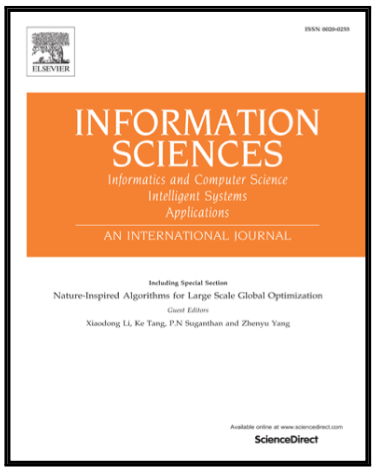

Cost-sensitive and Hybrid-attribute Measure Multi-decision Tree over Imbalanced Data Sets

Fenglian Li, Xueying Zhang, Xiqian Zhang, Chunlei Du, Yue Xu, Yu-Chu Tian

PII: S0020-0255(17)30478-4
DOI: 10.1016/j.ins.2017.09.013
Reference: INS 13117

To appear in: *Information Sciences*

Received date: 7 February 2017
Revised date: 15 July 2017
Accepted date: 5 September 2017

Please cite this article as: Fenglian Li, Xueying Zhang, Xiqian Zhang, Chunlei Du, Yue Xu, Yu-Chu Tian, Cost-sensitive and Hybrid-attribute Measure Multi-decision Tree over Imbalanced Data Sets, *Information Sciences* (2017), doi: 10.1016/j.ins.2017.09.013

Cost-sensitive and Hybrid-attribute Measure Multi-decision Tree over Imbalanced Data Sets

Fenglian Li[a,b], Xueying Zhang[a,*], Xiqian Zhang[a], Chunlei Du[a], Yue Xu[b], Yu-Chu Tian[a,b,*]

[a] *College of Information Engineering, Taiyuan University of Technology, Taiyuan, Shanxi 030024, China*
[b] *School of Electrical Engineering and Computer Science, Queensland University of Technology, GPO Box 2434, Brisbane QLD 4001, Australia*

Abstract

One of the most popular algorithms for classification is the decision tree. However, existing binary decision tree models do not handle well the minority class over imbalanced data sets. To address this difficulty, a Cost-sensitive and Hybrid attribute measure Multi-Decision Tree (CHMDT) approach is presented in this paper. It penalizes misclassification through a hybrid attribute measure, which is defined from the combination of the Gini index and information gain measure. It further builds a multi-decision tree consisting of multiple decision trees each with different root node information. The overall objective of the approach is to maximize the classification performance with the hybrid attribute measure while minimizing the total misclassification cost. Experiments are conducted over twelve KEEL imbalanced data sets to demonstrate the CHMDT approach. They show that the classification performance of the minority class is improved significantly without sacrifice of the overall classification accuracy of the majority class.

Keywords: Multi-decision tree, minority class, imbalanced data set, cost sensitivity, hybrid attribute measure

*Corresponding author: Xueying Zhang (email: tyzhangxy@163.com) and Yu-Chu Tian (email: y.tian@qut.edu.au).

Email addresses: ghllfl@163.com (Fenglian Li), tyzhangxy@163.com (Xueying Zhang), yue.xu@qut.edu.au (Yue Xu), y.tian@qut.edu.au (Yu-Chu Tian)