ISSN 2278-3091
Volume 8, No.6, November – December 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse63862019.pdf
https://doi.org/10.30534/ijatcse/2019/63862019

A Modified Adaptive Synthetic SMOTE Approach in Graduation Success Rate Classification

Hazel A. Gameng[1], Bobby D. Gerardo[2], Ruji P. Medina[3]
[1]Graduate Programs – Technological Institute of the Philippines - Quezon City, Philippines, hazel.gameng@dnsc.edu.ph
[2]West Visayas State University –Iloilo City, Philippines, bgerardo@wvsu.gov.ph
[3]Graduate Programs – Technological Institute of the Philippines - Quezon City, Philippines, ruji.medina@tip.edu.ph

ABSTRACT

In the real research situation, the oversampling method in data preprocessing is used to solve the problem in imbalanced data. This imbalance may lessen the capability of classification algorithms to identify instances of interest that lead to misclassification such as false positive generation. These imbalanced datasets come from fields of finance, health, education, among other areas. Academic related data such as graduate success rate on higher education are at times imbalanced. One of the established oversampling methods is the Synthetic Minority Oversampling Technique (SMOTE) with Adaptive Synthetic (Adasyn) SMOTE as one of its many variations. K-Nearest Neighbors (KNN) calculations using Euclidean distance is an embedded in Adasyn. In this study, Manhattan distance is utilized in the KNN calculations. The researchers correspondingly gathered actual data from open admission programs of Davao del Norte State College for the training and testing, which consists of 14 features and 897 records. This modified Adasyn was tested on an imbalanced and primary dataset on graduation success rate using logistic regression and random forest as the classification algorithms. This was evaluated in terms of the performance measurements on overall accuracy, precision, recall, and F1 score. Results showed that the modified Adasyn dominated on each performance metrics over SMOTE and Adasyn. Thus, proving that the modified Adasyn is reliable in decreasing misclassification on the graduate success rate dataset.

Key words: Adaptive Synthetic SMOTE, Classification, Graduate Rate, Manhattan, Distance, SMOTE

1. INTRODUCTION

Recent systems produce immense amounts of information in the field of data mining, which compels the advancement of computationally efficient solutions for their processing. These challenges cause difficulties, as this massive information can be affected by class imbalance [1]. Data set imbalance occurs when at least one of the target value classes is underrepresented in comparison with the other classes. Minority-to-majority imbalances can range from 1:10 to 1:1000 or beyond [2].

In the context of learning from imbalanced data, the Synthetic Minority Oversampling Technique (SMOTE) preprocessing algorithm is regarded to be an effective standard due to its procedure design simplicity and its robustness when utilized to various modes of problems. SMOTE algorithm does an oversampling method to balance the initial training set. Rather than simply replicating minority class cases, SMOTE's significant concept is the introduction of synthetic examples in the given region or neighborhood [3] where weighted distribution for diverse minority class instances is based on the learning level difficulty. More synthetic data is produced for minority class instances that are more difficult to learn in contrast to the easy to learn minority instances [4].

One downside of Adasyn is that it expands the positive occurrences region leading to increased false positive proportions that may be crucial to certain class imbalance issues and may lead to lower accuracy and F1 score [5]. In Adasyn, false negative values are lesser and false positive measures are higher. It allows the minority class to be recognized better, but it makes mistakes when identifying the majority group [6]. It can be misleading to calculate a classifier's output applied to imbalanced data using standard metrics on accuracy alone. It is possible to use other measurement metrics [7]. In such situations, threshold metrics such as precision and recall are used to determine a classifier's performance [8]. With these downsides, this study employs Manhattan distance to the modified Adasyn in the generation of synthetic data to increase the accuracy precision, recall and F1 score.

2. RELATED LITERATURE

2.1 Graduation Success Rate

Education itself is turning out to be multifaceted because of the numerous boards of education and divergence in the

curriculum and the teaching-learning progression. His also influences the students when they go for higher education [9].

Dropout in higher education is a worldwide problem in both developed and developing countries. Dropout is described as the interruption of higher education students enrolled for any length of time. This is irrespective of the change in university prior to completion. The reasons have social, academic, and demographic features [10] [11].

College choices are strongly related to the characteristics of students, in particular, their cognitive abilities. Findings suggest that dropouts from college may lack the skills or training needed for college graduation, policies that strengthen student readiness, such as remedial course work may be required [10]. All universities and colleges are concerned with their students' graduation rates and retention. Vast quantities of research focus on identifying significant predictor factors. Mathematical models include regression [12], To predict factors for successful college completion, Bayesian belief networks, discriminant analysis, support vector machines and neural networks, among many others are used [13].

In the Philippines, the K to 12 curriculum of the Department of Education commits to achieve the objectives of the Education for All 2015, which endeavors to prepare senior high school graduates for advance educational potentials to contribute to the country's competitive workforce [14][15].

2.2 Data Imbalanced in the Research Scenario

Data science comprises the preparation, analysis, and processing of both organized and unstructured massive information [16]. Evidence explaining real-world classification difficulties reveals imbalanced distribution in which one of the categories of decisions is underrepresented, often strongly compared to the other class [17]. Larger number of samples from one group would result in a classifier biased to the majority class [18]. The interest in imbalanced classes spans in varied fields such telecommunications, speech recognition, bioinformatics, satellite image recognition of oil spills, and many other areas [19]. Despite countless years of research, gaining from imbalanced remains a challenge in the framework of computational data. Pre-processing calculations are considered among the most effective methods committed to alleviating this concern [20]. The detection of data mining occurrences is a question of forecasting or classifying data. Rare occurrences are difficult to identify because of infrequency. Misclassification of rare occurrences can precede high costs in financial fraud recognition cases. Detection undertaking of rare events occurrences weakens imbalanced data classification [21].

2.3 Synthetic Minority Oversampling Technique

Pre-processing methods include random undersampling and oversampling. The considerable disadvantage of undersampling is that it can remove probable useful information that may be vital to the learning course. Oversampling may enhance the likelihood of overfitting as it makes accurate copies of current occurrences [22]. More cutting-edge methods have been developed including the Synthetic Minority Oversampling Technique (SMOTE) [23][24]. SMOTE preprocessing method is a forerunner in the research community's imbalanced classification. Since its launch, several variants have been developed to improve its effectiveness in different situations. It is also considered the most important algorithm for pre-processing data in data mining and machine learning [25]. SMOTE is also used in many applications. Nonetheless, SMOTE's drawback includes variation and generalization. To address these, SMOTE is merged with over-sampling and synthetically produces target class data [18].

2.4 Adaptive Synthetic (Adasyn) Sampling Technique

The algorithm of Adasyn follows the procedure, as shown in Figure 1 [4][26]. Adasyn's central idea is to employ a weighted distribution of various samples of minority class based on the learning difficulty level. More synthetic data is generated for samples of minority classes that are harder to learn contrasted to minority samples, which are simpler to learn. As a result, the Adasyn approach strengthens learning about data distribution in two modes: (1) reducing the bias set off by imbalanced of the class (2) adjusting the classification decision margin to the difficult cases [4][27]. In generating samples from the synthetic minority class, synthetic oversampling such as Adasyn techniques can improve classifiers' efficiency [28].

2.5 Evaluation Metrics on Imbalanced Data

Overall accuracy is inappropriate in imbalanced datasets as it considers the cost of misclassification and is skewed towards the dominant class and highly sensitive to class skews. Alternative measures such as the true positive rate or sensitivity and the true negative rate or specificity are metrics that independently measure the quality of identification on the majority and minority groups [29]. Researchers use other metrics to test the efficacy of imbalanced data classification solutions such as accuracy, F-measure, Geometric Mean (G-mean) and Area Under Curve (AUC) [30] because the overall accuracy of an imbalanced dataset is dominated by the majority group. Precision, recall, F1 measure, and G-mean Measures used to test learning from imbalanced datasets [31].

Adaptive Synthetic Algorithm

1. Compute the class imbalance degree where ms and ml are the numbers of minority and majority class examples, respectively. If d is lower than a certain threshold, initialize the algorithm.
$$d = m_s / m_l$$
2. Calculate the total number of synthetic minority data to generate for the minority class. β is the constraint applied to identify the balanced level desired after Adasyn.
$$G = (m_l - m_s)\beta$$
3. Obtain the k-Nearest neighbors of each minority instance and compute the r_i value. The r_i value indicates the dominance of the majority class in each specific neighborhood.
$$r_i = \#majority/k$$
4. Normalize the r_i values so that the sum of all r_i values equals to 1.
$$\hat{r}_i = r_i / \sum r_i \qquad \sum \hat{r}_i = 1$$
5. Calculate the amount of synthetic examples to generate per each minority example.
$$G_i = G\hat{r}_i$$
6. Produce G_i data per neighborhood. The minority samples for the neighborhood, x_i.is taken first Next, randomly select a different minority sample in that neighborhood, xz_i. λ is a random number in between 0–1. s_i is the new synthetic sample, x_i and xz_i are two minority examples in same neighborhood.
$$s_i = x_i + (x_{zi} - x_i)\lambda$$

Figure 1: Adaptive Synthetic Algorithm Procedure

3. SIMULATION ANALYSIS AND DISCUSSION

Figure 2 presents the simulation steps carried out in this study for the analysis. The graduation success rate imbalanced dataset was partitioned to 80% training set and 20% test set. This primary dataset consists of 897 student records of three open admission programs of a state college. This dataset has 14 variables such as sex, program, OLSAT [32] entrance exam test results of the verbal, nonverbal, and overall result and eight courses' ratings during the first semester of stay in the program and in the college. It also includes the status of whether they finish their respective programs on the prescribed duration.

SMOTE, Adasyn and modified Adasyn are the three oversampling SMOTE based methods used for the preprocessing. The default Adasyn uses the Euclidean distance while the modified Adasyn used the Manhattan distance [33] in the K-nearest neighbors' with distance computations shown in Figure 3. Each method output was then classified using logistic regression and random forest to obtain the confusion matrix shown in Figure 4. Figure 5 shows the equations to calculate the overall accuracy, precision, recall and F- measure or F1-score. All coding requirements are done in Jupyter notebook in Python using the sklearn and imblearn packages.

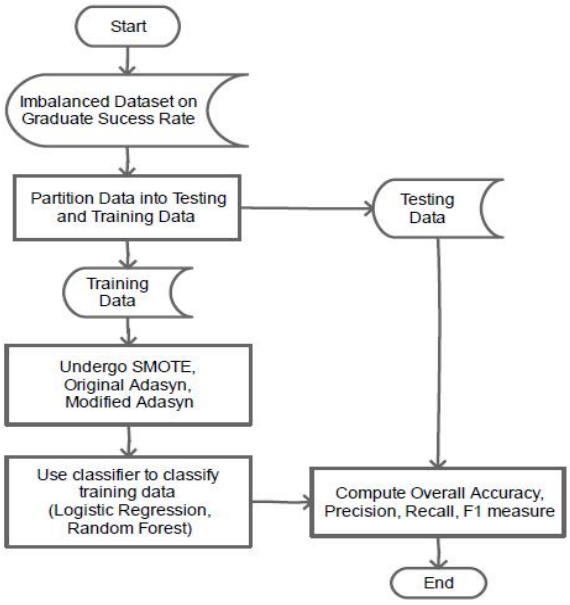

Figure 2: Simulation Steps Used in this Study

$$EuclideanDistance = sqrt(sum((x-y)\text{^}2)$$
$$Manhattan\,Distance = sum(|x-v|)$$

Figure 3: Scikit Learn Distance Metric Equations

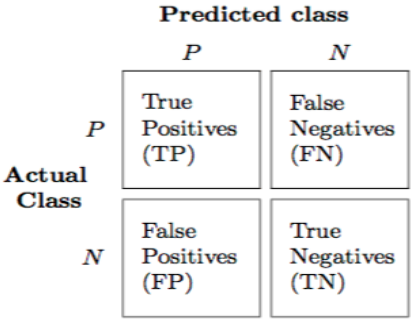

Figure 4: Confusion Matrix

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

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

$$F - Measure = \frac{2 * Recall * Precision}{Recall + Precision}$$

Figure 5: Performance Metrics Formula

4. EXPERIMENTAL RESULTS

The outcome of the performance metrics evaluation using the graduation success rate, three oversampling methods, two classifiers, and four performance metrics used in this study is shown in Table 1. Results show that the modified Adasyn dominated the outcomes on the four performance metrics using the two classifiers. Table 2 shows the percentage lead of the modified Adasyn over SMOTE and Adasyn on the four metrics. The highest lead is on recall metric over SMOTE with 3.283% and 18.326% for logistic regression and random forest, respectively. For this dataset, the classifying performance of random forest preceded over logistic regression. Thus showing that the modification done on the Adasyn proved to be reliable in the preprocessing.

Table 1: Evaluation Metrics Performance Comparison

Method	Classifier	Accuracy	Precision	Recall	F1 Score
SMOTE	Logistic Regression	73.016%	71.014%	77.778%	74.242%
	Random Forest	83.333%	85.593%	80.159%	82.787%
Adasyn	Logistic Regression	71.318%	69.333%	78.788%	73.759%
	Random Forest	82.171%	83.594%	81.061%	82.308%
Modified Adasyn	Logistic Regression	73.643%	71.333%	81.061%	75.887%
	Random Forest	99.225%	100.000%	98.485%	99.237%

Table 2: Percentage Lead of the Modified Adasyn Over SMOTE and Adasyn

Classifier	Method	Accuracy	Precision	Recall	F1 Score
Logistic Regression	SMOTE	0.627%	0.319%	3.283%	1.645%
	Adasyn	2.325%	2.000%	2.273%	2.128%
Random Forest	SMOTE	15.892%	14.407%	18.326%	16.450%
	Adasyn	17.054%	16.406%	17.424%	16.929%

5. CONCLUSION AND FUTURE WORK

In this study, the modified Adasyn utilizing Manhattan distance is fused in the embedded KNN for the generation of synthetic data. Outcomes of the metrics calculations for overall accuracy, precision, recall and F1 score for SMOTE, the original Adasyn, and the modified Adasyn oversampling methods on the graduation success rate primary dataset show that the modified Adasyn lead on the four performance metrics when subjected to the classifiers logistic regression and random forest. It shows better performance in random forest.

.

Imbalanced data learning is a challenge in the data mining field. In future studies, this modified Adasyn can be used to assess other binary classification data sets to assess its reliability further. Further study can be done in application to multinomial classification of this modified Adasyn.

ACKNOWLEDGEMENT

The authors are grateful to the reviewers for the positive comments and recommendations. This work is supported by the Philippine government through the Commission on Higher Education. The authors are thankful for the financial assistance provided for in this study.

REFERENCES

[1] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Prog. Artif. Intell.*, vol. 5, no. 4, pp. 221–232, 2016. https://doi.org/10.1007/s13748-016-0094-0

[2] S. J. Dattagupta, "A Performance Comparison of Oversampling Methods for Data Generation in Imbalanced Learning Tasks," Universidade Nova de Lisboa, 2017.

[3] Alberto Fernandez, Salvador Garcia, Francisco Herrera, and Nitesh V. Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary," *J. Artif. Intell. Res.*, vol. 61, pp. 863–905, 2018. https://doi.org/10.1613/jair.1.11192

[4] S. He, H., Bai, Y., Garcia, E., & Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In IEEE International Joint Conference on Neural Networks, 2008," *IJCNN 2008.(IEEE World Congr. Comput. Intell. (pp. 1322–1328)*, no. 3, pp. 1322– 1328, 2008.

[5] W. Siriseriwan and K. Sinapiromsaran, "Adaptive neighbor synthetic minority oversampling technique under 1NN outcast handling," vol. 39, no. 5, pp. 565–576, 2017.

[6] J. Ah-pine, J. Ah-pine, E. S. A. Study, and S. Oversampling, "A Study of Synthetic Oversampling for Twitter Imbalanced Sentiment Analysis To cite this version : HAL Id : hal-01504684," 2017.

[7] J. S. Akosa, "Predictive Accuracy : A Misleading Performance Measure for Highly Imbalanced Data,"

SAS Glob. Forum, vol. 942, pp. 1–12, 2017.

[8] A. Tharwat, "Classification assessment methods," *Appl. Comput. Informatics*, 2018. https://doi.org/10.1016/j.aci.2018.08.003

[9] R. Suganthan, C; Raja, "Differences in the Board of Education and its Impact in the Writings of Engineering Graduates," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 4, no. 2, pp. 2677–2679, 2015.

[10] L. Paura *et al.*, "Cause Analysis of Students' Dropout Rate in Higher Education Study Program," *Procedia - Soc. Behav. Sci.*, vol. 164, no. March 2017, pp. 1021–1030, 2017.

[11] A. Viloria, J. G. Padilla, C. Vargas-Mercado, H. Hernández-Palma, N. O. Llinas, and M. A. David, "Integration of Data Technology for Analyzing University Dropout," *Procedia Comput. Sci.*, vol. 155, no. 2018, pp. 569–574, 2019. https://doi.org/10.1016/j.procs.2019.08.079

[12] G. Lesinski, S. Corns, and C. Dagli, "Application of an Artificial Neural Network to Predict Graduation Success at the United States Military Academy," *Procedia Comput. Sci.*, vol. 95, pp. 375–382, 2016. https://doi.org/10.1016/j.procs.2016.09.348

[13] G. Lesinski and S. Corns, "Multi-objective evolutionary neural network to predict graduation success at the United States Military Academy," *Procedia Comput. Sci.*, vol. 140, pp. 196–205, 2018. https://doi.org/10.1016/j.procs.2018.10.329

[14] SEAMEO, *K-12 Toolkit*. 2012.

[15] R. Natividad, Marvee Cheska; Gerado, Bobby; Medina, "A Career Track Recommender System for Senior High School Students using Fuzzy Logic," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, no. 1, pp. 2512–2519, 2019. https://doi.org/10.30534/ijatcse/2019/97852019

[16] B. Manoj, K. V. K. Sasikanth, M. V. Subbarao, and V. Jyothi Prakash, "Analysis of data science with the use of big data," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 7, no. 6, pp. 87–90, 2018. https://doi.org/10.30534/ijatcse/2018/02762018

[17] S. Wojciechowski and S. Wilk, "Difficulty Factors and Preprocessing in Imbalanced Data Sets: An Experimental Study on Artificial Data," *Found. Comput. Decis. Sci.*, vol. 42, no. 2, pp. 149–176, 2017.

[18] C. De Souza, "Classification of Imbalanced Classes," vol. 4, no. 2, pp. 215–218, 2016.

[19] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, "SMOTEBoost: Improving Prediction of the Minority Class in Boosting," pp. 107–119, 2010.

[20] P. Skryjomski and B. Krawczyk, "Influence of minority class instance types on SMOTE imbalanced data oversampling," *Proc. Mach. Learn. Res.*, vol. 74, pp. 7–21, 2017.

[21] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, and H. Yuanyue, "Learning from class-imbalanced data : Review of methods and applications," *Expert Syst. Appl.*, vol. 73, pp. 220–239, 2017. https://doi.org/10.1016/j.eswa.2016.12.035

[22] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Inf. Sci. (Ny).*, vol. 250, pp. 113–141, 2013.

[23] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[24] H. Lee, J. Kim, and S. Kim, "Gaussian-based SMOTE algorithm for solving skewed class distributions," *Int. J. Fuzzy Log. Intell. Syst.*, vol. 17, no. 4, pp. 229–234, 2017. https://doi.org/10.5391/IJFIS.2017.17.4.229

[25] S. García, J. Luengo, and F. Herrera, "Tutorial on practical tips of the most influential data preprocessing algorithms in data mining," *Knowledge-Based Syst.*, vol. 98, pp. 1–29, 2016.

[26] R. Nian, "Fixing Imbalanced Datasets_ An Introduction to ADASYN (with code!)," 2018. [Online]. Available: https://medium.com/@ruinian/an-introduction-to-adasyn-with-code-1383a5ece7aa.

[27] A. Amin *et al.*, "Comparing Oversampling Techniques to Handle the Class Imbalance Problem : A Customer Churn Prediction Case Study," vol. 4, no. Ml, 2016.

[28] S. Barua, M. M. Islam, X. Yao, and K. Murase, "MWMOTE - Majority weighted minority oversampling technique for imbalanced data set learning," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 2, pp. 405–425, 2014. https://doi.org/10.1109/TKDE.2012.232

[29] V. García, J. S. Sánchez, R. Martín-Félez, and R. A. Mollineda, "Surrounding neighborhood-based SMOTE for learning from imbalanced data sets," *Prog. Artif. Intell.*, vol. 1, no. 4, pp. 347–362, 2012.

[30] H. Zhang and M. Li, "RWO-Sampling : A random walk over-sampling approach to imbalanced data classification," *Inf. Fusion*, vol. 20, pp. 99–116, 2014.

[31] G. Hoang, A. Bouzerdoum, and S. Lam, "Learning Pattern Classification Tasks with Imbalanced Data Sets," *Pattern Recognit.*, pp. 193–208, 2009. https://doi.org/10.5772/7544

[32] TestPrep, "What is The OLSAT Test_ Learn About The OLSAT 8 - TestPrep-Online." [Online]. Available: https://www.testprep-online.com/what-is-olsat.

[33] Sklearn, "SciKit Learn," *scikit-learn developers (BSD License)*, 2019. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.DistanceMetric.html?fbclid=IwAR3AagWXcYAUwTDcEO4tH5K1Uvksm2lq66yBbyi1oi0Yarg223qg_My-d5g. [Accessed: 20-Mar-2019].