ISSN 2278-3091
Volume 9, No.4, July – August 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse127942020.pdf
https://doi.org/10.30534/ijatcse/2020/127942020

Affinity Propagation SMOTE approach for Imbalanced dataset used in Predicting Student at Risk of Low Performance

Lanie B. Laureano[1], Ariel M. Sison[2], Ruji P. Medina[3]
[1]Graduate Programs-Technological Institute of the Philippines-Quezon City, Philippines, lanie.laureano@doscst.edu.ph
[2]School of Computer Studies- Emilio Aguinaldo College, Manila, Philippines, ariel.sison@eac.edu.ph
[3]Graduate Programs-Technological Institute of the Philippines-Quezon City, Philippines, ruji.medina@tip.edu.ph

ABSTRACT

Imbalanced datasets affect the performance of classification algorithms in predicting student performance. There are several techniques in combatting class imbalance and one of the most known is the Synthetic Minority Oversampling Technique (SMOTE). It is an oversampling technique that generates synthetic data along the line of the minority instances and its neighbors. However, it has a drawback on the distribution and generation of noisy samples which is the main reason of its many variations. In the cluster approaches for SMOTE, Affinity Propagation (AP) SMOTE is one of them. This approach uses affinity propagation to automatically produce clusters and cluster exemplars used to select the clusters to be oversampled. This way, the sparsity and generation of noisy samples will be avoided. The data used for the study is the student performance of freshman students of Davao Oriental State College of Science and Technology (DOSCST) as well as their enrolment data. The dataset comprises 10 features and 2112 instances, the imbalance ratio between majority and minority is 17.85. SMOTE and AP SMOTE are applied to the imbalanced dataset. The output is used in the J48 and Naïve Bayes classifiers to predict the student at risk of getting low performance in their freshman years in the college. The classifiers' performance is evaluated using f-measure, g-mean, and Areas Under the Curve (AUC). Results showed that AP SMOTE outperforms the original SMOTE with a percentage lead of .60%, .88%, 1.2% using the J48 classifier. The percentage lead for Naïve Bayes is 3.2%, 6.58%, 3.30%, respectively.

Key words : affinity propagation**,** EDM, imbalanced data, student at risk

1. INTRODUCTION

Data mining is one of the breakthroughs in discovering knowledge from the analysis of large volumes of dataset. The patterns, rules, predictions, and associations are the knowledge that may be discovered through this approach [1], [2]. It is applied in different areas like business, medicine, agriculture, engineering, economy, and even in education [3]. Data mining in education is known as Educational Data Mining (EDM), which aims to improve quality education by exploring educational datasets [4]. In the application of prediction in educational datasets, it is an assumption that there is an equal distribution of instances of each class. However, in the real scenario, imbalance exists in educational datasets. This happens when one of the classes is underrepresented in the whole dataset. The imbalance can cause misclassification of instances in the prediction phase, thus, balancing the dataset is essential.

There are several approaches in handling imbalanced data, including data level (preprocessing) and algorithm-based. One of the most known preprocessing methods is the Synthetic Minority Over Sampling Technique (SMOTE), which generates synthetic samples in between minority samples and its neighbors [5]. However, SMOTE has a downside on creating noisy samples because it oversamples all the minority instances regardless of position in the feature space. Thus, modification and extensions have been introduced including Affinity Propagation (AP) SMOTE [6].

AP SMOTE is an oversampling method that uses affinity propagation to automatically cluster the datasets and the clusters produced are filtered to determine which clusters are to be oversampled. The sparsity and density are used to calculate the number of samples to be created per filtered cluster. The said algorithm was used in classification using

UCI datasets. The application of affinity propagation SMOTE in the context of predicting the student at risk of getting low performance is the focus of this study.

2. RELATED LITERATURE

2.1 Predicting Student Performance

There are already several studies that addressed the prediction of student performance particularly on the identification of vulnerable students who are at risk of getting low performance or dropping out. Early detection may reduce the risk by undertaking timely actions and adopting pro-active measures [7].

Kovacic [8] tried to predict student outcome for newly enrolled students using only the enrolment data available in the enrolment form. The classification of students based on pre-enrollment information would allow the institution to identify students who would be at risk of dropping the course so that they could be given support systems such as orientation, advising, and mentoring.

In [9], Learning Management System (LMS) data are used to identify students at risk in the first week of the course to give proper strategies for motivation.

Data mining techniques are also used to predict students' outcomes based on early module performance and other student characteristics in [10]. Identifying indicators for better prediction was also made to determine which student characteristics are the best predictors.

2.2 Imbalanced Datasets in Educational Data Mining (EDM)

Imbalanced datasets that can affect the result of the classification process are usually overlooked in EDM [3]. In classification, it is assumed that the distribution of data is balanced among majority and minority classes. However, this is not the real scenario in the field of education. The imbalanced occurs when a class is underrepresented compared to other classes in educational data [11]. The classifier could be biased to the majority class; thus, handling the imbalance is done.

Prior studies have shown techniques on handling imbalanced datasets in EDM. For example, in [3], oversampling (SMOTE) and undersampling (OSS) approaches are used to balance a dataset distribution. In [12], the imbalanced educational dataset is handled using a hybrid resampling method in Weka. The hybrid technique combines the use of oversampling and undersampling to ensure that balanced will be achieved between the majority and minority class. The approaches done improves the result of the classification of the student performance classification.

2.3 Affinity Propagation SMOTE for Imbalanced Datasets

Affinity Propagation SMOTE is a modification of K-Means SMOTE [13] that uses affinity propagation to cluster the majority and minority instances and automatically generate the number of clusters and exemplars [6]. It has three main steps: First is clustering using affinity propagation which generates the clusters and cluster exemplars, second is to filter the cluster by choosing the clusters with 50% of minority instances for the generation of synthetic data and the last one is the oversampling step which uses the exemplar as the basis of synthetic data generation. The oversampling is still based on SMOTE which generates samples among the neighborhood [5]. The said approach was used in the UCI datasets which are medical-related.

3. METHODS

This study is geared to predict the students at risk of getting low performance in their freshman years based on enrolment data. Figure 1 shows the research process followed in this study. Available data from a public institution is used and processes such as preprocessing, classification, and evaluation were done. The details of each process are discussed in the sections below.

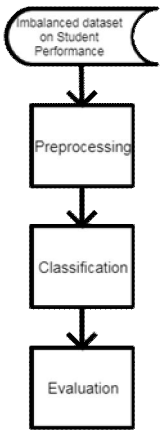

Figure 1: Research Process

3.1 Preprocessing

The imbalanced dataset used for this study is the student performance of 2112 freshman students of a Davao Oriental State College of Science and Technology (DOSCST) and their respective admission data. Admission data includes high school general average and State College Aptitude and Scholarship Test (SCAST) results. The details of the variables are in Table 1, and the sample dataset is shown in Figure 2.

Table 1: Variables of Student Performance dataset

Variable Code	Variable	Description
V1	HS – GA	Student's Highschool general average
V2	Vocabulary + Computation	Student's score for SCAST's vocabulary and computation
V3	Computation	Student's score for SCAST's computation
V4	Spatial Relations	Student's score for SCAST's spatial relations
V6	Word Comparison	Student's score for SCAST's word comparison
V7	Making Marks	Student's score for SCAST's making marks
V8	Raw Score Total	Students Raw Score total for all sections
V9	GWA	Student's general weighted average for the 1st semester of stay in the college.
V10	Class	The remarks for student's PERFORMANCE which is either Good Performance (GP) or Low Performance (LP) based on the GWA.

Relation: for visualization

No.	1: V1	2: V2	3: V3	4: V4	5: V5	6: V6	7: V7	8: V8	9: V9	10: Class
	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Nominal
1	87.0	16.0	8.0	8.0	5.0	48.0	62.0	147.0	2.5	GP
2	86.0	15.0	6.0	9.0	5.0	19.0	54.0	108.0	2.29...	GP
3	89.0	21.0	13.0	8.0	15.0	32.0	29.0	118.0	2.70...	GP
4	85.0	28.0	19.0	9.0	5.0	40.0	57.0	158.0	2.25	GP
5	84.0	18.0	9.0	9.0	6.0	31.0	41.0	114.0	2.375	GP
6	84.0	12.0	8.0	4.0	10.0	19.0	60.0	113.0	2.875	GP
7	84.0	24.0	17.0	7.0	13.0	35.0	50.0	146.0	2.5	GP
8	93.0	21.0	13.0	8.0	13.0	27.0	53.0	135.0	2.54...	GP
9	91.0	25.0	12.0	13.0	14.0	28.0	48.0	140.0	2.29...	GP
10	85.0	22.0	14.0	8.0	12.0	37.0	75.0	168.0	2.45...	GP
11	83.0	17.0	9.0	8.0	5.0	24.0	44.0	107.0	3.16...	LP
12	82.0	14.0	7.0	7.0	11.0	25.0	35.0	99.0	2.625	GP
13	87.0	19.0	18.0	1.0	10.0	45.0	82.0	175.0	2.25	GP
14	85.0	19.0	10.0	9.0	11.0	16.0	34.0	99.0	5.0	LP
15	90.0	21.0	15.0	6.0	11.0	38.0	53.0	144.0	2.25	GP
16	81.0	24.0	16.0	8.0	10.0	24.0	58.0	140.0	2.375	GP
17	84.0	23.0	12.0	11.0	8.0	20.0	38.0	112.0	2.625	GP
18	84.0	31.0	18.0	13.0	10.0	42.0	48.0	162.0	2.58...	GP
19	84.0	25.0	20.0	5.0	15.0	38.0	69.0	172.0	2.54...	GP
20	80.0	14.0	9.0	5.0	10.0	29.0	80.0	147.0	2.83...	GP

Figure 2: Sample dataset visualized in Weka

Figure 3 shows the imbalance between the classifications of student performance. The blue bar in the graph represents the majority, the students with good performance and the one in red are those who have low performance based on their GWA. The imbalanced ratio between the majority and minority class is 17.85. The imbalanced dataset was partitioned into 80% training dataset and for testing dataset 20%.

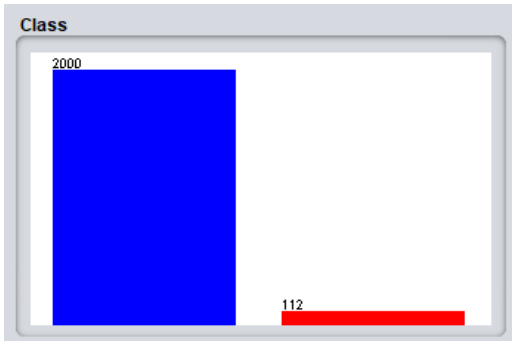

Figure 3: Imbalanced dataset visualization

SMOTE and AP SMOTE are used for the generation of synthetic data for the imbalanced dataset. Figure 4 shows the visualization of a balanced training dataset produced using AP SMOTE visualized through the Weka software.

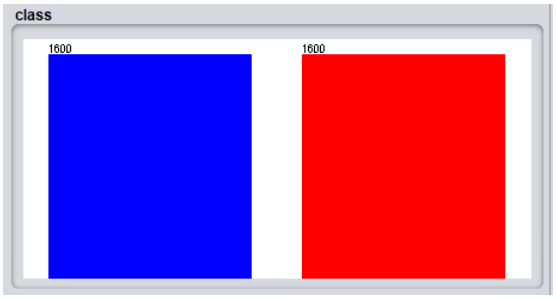

Figure 4: Balanced training dataset using AP SMOTE as an oversampling technique

3.2 Prediction of Student at Risk of Low Performance

After the application of SMOTE and AP SMOTE, the balanced datasets produced are used for the classification. The classifiers utilized are Naïve Bayes and J48 implemented in the WEKA software.

Both Naïve Bayes and J48 algorithms are commonly used in classifying datasets because of their simplicity and efficiency. In [14], the two classifiers are used and both gave good results for a bank dataset. Few examples of utilization of J48 in EDM are analysis of student performance [15] and classification of learning styles [16]. Naïve Bayes, on the other hand, is also used for bachelor's academic performance analysis [17] and classification based on educational qualification [18].

3.3 Evaluation

The binary classification matrix shown in Table 2 is usually used to derive the accuracy and error rate of classification algorithms. However, in the case of imbalanced data, the classifier gets biased towards the majority samples thus increasing the accuracy rate and lowering the error rate. For this reason, recall, precision, f-measure, g-mean, and AUC are utilized for the performance evaluation of classifiers [19].

Table 2: Confusion Matrix

	Predicted Negative	Predicted Positive
Target Negative	TN	FP
Target Positive	FN	TP

The previously mentioned measures are defined through the following equations:

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

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F-measure = \frac{2*Recall*Precision}{Recall+Precision} \quad (4)$$

$$g-mean = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}} \quad (5)$$

$$AUC = \left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) / 2 \quad (6)$$

Precision and recall evaluate the classifiers by concentrating on the minority class. Precision measures the proportion of positive identification that was actually correct, while recall measures the proportion of actual positives which was identified correctly. The harmonic mean of precision and recall is measured through the F-measure or F-1 score. To measure the balanced performance of a classifier, Geometric mean, or g-mean is used. AUC is Area Under the ROC Curve which is used to evaluate the performance of the model. The classifier performs better when the AUC is higher [20], [21].

4. RESULTS AND DISCUSSION

In tables 3 and 4 below, the result of the performance metrics evaluation of two classifiers, J48 and Naïve Bayes are shown. In table 3, results show that AP SMOTE dominated the outcomes of the performance metrics using the two classifiers. The percentage lead of AP SMOTE to SMOTE in terms of accuracy, precision, and recall using the J48 classifier is .56%, .60%, .60%, respectively. While using Naïve Bayes as a classifier, the percentage lead is 3.32%, 4.60%, 3.30%.

On the other hand, table 4 shows the result of classifiers' performance in terms of F-measure, g-mean, and AUC. AP SMOTE outperformed SMOTE with a percentage lead of .60%, .88%, 1.2% for the 3 measures using the J48 Classifier. While using Naïve Bayes, the percentage lead is 3.2%, 6.58%, 3.30%, respectively.

With these results, for this educational dataset, the imbalanced data applied with AP SMOTE has improved the performance of J48 and Naïve Bayes classifiers.

Table 3: Performance Comparison in terms of Accuracy, Precision, and Recall

Methods	Classifier	Accuracy	Precision	Recall
SMOTE	J48	91.59	91.60	91.60
	Naïve Bayes	68.80	70.80	68.80
AP SMOTE	J48	**92.15**	**92.20**	**92.20**
	Naïve Bayes	**72.116**	**75.40**	**72.10**

Table 4: Performance Comparison in terms of F-measure, G-mean, and AUC

Methods	Classifier	F-measure	G-Mean	AUC
SMOTE	J48	91.60	91.82	91.55
	Naïve Bayes	68	80.69	68.80
AP SMOTE	J48	**92.20**	**92.70**	**92.15**
	Naïve Bayes	**71.20**	**87.27**	**72.10**

5. CONCLUSION AND FUTURE WORKS

The main objective of this paper is to properly handle an imbalanced dataset in predicting student performance. AP SMOTE as an oversampling technique, is used for the generation of synthetic data. Based on the performance comparison of classifiers in terms of accuracy, recall, precision, f-measure, g-mean, and AUC, the dataset applied with AP SMOTE shows better performance for the Naïve Bayes and J48 classifiers. Thus, improves the prediction of students at risk of low performance. With early prediction, early interventions can be given by the administration.

Imbalanced learning occurs in different areas in the data mining field and techniques for handling it are still an interest. In future studies, the exploration of using AP SMOTE for multiclass datasets could be done.

ACKNOWLEDGEMENT

The authors would like to thank the proponents of a DOSCST Institutional Study for the dataset provided for this paper. The Philippine government gives this work support through the Commission on Higher Education.

REFERENCES

1. S. Borkar and K. Rajeswari, “Attributes Selection for Predicting Students' Academic Performance using Education Data Mining and Artificial Neural Network,” *Int. J. Comput. Appl.*, vol. 86, no. 10, pp. 25–29, 2014, doi: 10.5120/15022-3310.
2. I. M. Tarun, “Prediction Models for Licensure Examination Performance using Data Mining Classifiers for Online Test and Decision Support System,” *Asia Pacific J. Multidiscip. Res.*, vol. 5, no. 3, pp. 10–21, 2017.
3. Y. Pristyanto, I. Pratama, and A. F. Nugraha, “Data level approach for imbalanced class handling on educational data mining multiclass classification,” in *Proc. 2018 Int. Conf. Inf. Commun. Technol. ICOIACT 2018*, vol. 2018-Janua, pp. 310–314, 2018 doi: 10.1109/ICOIACT.2018.8350792.
4. J. S. Gil, A. J. P. Delima, and R. N. Vilchez, “Predicting students’ dropout indicators in public school using data

mining approaches," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 1, pp. 774–778, 2020.

5. A. Fernández, S. García, F. Herrera, and N. 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 doi: 10.1613/jair.1.11192.
6. L. B. Laureano, A. M. Sison, and R. P. Medina, "Handling imbalanced data through affinity propagation and SMOTE," in *Proc. ACM Int. Conf. Proceeding Ser.*, pp. 22–26, 2019, doi: 10.1145/3366650.3366665.
7. M. A. Timbal, "Analysis of Student-at-Risk of Dropping out (SARDO) Using decision tree: An Intelligent predictive model for reduction," *Int. J. Mach. Learn. Comput.*, vol. 9, no. 3, pp. 273–278, 2019 doi: 10.18178/ijmlc.2019.9.3.798.
8. Z. J. Kovacic, "Early Prediction of Student Success: Mining Students Enrolment Data," *Proc. 2010 InSITE Conf.*, pp. 647–665, 2010, doi: 10.28945/1281.
9. J. D. John Milne, L. M. Jeffrey, G. Suddaby, and A. Higgins, "Early identification of students at risk of failing," *ASCILITE 2012 - Annu. Conf. Aust. Soc. Comput. Tert. Educ.*, 2012.
10. Z. Alharbi, J. Cornford, L. Dolder, and B. De La Iglesia, "Using data mining techniques to predict students at risk of poor performance," in *Proc. 2016 SAI Comput. Conf. SAI 2016*, pp. 523–531, 2016 doi: 10.1109/SAI.2016.7556030.
11. H. A. Gameng, B. D. Gerardo, and R. P. Medina, "A modified adaptive synthetic smote approach in graduation success rate classification," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, no. 6, pp. 3053–3057, 2019, doi: 10.30534/ijatcse/2019/63862019.
12. V. T. N. Chau and N. H. Phung, "Imbalanced educational data classification: An effective approach with resampling and random forest," in *Proc. - 2013 RIVF Int. Conf. Comput. Commun. Technol. Res. Innov. Vis. Futur. RIVF 2013*, no. January, pp. 135–140, 2013, doi: 10.1109/RIVF.2013.6719882.
13. G. Douzas, F. Bacao, and F. Last, "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE," *Inf. Sci. (Ny).*, vol. 465, pp. 1–20, 2018, doi: 10.1016/j.ins.2018.06.056.
14. T. R. Patil and S. Sherekar, "Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification," *Int. J. Intell. Syst. Appl. Eng.*, vol. 6, no. 2, pp. 88–91, 2013, doi: 10.18201/ijisae.2019252786.
15. M. Mythili and A. Shanavas Mohamed, "An Analysis of students' performance using classification algorithms," *IOSR J. Comput. Eng.*, vol. 16, no. 1, pp. 63–69, 2014, doi: 10.9790/0661-16136369.
16. R. R. Maaliw and M. A. Ballera, "Classification of learning styles in virtual learning environment using J48 decision tree," in *Proc 14th Int. Conf. Cogn. Explor. Learn. Digit. Age, CELDA 2017*, no. Celda, pp. 149–156, 2017.
17. F. Razaque *et al.*, "Using naïve bayes algorithm to students' bachelor academic performances analysis," *4th IEEE Int. Conf. Eng. Technol. Appl. Sci. ICETAS 2017*, vol. 2018-January, pp. 1–5, 2018 doi: 10.1109/ICETAS.2017.8277884.
18. S. Karthika and N. Sairam, "A Naïve Bayesian Classifier for Educational Qualification," *Indian J. Sci. Technol.*, vol. 8, no. 16, 2015 doi: 10.17485/ijst/2015/v8i16/62055.
19. H. Guo, J. Zhou, and C. A. Wu, "Imbalanced learning based on data-partition and SMOTE," *Inf.*, vol. 9, no. 9, 2018, doi: 10.3390/info9090238.
20. I. Nekooeimehr and S. K. Lai-Yuen, "Adaptive semi-unsupervised weighted oversampling (A-SUWO) for imbalanced datasets," *Expert Syst. Appl.*, vol. 46, pp. 405–416, 2016, doi: 10.1016/j.eswa.2015.10.031.
21. Y. Ge, D. Yue, and L. Chen, "Prediction of wind turbine blades icing based on MBK-SMOTE and random forest in imbalanced data set," in *Proc. 2017 IEEE Conf. Energy Internet Energy Syst. Integr. EI2 2017 - Proc.*, vol. 2018-Janua, pp. 1–6, 2018, doi: 10.1109/EI2.2017.8245530.