IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 31 October 2022, accepted 15 November 2022, date of publication 28 November 2022, date of current version 6 January 2023.

Digital Object Identifier 10.1109/ACCESS.2022.3225404

TOPICAL REVIEW

Imbalanced Classification Methods for Student Grade Prediction: A Systematic Literature Review

SITI DIANAH ABDUL BUJANG[1,2], ALI SELAMAT[1,3,4], (Member, IEEE), ONDREJ KREJCAR[1,4], FARHAN MOHAMED[3], (Senior Member, IEEE), LIM KOK CHENG[5], (Member, IEEE), PO CHAN CHIU[6], AND HAMIDO FUJITA[4,7,8,9], (Life Senior Member, IEEE)

[1]Malaysia Japan International Institute of Technology (MJIIT), Universiti Teknologi Malaysia Kuala Lumpur, Kuala Lumpur 54100, Malaysia
[2]Faculty of Information and Communication Technology, Politeknik Sultan Idris Shah, Sungai Ayer Tawar, Selangor 45100, Malaysia
[3]School of Computing, Faculty of Engineering, Malaysia & Media and Games Center of Excellence (MagicX), Universiti Teknologi Malaysia, Johor Bahru 81310, Malaysia
[4]Faculty of Informatics and Management, University of Hradec Kralove, 50003 Hradec Kralove, Czech Republic
[5]College of Computing and Informatics, Universiti Tenaga Nasional Malaysia, Kajang, Selangor 43000, Malaysia
[6]Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, Kota Samarahan, Sarawak 94300, Malaysia
[7]Faculty of Software and Information Science, Iwate Prefectural University, Takizawa, Iwate 020-0693, Japan
[8]Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, 18011 Granada, Spain
[9]i-SOMET Incorporated Association, Morioka 020-0104, Japan

Corresponding authors: Ali Selamat (aselamat@utm.my), Siti Dianah Abdul Bujang (sdianah84@gmail.com), and Ondrej Krejcar (Ondrej_krejcar@uhk.cz)

This work was supported in part by the Fundamental Research Grant Scheme (FRGS) through the Ministry of Education under Grant FRGS/1/2022/ICT08/UTM/01/1; and in part by the Faculty of Informatics and Management, University of Hradec Kralove, through the Specific Research Project (SPEV) under Grant 2102/2022.

ABSTRACT Student success is essential for improving the higher education system student outcome. One way to measure student success is by predicting students' performance based on their prior academic grades. Concerning the significance of this area, various predictive models are widely developed and applied to help the institution identify students at risk of failure. However, building a high-accuracy predictive model is challenging due to the dataset's imbalanced nature, which caused biased results. Therefore, this study aims to review the existing research article by providing a state-of-the-art approach for handling imbalanced classification in higher education, including the best practices of dataset characteristics, methods, and comparative analysis of the proposed algorithms, focusing on student grade prediction context problems. The study also presents the most common balancing methods published from 2015 to 2021 and highlights their impact on resolving imbalanced classification in three approaches: data-level, algorithm-level, and hybrid-level. The survey results reveal that the data-level approach using SMOTE oversampling is broadly applied in determining imbalanced problems for student grade prediction. However, the application of hybrid and feature selection methods supporting the generalization of the predictive model to boost student grade prediction performance is generally lacking. Other than that, some of the strengths and weaknesses of the proposed methods are discussed and summarized for the direction of future research. The outcomes of this review will guide the professionals, practitioners, and academic researchers in dealing with imbalanced classification, mainly in the higher education field.

INDEX TERMS Imbalanced classification, prediction model, machine learning, student grade prediction, education, systematic literature review.

I. INTRODUCTION

Student grade prediction is one of the essential areas that can determine and monitor student performance in higher educational institutions (HEI). This area has gained significant attention in the education sector over the years as many studies have been interested and proven the reliability of student grade prediction with many help of the existing machine learning algorithms to enhance student success [1], [2], [3].

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Liu.

The aim is to facilitate the educational sector to evaluate the risk of academic failure and provide feedback to improve student outcomes for each semester. With early grade prediction, the development and progress of students can be assessed more effectively [4], [5]. Other than that, student grade prediction is also one of the common indicators to determine student performance success [6]. The high accuracy of students' grade performances is beneficial and helps the HEI identify the students at risk of failure early in academics. However, as the student dataset becomes more extensive and complex, the effects of imbalance distribution on the target class become higher, which results in poor performance on the predictive model [7]. Therefore, knowing the imbalanced classification methods is significant for building an effective predictive model to improve students' future teaching and learning performance.

There have been various published reviews regarding predicting student academic performances and their relevance, as presented in TABLE 1. Nevertheless, based on the results, few studies highlight the algorithm involved in boosting the predictive model accuracy in predicting student performances. Most of the existing surveys were focused on the machine learning methods and summarized the prominent findings with interesting future directions on student performances, but the review of the algorithm resolving the imbalanced dataset was not discussed comprehensively enough. In this paper, the useful methods to resolve the imbalanced classification to improve the efficiency of the predictive model will be discussed in more detail in three different approaches; data-level, algorithm-level, and hybrid-level.

Therefore, this paper thoroughly reviews and summarizes the most common methods for addressing imbalanced classification in the education domain, focusing on improving the performance of student grade prediction. The contributions of this comprehensive review are summarized and highlighted as follows:

1. This SLR analyzes and summarizes the imbalanced classification methods in detail from three different approaches, data-level, algorithm-level, and hybrid-level, to improve the accuracy of predictive models.
2. Provide a taxonomy of current imbalanced classification methods used for predicting student grades to highlight the most applied algorithms in the education field that will ease the professionals, practitioners, and academic researchers to understand the significance of this technique.
3. A comparative study of existing balancing methods with their classifiers in both aspects (binary and multi-class) and accuracy scores more comprehensively that can be used for future educational research.
4. Provide an overview of the existing evaluation performance metrics applied for an imbalanced classification problem to improve predictive model performances in student grade prediction.

The rest of the paper is organized as follows. Section II gives background information to this research for the reader's basic understanding. Section III describes the review method and how the SLR was conducted to formulate the selected articles' research questions and search strategy. Section IV provides data extraction and synthesis of SLR results. Section V discusses the results of the overall findings. Section VI discusses the future direction of this research, and finally, the study is summarized and concluded in Section VII.

II. IMBALANCED CLASSIFICATION IN STUDENT GRADE PREDICTION

Data-driven in education is a new trend accelerated by global changes lately. The knowledge and insightful information gained from this area provide many advantages that can improve HEI decision-making. To achieve this, educational datasets are collected from various online databases and platforms such as Course Management and Learning Management Systems (LMS) or known as Moodle, Massive Open Online Courses (MOOC), Open Course Ware (OCW), Open Educational Resources (OER), and social media sites such as Twitter, Facebook, YouTube and Personal Learning Environments (PLE) [10].

Student grade prediction uses machine learning to predict the final score to improve student academic performance by the end of the semester [11]. The aim is to help educators determine the potential students at risk of low results and help them overcome their learning difficulties. Hence, identifying the relevant factors, including student background, academic information, environmental factors, test scores, and Grade Point Average (GPA) or Cumulative Grade Point Average (CGPA), are significant in predicting student performance [6]. However, when a tremendous amount of data is collected and analyzed without being classified in a balanced way, it becomes a significant problem for predicting students' grades.

During the training phases of student grade prediction, imbalanced classification appears when there is an unequal distribution of instances within the target class in the training dataset [12], [13]. Most datasets involve binary classification consisting of two target outputs: the "pass" class as the majority and the "fail" class as a minority. In contrast, some of it comprises more than two different classes, known as multi-class classification. When one class significantly outnumbers the class of the other, the training model usually spends more time processing on the majority classes than the minority ones, which could be less informative. Consequently, it usually leads classifiers to become biased and produce high erroneous. Due to this, many empirical studies are interested in exploring various methods to enhance student grade prediction performance [14], [15], [16], [17]. However, the methods and algorithms used in dealing with various class-imbalanced distributions to predict student grades are not being highlighted and are not comprehensive enough.

Several approaches have been proposed to handle class imbalance to improve the prediction model's performance. These approaches can be categorized into three levels of solutions: