Imbalance Data Handling using Neighborhood Cleaning Rule (NCL) Sampling Method for Precision Student Modeling

by Prawidya Destarianto

Submission date: 22-Dec-2020 08:47PM (UTC+0700)
Submission ID: 1480520223
File name: agustianto2019.pdf (389.02K)
Word count: 2601
Character count: 13581

Imbalance Data Handling using Neighborhood Cleaning Rule (NCL) Sampling Method for Precision Student Modeling

Khafidurrohman Agustianto
Jurusan Teknologi Informasi
Politeknik Negeri Jember
Jember, Indonesia
agustianto.khafid@gmail.com

Prawidya Destarianto
Jurusan Teknologi Informasi
State of Polytechnic Jember
Jember, Indonesia
prawidya@polije.ac.id

***Abstract*—Student modeling has an important role in every educational process. In general, educational process begins with student admission process, teaching and learning process, and assessement of the learning outcomes. These sequential processes can be represented as a data or called as the Educational Data (ED). However, in real life, the Educational Data has unbalanced characteristics. To overcome the imbalanced issue, some balancing methods are applied. The balancing process basically divided into three methods: undersampling, oversampling and hybrid of oversampling and undersampling. In this paper, we focus on balancing the Educational Data using the undersampling approach Neighborhood Cleaning Rule (NCL) to obtain the Precision Student Modeling. Data that has been undersampled using NCL is then classified pusing the Decision Tree C4.5 algorithm. While, the performance evaluation is processed pusing the accuracy calculations. The test result using NCL shows an accuracy value of 91.37%. The value of accuracy from the research is represented of the student who fail and succeed academically, so that appropriate treatment can be given. This accounting value obtains the standard error in the educational application (10%).**

Keywords—student modeling, educational data, NCL, imbalance data

I. INTRODUCTION

Student modeling has an important part of the education process. The Educational process is started by student admission process, teaching and learning process, and assessement of the learning outcomes. The student modeling can be used as a reference or analysis material to determine the success of the academic process of students and institutions.

The description of students can be used by the lecturers to provide the appropriate treatment, so that, things that are not relevant with the goals of education can be avoided. The learning objectives at the university are seen from the level of student graduation, students who have completed all credits reflecting Learning Outcomes (LO), the students have fulfilled the objectives of education at a university. However, in the process there are various obstacles so that students cannot finish their education at the university on the right time.

Students who graduate more than the time limit are necessary to be handled, because if there are many students who are not able to complete their education on time, the purpose of university education will not be achieved properly. This lack of impact can affect many sectors, such as the labor sector and the economic sector. These diverse parameters have the high probabilitas that imbalance will occur in the data. Usually, in Educational data, this imbalanced phenomenon can be indicated from the number of students who succeed ari greater than the number of students that fail in completing lectures at the University.

The high gap in the number of students who succeeded and fail in completing their education at the University will cause inaccurate classification process. This is related to the formation of a class (academic failure and success) which is very much referring to the Educational Data (ED), so that when the ED is not representative enough to worry about the classification process does not work properly.

ED as a basis for classification needs to be balanced. How to outline data in broad outlines is divided into three parts: undersampling, oversampling and hybrid. The study aims to balance the data using the undersampling approach: Neighborhood Cleaning Rule (NCL) for Precision Student Modeling. Data that has been processed with NCL is then classified. The calcification process in this study uses the Decision Tree C4.5 method. Performance evaluation is processed using accuracy calculations. The test results using NCL show an accuracy value of 91.37%.

The value of accuracy from the research is used as the basis of the implementation of the algorithm in modeling failed academic and successful students, so that appropriate treatment can be given. This accounting value meets the standard error in the educational application (10%).

II. RELATED WORK

Student modeling [1] is one that any part of research on the students activity [2][3], such as Kardan et al. [4] that using the technique of algorithms two levels AL called ACO-Map, in order to obtain the output of concept folders for each group based on their needs.

Jugo et al. [5] using the data mining to make the adaptive learning, this study obtain the recommendation based on

978-1-7281-3436-9/19/$31.00 ©2019 IEEE

patterns derived from the domain of knowledge of students. The application of student modeling is closely related to data usage. The generally primary is used (data directly obtained in the field directly by researchers), in this type of data has a great opportunity to be unbalanced. Thus giving rise to research Ike [6], online class imbalance learning is a new learning problem, this research propose two new ensemble methods that maintain both OOB and UOB with adaptive weights for final predictions, called WEOB1 and WEOB2. They are shown to possess the strength of OOB and UOB with good accuracy and robustness. The application of pre-processing data like this is commonly found to ensure the data used better reflects the research objectives. as in research [7], [8] who uses the undersampling approach, [9] using the hybrid method.

Undersampling as one of the solutions to produce precision student modeling is the basis of this study, so that this study aims do balance the data using the undersampling approach: Neighborhood Cleaning Rule (NCL). Data that has been processed with NCL is then classified. In this study, the classification method that Led is the Decision Tree C4.5 method. Performance evaluation is processed using accuracy calculations.

III. Imbalance Data Handling for Educational Modeling Precision

Educational Data (ED) that has an imbalanced characteristics are needed to be balanced. How to outline data in broad outlines is divided into three parts: undersampling, oversampling and hybrid of oversampling and undersampling. The study aims to balance the data using the undersampling approach: Neighborhood Cleaning Rule (NCL) while undersampling is a method to reduce the amount of dominant data, so that amound data of minor class and dominant class can be balanced.

NCL [9] is an undersampling method to overcome the imbalance class distribution by reducing data based on cleaning. One of the advantages of NCL that it considers the quality of the data to be deleted by not focusing only on data reduction but focusing on cleaning data. The data cleaning process is intended not only for samples in the majority class but also for minority classes. Basically, the principle of NCL is based on the concept of One-Sided Selection (OSS), which is one technique for reducing data based on the instances to reduce classes carefully.

The cleaning data process on NCL [9] is applied to the majority and minority samples separately. NCL adopted the Edited Nearest Neighbor (ENN) method to clear data in the majority class. For example, there is an E1 sample in the training set, then find the three closest neighbors of each sample. If E1 is included as the majority class and the classification result turns out to be the opposite of the original class at E1, then E1 will be deleted. Conversely, if E1 is a minority class and the three neighbors are classified as opposite (majority), then the nearest neighbor will be deleted.

The research method used is shown in Figure 1. Literature Review is used to ensure that the effectiveness of the method that used in research, this stage is then used as State of the Art. State of The Art is intended to ensure the novelty of the research. After finding the novelty, the research continues on the next stage, namely the Design and Implementation of the Method.

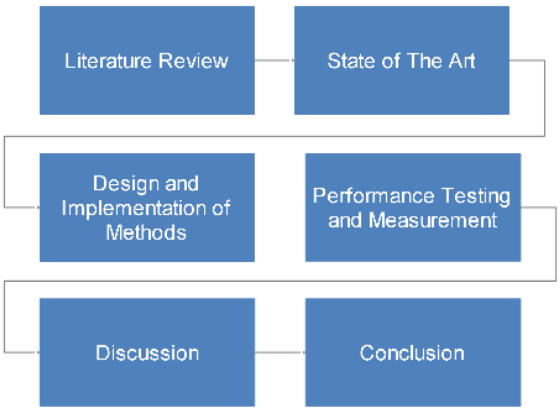

Fig. 1. Research Methods.

Design and Implementation of the method in this study is an important stage. This process consists of the step-by-step testing process until it obtains the right output. The next stage is Testing and Performance Measurement. At this stage, accuracy testing of data that has been processed with NCL will be carried out. The results obtained at the Testing and Measurement stages will then be analyzed in the Discussion phase. The last step carried out in this study was to make a conclusion.

IV. Result and Discussion

Neighborhood Cleaning Rule (NCL) is an undersampling method to overcome imbalance class distribution by reducing the data based on cleaning. In the NCL process, it is assumed that there is a T dataset where C is the class of interest with a small amount of data and O is the majority class obtained from the reduction O = T - C. NCL uses the rules of Edited Nearest Neighbor (ENN) proposed by (More, 2016) to reduce O by removing noise A1 data on O. In addition, ENN deletes the data that has a different class with the majority class (misclassify). Then in the NCL method, the cleaning process is improvised by removing the three closest neighbors from the data on C which are incorrectly classified and still part of O. The three closest neighbors that were deleted are made as sets A2. In this study NCL is used to answer the problem of students who cannot complete college on time need to be to be handled, because if there are many students who cannot complete their education on time, the purpose of university education cannot be achieved properly. This lack of impact in many sectors, starting from the labor sector to the economic sector.

Parameters that are so numerous and concerning education have the opportunity for high imbalance, this can be seen from the number of students who succeed and who do not succeed in completing lectures at the University. The number of students who succeeded in completing far more than those who did not.

The difference in the number of students who succeeded and did not succeed in completing their education at the University resulted in the classification process not being able to run

properly. This is related to the formation of a class (academic failure and success) which is very much referring to the Educational Data (ED), so that when the ED is not representative enough to worry about the classification process does not work properly.

The Educational Data (ED) used for the testing process in the study is shown in Figure 2. The imbalanced data that describes two different classes, the failed student class and the succeed student class. The data is then processed by the NCL method, so that the resulting ED distribution is shown in Figure 3. From the distribution it can be seen that the dominant class has been reduced, so that the data of the two classes becomes more proportional.

Performance evaluation is processed using accuracy calculations in Equation 1.

$$Accuration = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Figure 2 is represented as the plot of original education data in one month (Desember). While Figure 3 is represented the ED plotting after undersampling process using Neighborhood Cleaning Rule (NCL). The plotting figure is divided into three colors: blue is represented as class 0 (majority class), while the orange region is described the minority class (class 1). Then, the removed data of majority instances is shown in green points. As can be seen in Table 1, the original data shows an accuracy as 90.72% wh 8 the test result using NCL shows an accuracy value of 91.37%. The visualization of this result is shown in Figure 3. The results of this good accuracy are used as parameters for the success of NCL in balancing ED, so the classification process can produce good accuracy.

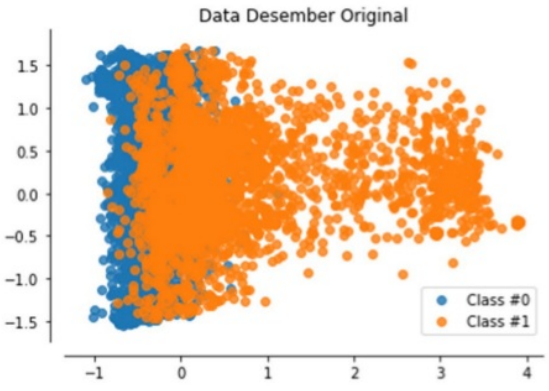

Fig. 2. Plotting of Original Education Data

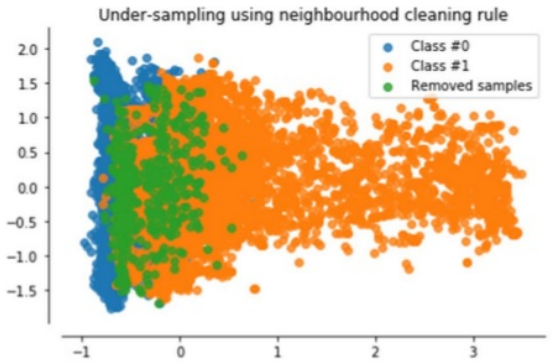

Fig. 3. Plotting of Undersampled Eduction Data

TABLE I. THE ACCURACY OF ORIGINAL AND UNDERSAMPLED DATA

Accuracy of Original Data	Accuracy of Undersampled Data
90.72	91.37

According to Sugiyono [10], research in the field of education has an error tolerance of up to 10%. So, if a product that is able to provide a value of at least 90% accuracy, then the application can be accepted and applied to the education sector.

Based on this results, this implementation are expected to improve quality of description of Jember State Polytechnic student. The description of students can be used by lecturers to provide appropriate treatment, so that things that are not in accordance with the goals of education can be avoided. The learning objectives at the university are seen from the level of student graduation, students who have completed all credits reflecting Learning Outcomes (CP), the students have fulfilled the objectives of education at a university. However, in the process there are various obstacles so that students cannot finish their education at the university on time or even carry the status of Drop Out (DO).

V. CONCLUSION

Performance evaluation is processed using accuracy calculation. NCL test results show an accuracy value of 91.37%. Visualization of these results is shown in Figure 3. The test results show that NCL is suitable for balancing ED, this is indicated by the accuracy value obtained reaching more than 90%.

According to Sugiyono, research in the field of education has an error tolerance of up to 10%. So, if a product that is able to provide a value of at least 90% accuracy, then the application can be accepted and applied to the education sector. The fulfillment of standard errors, it is hoped that this method/ application can truly reach the precision student modeling. The final results of this implementation are expected to improver Precision Student Modeling quality of Jember State Polytechnic.

REFERENCES

[1] Y. Amaya, E. Barrientos, and D. Heredia, "Student Dropout Predictive Model Using Data Mining Techniques," *IEEE Lat. Am. Trans.*, vol. 13, no. 9, pp. 3127–3134, 2015.

[2] E. Gaudioso, M. Montero, and F. Hernandez-del-Olmo, "Supporting teachers in adaptive educational systems through predictive models: A proof of concept," *Expert Syst. Appl.*, vol. 39, no. 1, pp. 621–625, Jan. 2012.

[3] K. Agustianto, A. E. Permanasari, S. S. Kusumawardani, and I. Hidayah, "Design adaptive learning system using metacognitive strategy path for learning in classroom and intelligent tutoring systems," in *AIP Conference Proceedings*, 2016, vol. 1755.

[4] A. Kardan, "A Novel Adaptive Learning Path Method," pp. 20–25, 2013.

[5] I. Jugo, B. Kovačić, and V. Slavuj, "Using Data Mining for Learning Path Recommendation and Visualization in an Intelligent Tutoring System," no. May, pp. 26–30, 2014.

[6] S. Wang, L. L. Minku, and X. Yao, "Resampling-Based Ensemble Methods for Online Class Imbalance Learning," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 5, pp. 1356–1368, 2015.

[7] F. Hamdi, M. Lebbah, and Y. Bennani, "Topographic Under-Sampling for Unbalanced Distributions," *Proc. Int. Jt. Conf. Neural Networks*, 2010.

[8] G. Adluru, L. Chen, and E. V. R. Dibella, "Undersampled Free Breathing Cardiac Perfusion MRI Reconstruction Without Motion Estimation," *Proc. - Int. Symp. Biomed. Imaging*, pp. 97–100, 2011.

[9] S. Choirunnisa, B. Meidyani, and S. Rochimah, "Software Defect Prediction using Oversampling Algorithm: A-SUWO," *2018 Electr. Power, Electron. Commun. Control. Informatics Semin. EECCIS 2018*, pp. 337–341, 2019.

[10] Sugiyono, *Metode Penelitian Kuantitatif, Kualitatif, dan R&D.* Bandung: Alfabeta, 2011.

Imbalance Data Handling using Neighborhood Cleaning Rule (NCL) Sampling Method for Precision Student Modeling

ORIGINALITY REPORT

10%	7%	7%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	aip.scitation.org Internet Source	3%
2	www.epics-project.eu Internet Source	2%
3	Nugroho Setyo Wibowo, Prawidya Destarianto, Hendra Yufit Riskiawan, Khafidurrohman Agustianto, Syamsiar Kautsar. "Development of Low-Cost Autonomous Surface Vehicles (ASV) for Watershed Quality Monitoring", 2018 6th International Conference on Information and Communication Technology (ICoICT), 2018 Publication	1%
4	Submitted to CSU, San Jose State University Student Paper	1%
5	sinta3.ristekdikti.go.id Internet Source	1%
6	Manuel Torres-Vásquez, Oscar Chávez-Bosquez, Betania Hernández-Ocaña, José	1%

Hernández-Torruco. "Classification of Guillain–Barré Syndrome Subtypes Using Sampling Techniques with Binary Approach", Symmetry, 2020
Publication

---

7 Khafidurrohman Agustianto, Prawidya Destarianto, Wahyu Kurnia Dewanto. "Development of real-time motion autonomous surface vehicle controlling for coral reef conservation and fisheries", IOP Conference Series: Earth and Environmental Science, 2020 <1%
Publication

---

8 P Destarianto, B Etikasari, K Agustianto. "Developing Automatic Student Motivation Modeling System", Journal of Physics: Conference Series, 2018 <1%
Publication

---

9 Intelligent Systems Reference Library, 2015. <1%
Publication

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		