MAHAKIL: Diversity based Oversampling Approach to Alleviate the Class Imbalance Issue in Software Defect Prediction

Extended Abstract*

Kwabena E. Bennin[1], Jacky Keung[1], Passakorn Phannachitta[2], Akito Monden[3] and Solomon Mensah[1]

[1]Department of Computer Science, City University of Hong Kong, Hong Kong
[2]College of Arts, Media and Technology, Chiang Mai University, Thailand
[3]Graduate School of Natural Science and Technology, Okayama University, Japan
{kebennin2-c,Jacky.Keung,smensah2-c}@my.cityu.edu.hk,passakorn.p@cmu.ac.th,monden@okayama-u.ac.jp

ABSTRACT

This study presents MAHAKIL, a novel and efficient synthetic oversampling approach for software defect datasets that is based on the chromosomal theory of inheritance. Exploiting this theory, MAHAKIL interprets two distinct sub-classes as parents and generates a new instance that inherits different traits from each parent and contributes to the diversity within the data distribution. We extensively compare MAHAKIL with five other sampling approaches using 20 releases of defect datasets from the PROMISE repository and five prediction models. Our experiments indicate that MAHAKIL improves the prediction performance for all the models and achieves better and more significant *pf* values than the other oversampling approaches, based on robust statistical tests.

CCS CONCEPTS

• **Software and its engineering → Empirical software validation**; *Software defect analysis*; *Search-based software engineering*;

KEYWORDS

Software defect prediction, Class imbalance learning, Synthetic sample generation, Data sampling methods, Classification problems

ACM Reference Format:
Kwabena E. Bennin[1], Jacky Keung[1], Passakorn Phannachitta[2], Akito Monden[3] and Solomon Mensah[1]. 2018. MAHAKIL: Diversity based Oversampling Approach to Alleviate the Class Imbalance Issue in Software Defect Prediction: Extended Abstract. In *ICSE '18: ICSE '18: 40th International Conference on Software Engineering , May 27-June 3, 2018, Gothenburg, Sweden.* , Article 4, 1 page. https://doi.org/10.1145/3180155.3182520

1 PROBLEM AND MOTIVATION

Highly unbalanced data typically makes accurate predictions difficult. Unfortunately, software defect datasets naturally contain fewer defective modules than the non-defective modules. Synthetic oversampling techniques address this concern by creating new minority defective modules to balance the minority and majority class distribution. Notwithstanding the successes attained by these techniques, they sometimes generate erroneous data instances, which lead to high false positives. They also tend to generate less diverse data points and data points very similar to existing samples because they consider only nearest neighbor samples. We propose a novel oversampling technique that generates diverse synthetic data restricted within the minority class region.

2 APPROACH AND UNIQUENESS

Our proposed oversampling approach takes inspiration from the field of biology by considering how diverse populations of living organisms differ although they are of the same species. This phenomenon, explained by the Chromosomal Theory of Inheritance [1], justifies how an offspring inherits traits from their parents by obtaining chromosomes from each parent in equal quantity. The proposed approach named MAHAKIL (oversampling based on the theory of inheritance and the Mahalanobis distance), increases the diversity within the minority class by uniquely creating new synthetic minority instances based on a "typical" case (having a small diversity measure distance value) and an "atypical" case (having a large diversity measure distance value) so that the resultant instance becomes not too typical and not too atypical. New samples are generated by computing the average between two instances merged together based on the Mahalanobis distance disparity.

For empirical validations, twenty imbalanced defect datasets from the PROMISE repository and five classification algorithms (C4.5, NNET, KNN, RF, SVM) are considered and evaluated with the *pd* and *pf* measures. Based on robust Brunner statistical test and *win-tie-loss* statistics, our results show that MAHAKIL significantly improves the recall (*pd*) performance and reduces the *pf* for all five models over SMOTE, Borderline-SMOTE, ADASYN and ROS.

ACKNOWLEDGMENTS

This work is supported in part by the General Research Fund of the Research Grants Council of Hong Kong (No. 11208017 and 11214116), the research funds of City University of Hong Kong (No. 7004683) and JSPS KAKENHI 17K00102.

REFERENCES

[1] Walter S Sutton. 1903. The chromosomes in heredity. *The Biological Bulletin* 4, 5 (1903), 231–250.

*The full version is available on http://ieeexplore.ieee.org/document/7990590/

ICSE '18, May 27-June 3, 2018, Gothenburg, Sweden
© 2018 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-5638-1/18/05.
https://doi.org/10.1145/3180155.3182520