University of Rijeka
FACULTY OF ENGINEERING

hrzz
Hrvatska zaklada za znanost

Software Defect Classification with a Variant of NSGA-II and Simple Voting Strategies

Emil Rubinić, Goran Mauša and Tihana Galinac Grbac

{erubinic, gmausa, tgalinac}@riteh.hr

SEIPLAB: Software Engineering and Information Processing Laboratory

Motivation for this work

- Student project: Bug-Code analyser Buco tool
- Research project funded by Croatian Science Foundation:
  - **Evolving Software Systems: Analysis and Innovative Approaches for Smart Management, EVOSOFT**
    - http://www.seiplab.riteh.uniri.hr/?page_id=712&lang=en

Introduction

- Increase in software complexity results in huge verification space
  – Finding faults at early stage is important!
- Can we build a good model based on software project datasets for early fault prediction?
  – Machine based learning algorithms do not provide good results for imbalanced datasets

Related work

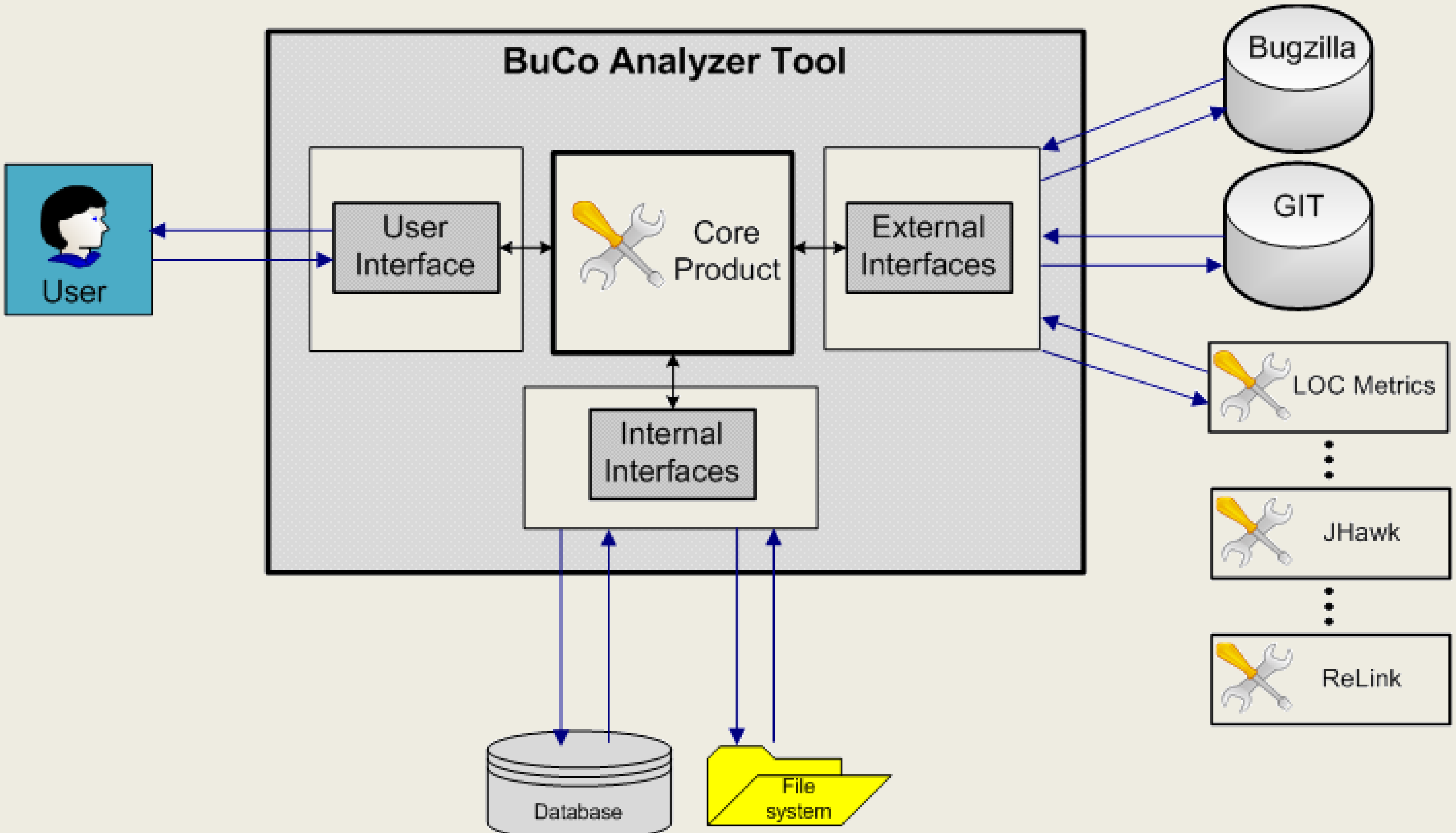
- Genetic programing may provide a solution
  - Bhowan, U., Johnston, M., Zhang, M., Yao, X.: Evolving diverse ensembles using genetic programming for classification with unbalanced data. IEEE TEC 17(3), 368–386 (2013)
- Several approaches have been studied and AdaBoost.NC gives best performance
  - Wang, S., Yao, X. "Using Class Imbalance Learning for Software Defect Prediction," IEEE Transactions on Reliability, 62(2):434-443, June 2013

Experiment

Stages of the experiment

1. Data collection and preparation
2. Building a model for classification
3. Simple voting strategies
4. Comparing results

Data collection using Bug-Code Analyzer **BUCO** tool



http://www.seiplab.riteh.uniri.hr

- Datasets used in article
  - Eclipse Plug-in Development Environment (PDE) project versions - 2.0 , 2.1 and 3.0

Version	**Whole set**				
	Attributes		FP	NFP	Total
	No.	Type	(%)	(%)	No.
$PDE_{2.0}$	48	Integer Decimal	19%	81%	576
$PDE_{2.1}$			16%	84%	761
$PDE_{3.0}$			31%	69%	881

- Default task – binary classification
  - File is classified as Fault Prone (FP) if contains at least one fault, otherwise it is Non fault Prone (NFP)
- Each version was randomly divided 50 times:
  - 50% for training and 50% for test

Building a model for SDC

- Matlab variant of NSGA-II (mNSGA-II)
  - 97 decision variables:

$$X = [w_1, w_2, \dots, w_{48}, o_1, o_2, \dots, o_{47}, o_{48}, \varepsilon]$$

a – dataset attribute; w – weight; o – arithmetical operator from {+, - , · , / } ; ε - noise

  - If C>1 the file is classified as FP, otherwise it is NFP

$$C = [(w_1 \cdot a_1) \cdot o_1 \cdot (w_2 \cdot a_2) \cdot o_2 \dots (w_{48} \cdot a_{48})] \cdot o_{48} \cdot \varepsilon$$

– Two objectives:

- Sensitivity (TPR) and Specificity (TNR):

$$TPR = \frac{TP}{FN + TP}$$

$$TNR = \frac{TN}{TN + FP}$$

T – True		Real State	
F – False		1	0
Prediction	1	TP	FP
	0	FN	TN

– mNSGA-II minimizes multiple functions:

$$minimize(1 - TPR) \qquad minimize(1 - TNR)$$

– mNSGA-II settings:

- 3 sub-populations of size 200
- Algorithm runs for max . 100 generations
- Each run returns one Pareto Approximated (PA) front

mNSGA-II results

- For evaluating evolved fronts trapezoidal numerical integration was used – hyperarea
  - Best fronts for $PDE_{2.0}$ (0.81 ± 0.02)
    - Smallest dataset
  - $PDE_{2.1}$ (0.74 ± 0.03) and $PDE_{3.0}$ (0.74 ± 0.01) – similar results
    - $PDE_{3.0}$ is larger than $PDE_{2.1}$ but more balanced

Making use of population

- For each mNSGA-II output majority voting
  1. Individuals on PA front (PF vote)
  2. Individuals on PA front *without individuals with TPR or TNR less than 0.5* (RPF vote)
  3. Final population (FP vote)
  4. Final population *without individuals with TPR or TNR rate than 0.5* (RFP vote)

Voting strategies results

- Evaluation was made in terms of zenith point (z):

$$z = \sqrt{(1-TPR)^2 + (1-TNR)^2}$$

- TNR is greater than TPR in most tasks
  - Dataset is imbalanced!
- RPF and RFP
  - More balanced and less dispersed results
- RFP-vote has produced best overall results

Conclusion and future work

- Use of entire final population together with removing border solutions can led to better model creation
  - Solutions in the middle region are more desirable
- **Future work:**
  - In this study mNSGA-II is used, but there exist MOEAs which tends to create more solution on middle region
    - SPEA-II
  - Explore other objective formulations (etc. AUC)
  - Extending datasets

University of Rijeka
FACULTY OF ENGINEERING

Hrvatska zaklada za znanost

Questions ?