**Loyalty-SMOTE: Data Synthesis Algorithm for Effective Imbalanced Data Classification**

Shengquan Hu a, Junfei Li b, Zefeng Li a, Zihao Zhang a,Yan Feng a\*,K. L. Eddie Law c

*a College of Information Engineering, Northwest A&F University, Yangling, Shaanxi,* *China*

*b Advanced Robotics and Intelligent Systems Laboratory, School of Engineering, University of Guelph, Guelph, Canada*

*c Faculty of Applied Sciences, Macao Polytechnic University, Macao, China*

**Abstract**

Imbalanced datasets are always problematic in training machine learning models, such that classifiers often struggle to achieve satisfactory performance. Numerous approaches have been developed to tackle imbalanced data problems. Among them, some data-level methods perform linear interpolations between neighboring minority class samples to genrate new data points, while others focus on oversampling boundary samples which are specific to certain classes. However, many methods fail to consider scenarios involving noise susceptibility. In this paper, we propose a novel data-level method called the Loyalty-SMOTE algorithm. We introduce the concept of Loyalty to identify noise and boundaries within datasets. After identifying potential noisy datapoints, SMOTE(Synthetic Minority Oversampling Technique) algorithm is applied to oversample the minority class boundary data. Sub sequently, a denoising process based on Loyality is conducted to obtain a balanced dataset. To extend our design, the concept of Attraction is introduced to generalize the denoising technique for multiclass problems. In our study, the SVM (Support Vector Ma chine) classifier is used as our base learner, extensive experiments are performed to evaluate and compare different algorithms. Our results demonstrate that Loyalty-SMOTE achieved superior performance across multiple metrics on both binary and multiclass UCI datasets. For 25 binary datasets, it achieved the highest scores in 21 datasets (84%) for F1-score, 24 datasets (96%) for AUROC, 21 datasets (84%) for recall, and 22 datasets (88%) for G-mean. For 5 multiclass datasets, our design achieved scores of 0.8317, 0.6153, 0.8537, and 0.6717, respectively.

*Keywords:* Classification, imbalanced data, Loyalty-SMOTE, SMOTE

# Author:

Shengquan Hu

College of Information Engineering, Northwest A&F University

3 Tai Cheng Road, Yangling Demonstration Zone

Xianyang, Shaanxi, China

Email: [hushengquan@nwafu.edu.cn](mailto:hushengquan@nwafu.edu.cn)

Junfei Li

Advanced Robotics and Intelligent Systems Laboratory, School of Engineering, University of Guelph

50 Stone Road East

Guelph, Ontario, Canada

Email: [jli64@uoguelph.ca](mailto:jli64@uoguelph.ca)

Zefeng Li

College of Information Engineering, Northwest A&F University

3 Tai Cheng Road, Yangling Demonstration Zone

Xianyang, Shaanxi, China

Email: [15648169267@163.com](mailto:15648169267@163.com)

Zihao Zhang

College of Information Engineering, Northwest A&F University

3 Tai Cheng Road, Yangling Demonstration Zone

Xianyang, Shaanxi, China

Email: [zihao.zhang@nwafu.edu.cn](mailto:zihao.zhang@nwafu.edu.cn)

K. L. Eddie Law

Faculty of Applied Sciences, Macao Polytechnic University

Macao, China

Email: [eddielaw@mpu.edu.mo](mailto:eddielaw@mpu.edu.mo)

# Corresponding Author:

Yan Feng

College of Information Engineering, Northwest A&F University

3 Tai Cheng Road, Yangling Demonstration Zone

Xianyang, Shaanxi, China

Email: [yanfeng52@126.com](mailto:yanfeng52@126.com)

**Running Title:** A Loyalty-SMOTE Algorithm for Imbalanced Data Classification

**Word Count:** 7409

**Table Count:** 22

**Figure Count:** 11

**Data Availability:** Data is openly available in a public repository.

**Funding:** This study was funded by the Key Research and Development Program of Shaanxi(2020NY-129) of the People’s Republic of China.

**Acknowledgments:** We sincerely appreciate the contributors of the UCI and Kaggle databases for providing the data source for the experiments in this paper.

**Conflicts of Interest:** The authors declare no conflict of interest.