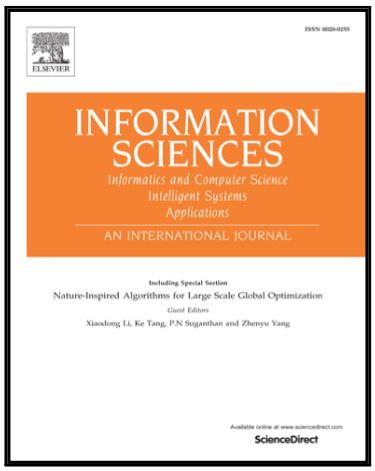

Novel Algorithms for Cost-Sensitive Classification and Knowledge Discovery in Class Imbalanced Datasets with an Application to NASA Software Defects

Michael J. Siers, Md Zahidul Islam

PII: S0020-0255(18)30401-8
DOI: 10.1016/j.ins.2018.05.035
Reference: INS 13670

To appear in: *Information Sciences*

Received date: 9 January 2018
Revised date: 27 March 2018
Accepted date: 13 May 2018

Please cite this article as: Michael J. Siers, Md Zahidul Islam, Novel Algorithms for Cost-Sensitive Classification and Knowledge Discovery in Class Imbalanced Datasets with an Application to NASA Software Defects, *Information Sciences* (2018), doi: 10.1016/j.ins.2018.05.035

ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Inf. Sci

Information Sciences 00 (2018) 1–25

Novel Algorithms for Cost-Sensitive Classification and Knowledge Discovery in Class Imbalanced Datasets with an Application to NASA Software Defects

Michael J. Siers & Md Zahidul Islam

School of Computing and Mathematics, Charles Sturt University, Bathurst, NSW, Australia

msiers@csu.edu.au, zislam@csu.edu.au

Abstract

Software defect prediction (SDP) involves using machine learning to locate bugs in source code. Datasets used for SDP are typically affected by an issue called class imbalance. Traditional learning algorithms do not perform well on class imbalanced datasets. Cost-sensitive learning has been used in SDP to minimise the monetary costs incurred by predictions. We propose a framework which produces cost-sensitive predictions and also mitigates class imbalance. Since our algorithm builds a decision forest classifier, knowledge can be extracted by manual inspection of the individual decision trees. To enhance this knowledge discovery process, we propose an algorithm for extracting the most interesting patterns from a decision forest. Our algorithm calculates interestingness as the potential financial gain of knowing the pattern. We then present a process which combines the above-mentioned techniques into an end-to-end cost-sensitive knowledge discovery process. This process is demonstrated by extracting knowledge from four software projects undertaken by the National Aeronautics and Space Administration (NASA).

Keywords: Software Defect Prediction, Class Imbalance, Cost-Sensitive, Decision Forest, Knowledge Discovery

1. Introduction

Predicting which sections of code contain bugs is a process called Software Defect Prediction (SDP) [14, 17, 11, 20, 12, 19, 21, 22]. These sections of code are referred to as modules. If a module contains at least one bug, it is considered *defective*. In many SDP studies, each C/Java function is considered as one module [20]. Therefore, these studies separated a software project's source code into each function, then predicted which function contained bugs.

SDP studies either take a non cost-sensitive approach [14, 17, 11] or a cost-sensitive approach [12, 19, 21, 22]. When performed non-cost-sensitively, the aim is to make as many correct predictions as possible. Below, we explain how certain types of predictions incur different real life costs. When SDP is performed cost-sensitively, the aim is to minimise costs which are incurred by the predictions. This study focuses on cost-sensitive SDP.