Cost-Sensitive Extensions for Global Model Trees: Application in Loan Charge-Off Forecasting

Marcin Czajkowski[1], Monika Czerwonka[2], and Marek Kretowski[1]

[1] Faculty of Computer Science, Bialystok University of Technology, Wiejska 45a, 15-351 Bialystok, Poland
{m.czajkowski,m.kretowski}@pb.edu.pl
[2] Collegium of Management and Finance, Warsaw School of Economics, Al. Niepodleglosci 162, 02-554 Warsaw, Poland
monika.czerwonka@sgh.waw.pl

Abstract. Most of regression learning methods aim to reduce various metrics of prediction errors. However, in many real-life applications it is prediction cost, which should be minimized as the under-prediction and over-prediction errors have different consequences. In this paper, we show how to extend the evolutionary algorithm (EA) for global induction of model trees to achieve a cost-sensitive learner. We propose a new fitness function which allows minimization of the average misprediction cost and two specialized memetic operators that search for cost-sensitive regression models in the tree leaves. Experimental validation was performed with bank loan charge-off forecasting data which has asymmetric costs. Results show that Global Model Trees with the proposed extensions are able to effectively induce cost-sensitive model trees with average misprediction cost significantly lower than in popular post-hoc tuning methods.

Keywords: cost-sensitive regression, asymmetric costs, evolutionary algorithms, model trees, loan charge-off forecasting.

1 Introduction

In the vast number of contemporary systems, information including business, research and medical issues is collected and processed. In real-life data mining problems, the traditional minimization of prediction errors may not be the most adequate scenario. For example, in medical domain misclassifying an ill patient as a healthy one is usually much more harmful than treating a healthy patient as an ill one and sending him for additional examinations. In finance, investors tend to sell winning stocks more readily than losing stocks in the sense that they realize gains relatively more frequently than losses. The sadness that one experiences in losing the money appears to be greater than the pleasure of gaining the same amount of money. This strong loss aversion was explained and described in the prospect theory by Kahneman and Tversky [14] and applied to finance practice by Shefrin and Statman [25].

J. Świątek et al. (eds.), *Advances in Systems Science*,
Advances in Intelligent Systems and Computing 240,
DOI: 10.1007/978-3-319-01857-7_30,

In this paper, we want to tackle the cost-sensitive regression methods. We focus on extending the existing *EA* for model tree induction to handle data with asymmetric costs.

1.1 Background

The decision trees [22] are one of the most widely used prediction techniques. Ease of application, fast operation and what may be the most important, effectiveness of decision trees, makes them powerful and popular tool [15]. Regression and model trees [13] may be considered as a variant of decision trees, designed to approximate real-valued functions instead of being used for classification tasks. The main difference between regression tree and model tree is that, in the latter, constant value in the terminal node is replaced by a regression plane. Each leaf of the model tree may hold a linear (or nonlinear) model whose output is the final prediction.

Problem of learning an optimal decision tree is known to be NP-complete. Consequently, classical decision-tree learning algorithms are built with a greedy top-down approach [21] which usually leads to suboptimal solutions. Recently, application of *EA*s [18] to the problem of decision tree induction [2] become increasingly popular alternative. Instead of local search, *EA* performs a global search in the space of candidate solutions. Trees induced with *EA* are usually significantly smaller in comparison to greedy approaches and highly competitive in terms of prediction accuracy [17,7]. On the other hand, the induction of global regression and model trees is much slower [8]. One of the possible solutions to speed up evolutionary approach is a combination of *EA*s with local search techniques, which is known as Memetic Algorithms [12].

Cost-sensitive prediction is the term which encompasses all types of learning where cost is considered [28,10] e.g., costs of tests (attributes), costs of instances, costs of errors. In this paper, we only focus on asymmetric costs, which are associated with different types of prediction errors.

The vast majority of data mining algorithms is applied only to the classification problems [27] while cost-sensitive regression is not really studied outside of statistic field [3]. In induction of cost-sensitive classification trees, three techniques are popular:

- convert classical decision tree into cost-sensitive one, mainly by changing the splitting criteria and/or adopting pruning techniques for incorporating misclassification costs (e.g. [4]);
- application of *EA*s that induce cost-sensitive trees [16];
- application of universal methods like: cost instance-weighting [26] or post-hoc tuning solutions e.g. MetaCost [9].

One of the earliest studies of asymmetric costs in regression was performed by Varian [30]. Author propose *LinEx* loss function which is approximately linear on one side and exponential on the other side as an alternative to popular least squared procedures. Application of different loss functions was later extended