Contents lists available at ScienceDirect

Information Sciences

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

Time-constrained cost-sensitive decision tree induction

Yen-Liang Chen[a,*], Chia-Chi Wu[b], Kwei Tang[c]

[a] Department of Information Management, National Central University, Chung-Li 320, Taiwan, ROC
[b] Data Analytics Technology & Applications Research Institute, Institute for Information Industry, Taipei 105, Taiwan, ROC
[c] Department of Business Administration, National Chengchi University, Taipei 116, Taiwan, ROC

ARTICLE INFO

Article history:
Received 7 July 2015
Revised 2 March 2016
Accepted 10 March 2016
Available online 18 March 2016

Keywords:
Data mining
Decision analysis
Cost-sensitive learning
Classification
Decision tree

ABSTRACT

A cost-sensitive decision tree is induced for the purpose of building a decision tree from training data that minimizes the sum of the misclassification cost and test cost. Although this problem has been investigated extensively, no previous study has specifically focused on how the decision tree can be induced if the classification task must be completed within a limited time. Accordingly, we developed an algorithm to generate a time-constrained minimal-cost tree. The main idea behind the algorithm is to select the attribute that brings the maximal benefit when time is sufficient, and to select the most time-efficient attribute (i.e., the attribute that provides maximal benefit per unit time) when time is limited. Our experimental results show that the performance of this algorithm is highly satisfactory under various time constraints across distinct datasets.

1. Introduction

Classification is a data analysis method that can be used to extract models from training data in order to predict future data classes [7]. Various classification techniques have been proposed previously, such as decision tree classifiers [10,22,24], Bayesian classifiers [7], neural networks [27], and case-based reasoning [13]. Among these, decision trees are probably the most widely recognized and commonly used classification models because they offer several advantages, such as ease of interpretation, computational efficiency, and ability to generate understandable classification rules [7,17,25].

Decision trees enable representation of the rules underlying data: decision trees are hierarchical and sequential structures that recursively partition data [17]. In a decision tree, each internal node denotes a test on an attribute, each branch represents an outcome of this test, and each leaf node represents a class or a label. Decision tree classifiers have been successfully used in a wide range of applications, and have generated numerous studies of diverse types; previous studies have reported that decision trees can be induced using ID3 [22], C4.5 [24], QUEST [11], and GATree [18]; methods of pruning trees, which address the overfitting problem, have been proposed in [21,22,23,38]; and the matter of scalability has been addressed in [3,6,14,15,26,28]. Other related topics investigated include the production of learning trees from multivalued and multilabeled data [40] and by using a semisupervised approach [30]. Almost all of the aforementioned studies were aimed at maximizing classification accuracy and minimizing classification error. Because a classification task might incur various types of cost, increasing attention has been devoted recently to the construction of cost-sensitive decision trees [12].

Of the various types of costs recognized in previous studies, the two most commonly observed are misclassification costs and test costs. In most studies on this subject, misclassification cost has been defined as the cost of inaccurately assigning

[*] Corresponding author.
E-mail address: ylchen@mgt.ncu.edu.tw (Y.-L. Chen).

http://dx.doi.org/10.1016/j.ins.2016.03.022

a case to Class i when it belongs to Class j. The cost of misclassification error might be highly unbalanced. For example, classifying a sick patient as a healthy patient is often considerably more costly than labeling a healthy patient as a sick patient [1]. In addition to the misclassification cost, each test performed might include an associated cost; for example, in medical diagnosis, a blood test incurs a cost [32].

In various real-world applications, a task must be completed within a time limit. For example, whether a customer's credit is good or bad must be determined in time to support decision-making. Similarly, the results of a patient's medical tests must be obtained before subsequent treatment or the next clinical appointment. Therefore, in addition to identifying misclassification cost and test costs, it is critical to determine how a test sequence can be prevented from being exceedingly time-consuming. However, to the best of our knowledge, no previous research has addressed this matter.

In this paper, we propose a new algorithm for time-constrained cost-sensitive decision tree induction. This algorithm is designed specifically for minimizing the misclassification cost and test cost under a time limit, and is based on this main concept: the attribute that brings the maximal benefit is selected when time is sufficient, whereas the most time-efficient attribute (i.e., the attribute that brings maximal benefit per unit time) is selected when time is limited. Because the time required for a classification task is extremely challenging to estimate before the task is completed, our algorithm first builds an initial decision tree based on the most time-efficient criterion. Subsequently, the focus is on the nodes that have time remaining to further improve the decision tree based on the maximal benefit criterion.

The contribution of this paper is twofold. First, we propose a new cost-sensitive decision tree induction problem by adding a time limit into the traditional cost-sensitive model. Second, an algorithm is developed to resolve the proposed problem. The remainder of this paper is organized as follows: first, we review related studies in Section 2, and then formalize the addressed problem in Section 3. The proposed algorithm is introduced in Section 4, and its performance evaluation based on a real-world case study is described in Sections 5 and 6. In Section 7, we outline our conclusions and suggestions for future work.

2. Related work

Cost-sensitive classification has been investigated extensively [12]. Among all cost components, misclassification cost and test cost are probably the most critical factors examined previously. For example, misclassification cost was considered in [4,19,20,34] and test cost in [16,39]. Moreover, both costs were considered concurrently in several studies. For example, Turney [31] introduced ICET (Inexpensive Classification with Expensive Tests), a genetic algorithm for cost-sensitive classification whose fitness function is the average cost of classification, which includes both test costs and classification-error costs. As the splitting criterion for selecting attributes, Ling et al. [8] used the sum of the misclassification cost and test cost instead of information gain. Chai et al. [2] used the test cost and misclassification cost to train naïve Bayesian classifiers, whereas Yang et al. [39] integrated the works of Chai et al. [2] and Ling et al. [8] to build a TCSL (Test-Cost-Sensitive Learning) framework. Ling et al. [9] proposed a lazy decision tree learning algorithm that minimizes the total cost of testing and misclassification and also proposed several novel test strategies. Sheng and Ling [29] considered the problem that distinct delay times might be involved in obtaining the results of medical tests from the laboratories, and proposed an algorithm to minimize the sum of the attribute-acquisition costs, delay costs, and misclassification costs. Zhang [41] also considered the delay cost, and thus selected the splitting attributes according to the ratio of returns (reduction of misclassification cost) and investments (attribute cost and delay cost). Deng and Jeong-Young [5] defined the concept of "correct classification benefit," and then built a novel decision tree based on the cost and benefit dual-sensitive (CBDSDT) process; in this method, the optimal classification result featuring lowest costs and highest benefits is obtained by considering the test cost, misclassification cost, attribute information, and correct classification benefit.

Cost optimization was considered in all of the aforementioned studies on classification. However, in several real-world applications, a classification task must additionally be completed within a specified period. For example, as noted in the preceding section, the results of medical screening for a patient must be available before subsequent treatment or the next clinical appointment. Similarly, decisions on fraud detection must be made quickly in order to limit potential damages. Given this requirement to include time constraints in classification tasks, this study was devoted to minimizing misclassification costs and test costs within a limited time.

3. Problem definition

3.1. Training data and decision trees

A classifier is learned through a learning algorithm from training data, each record of which contains a set of attribute values and a label. Let A denote the attribute set, and a_x the xth attribute in A. The decision tree $\boldsymbol{T}$ is a directed acyclic graph. Each decision tree contains a root, a finite set of nodes (internal nodes and leaf nodes), and a set of edges that link two nodes. An example of a decision tree is shown in Fig. 1, where $H(\boldsymbol{T})$, the height of the decision tree $\boldsymbol{T}$, is defined as the number of levels in $\boldsymbol{T}$.

In the decision tree, let n_i be the node numbered i. If n_i is an internal node, an attribute $a(n_i)$ will be associated with n_i, where $a(n_i)$ is the test attribute used in the node n_i. Conversely, a leaf node in the decision tree is associated with a label.