Available online at www.sciencedirect.com

ScienceDirect

Decision Support Systems 43 (2007) 809 – 826

Decision Support Systems

www.elsevier.com/locate/dss

A multi-objective genetic programming approach to developing Pareto optimal decision trees

Huimin Zhao *

Sheldon B. Lubar School of Business, University of Wisconsin-Milwaukee, P. O. Box 742, Milwaukee, WI 53201, United States

Received 1 March 2006; received in revised form 5 December 2006; accepted 15 December 2006
Available online 28 December 2006

Abstract

Classification is a frequently encountered data mining problem. Decision tree techniques have been widely used to build classification models as such models closely resemble human reasoning and are easy to understand. Many real-world classification problems are cost-sensitive, meaning that different types of misclassification errors are not equally costly. Since different decision trees may excel under different cost settings, a set of non-dominated decision trees should be developed and presented to the decision maker for consideration, if the costs of different types of misclassification errors are not precisely determined. This paper proposes a multi-objective genetic programming approach to developing such alternative Pareto optimal decision trees. It also allows the decision maker to specify partial preferences on the conflicting objectives, such as false negative vs. false positive, sensitivity vs. specificity, and recall vs. precision, to further reduce the number of alternative solutions. A diabetes prediction problem and a credit card application approval problem are used to illustrate the application of the proposed approach.

Keywords: Data mining; Binary classification; Decision tree; Cost-sensitive classification; Genetic programming; Multi-objective optimization; Pareto optimality

1. Introduction

As modern organizations can nowadays collect and maintain huge volumes of data, many of them have started to employ data mining techniques to mine their operational data for interesting patterns and decision models that could be used to support their decision making. Classification is a frequently encountered predictive data mining problem where a categorical dependent variable needs to be predicted based on a set of independent variables. Various classification techniques have been developed in such fields as multivariate statistical analysis, machine learning, and artificial neural networks [56,22] and applied to solve a variety of classification problems, such as workplace Web usage profiling [1], purchase behavior prediction [23], sales profile forecasting [52], deception detection [62], credit evaluation [18,48], bankruptcy prediction [45,51], bank failure prevention [46], intrusion detection [28,63], and medical diagnosis [33]. These classification techniques automatically induce prediction models, called classifiers, based on training examples with known outcomes. The trained classifiers can then be applied to predict the outcomes of new problem instances in the future.

Decision tree techniques have been widely used in building classification models as such models closely

* Tel.: +1 414 229 6524; fax: +1 414 229 5999.
E-mail address: hzhao@uwm.edu.

0167-9236/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2006.12.011

resemble human reasoning and are easy to understand [56]. Decision trees are sequential models, which logically combine a sequence of simple tests; each test compares a numeric attribute against a threshold value or a nominal attribute against a set of possible values. Such symbolic classifiers have an advantage over "black-box" models, such as neural nets, in terms of comprehensibility. The logical rules followed by a decision tree are much easier to interpret than the numeric weights of the connections between the nodes in a neural network. Decision makers tend to feel more comfortable to use models that they can understand.

Many real-world classification problems are cost-sensitive, meaning that different types of misclassification errors (false positive and false negative in a binary classification problem) are not equally costly [9,39,48]. For example, to a bank, approving a potentially bad loan is more costly than denying a good loan. Similarly, to a banking regulatory agent, leaving a problem bank unnoticed has more serious consequences than predicting a healthy bank as being problematic and thus scheduling unnecessary on-site examination. Such asymmetric costs need to be incorporated into the classifier training process such that the expected misclassification cost, rather than plain error rate, is sought to be minimized. A general method for cost-sensitive classifier training is to weight the different classes of training examples according to the costs such that more costly errors are penalized more heavily [6,53,58]. Using this method, different cost settings tend to result in very different decision trees, each of which may be superior to others in a particular range of cost settings.

It is often very difficult for decision makers to precisely pinpoint the costs of different types of misclassification errors, although they may be more confident in specifying a reasonable range for the cost ratio between false positive and false negative. For example, a banker may consider approving a bad loan to be between ten and twenty times more costly than denying a good loan. It is also possible that a decision maker's assessment of the cost ratio is dependent on the context of decision making. For example, a banking regulatory agent may assign a more extreme cost ratio when scheduling on-site examinations to prevent likely bank failures than when preparing for the consequences of eventual bank failures. Although bank failures are very costly, most of them can be prevented via supervision actions and only a few of them do occur. Thus, the decision maker may need to consider a range of possible cost ratios.

Since different decision trees may excel under different cost ratios, a set of decision trees should be developed and presented to the decision maker for consideration, if the costs of different types of misclassification errors are not precisely determined. In this paper, we propose a multi-objective genetic programming approach to developing such alternative decision trees. These alternatives are said to be non-dominated or Pareto optimal, in that each of them is better than any other on at least one of two conflicting objectives, e.g., minimizing false negative rate vs. minimizing false positive rate. The system we have implemented also allows the decision maker to specify partial preferences on the two conflicting objectives to further reduce the number of alternative solutions. The preference (or tradeoff) can be similarly made on other pairs of performance measures, such as sensitivity vs. specificity and recall vs. precision, which have been typically employed in some domains, such as medical diagnosis and information retrieval. We have applied the system on several binary classification datasets publicly available from the UCI machine learning repository [37] and will present the results on two of these datasets. This paper makes a unique contribution by formulating cost-sensitive classification as a multi-objective optimization problem and providing an evolutionary computation approach.

The rest of the paper is organized as follows. We first briefly review the background and related literature on cost-sensitive classification and multi-objective evolutionary computation in the next section. We then present a diabetes prediction problem and a credit card application approval problem as motivational examples in Section 3. We then propose a multi-objective genetic programming approach to developing Pareto optimal decision trees and illustrate its application in the diabetes prediction and credit card application approval examples in Sections 4 and 5. Finally, we summarize the contributions of this work and discuss future research directions in Section 6.

2. Background and related work

In this section, we briefly review the background and related literature on cost-sensitive classification and multi-objective evolutionary computation.

2.1. Cost-sensitive classification

In this paper, we consider *binary* classification, where a binary dependent variable y, referred to as the class, needs to be predicted based on a vector of independent variables (also called attributes) $\mathbf{x}=(x_1,x_2,\ldots, x_m)$. A problem instance is considered positive (or negative) if