Figure 1.

The gray level color of each region represents a different class label (for a classification tree), and the height corresponds to the value of the prediction function (regression and model trees).

Inducing optimal decision tree is known to be NP-complete (Naumov, 1991). Consequently, practical decision-tree learning algorithms are based on heuristics such as greedy algorithms where locally optimal splits are made in each node. The most popular tree-induction is based on the top-down approach (Rokach & Maimon, 2005). Top-down induction starts from the root node, where locally best split (test) is searched, according to the given optimality measure (e.g. Gini, towing or entropy rule for classification tree and least squared or least absolute deviation error criterion for regression tree). Next, the training instances are redirected to the newly created nodes and this process is repeated for each node until some stopping-rule is satisfied. The recursive partitioning of the dataset may lead to the data over-fitting, therefore, the decision tree pruning (Esposito, Malerba & Semeraro, 1997) is applied to improve the generalization power of the predictor. For an alternative approaches to greedy decision tree induction, like e.g. bottom-up, please refer to framework proposed by Barros et al. (2014).

Most of tree inducing algorithms partition the feature space with axis decision borders (Sheth & Deshpande, 2015). These types of trees are called univariate decision trees. Split at each non-terminal node usually involves single feature. For a continuous-valued feature usually an inequality test with binary outcomes is applied and for a nominal attribute mutually exclusive groups of attribute values are associated with outcomes. One of the first and most well-known solution that can be applied to classification and regression problem is CART (Breiman, Friedman, Olshen & Stone, 1984) system. Good representatives of univariate inducers are also systems developed by Quinlan: C4.5 (1993) for classification and M5 (1992) for regression.

When more than one feature is taken into account to build a test in non-terminal node, then we deal with multivariate decision trees. The most common form of such a test is an oblique split, which is based on a linear combination of features (hyper-plane). The decision tree which applies only oblique tests is often called oblique or linear, whereas heterogeneous trees with univariate, linear and other multivariate (e.g., instance-based) tests can be called mixed decision trees (Llora & Wilson, 2004). It should be emphasized that computational complexity of the multivariate induction is generally significantly higher than the univariate induction. OC1 (Murthy, Kasif & Salzberg, 1994) is a good examples of multivariate decision tree system.

Inducing the decision tree with greedy strategy usually leads to suboptimal solutions. They search only for locally best splits at each node which does not guarantee the globally best solution. One of the alternatives is the ensemble of trees (Seni & Elder, 2010), which is created by the induction of different trees from the training sample. Ensemble classifiers like Random Forests (Breiman, 2001) induce many decision trees whose predictions