Data Science and Engineering (DSE) Record, Volume 1, issue 2.

A Preliminary Study of Risk Prediction Model Development for Road Traffic Injury in Drunk-Drivers During Festivals in Thailand: An Approach of Imbalanced Public Health Data for Classification Model

Wachiranun Sirikul [1,2] and Trasapong Thaiupathump [3]

[1] Data Science Program, Chiang Mai University, Chiang Mai, Thailand
[2] Community Medicine, Chiang Mai University, Chiang Mai, Thailand
[3] Faculty of Engineering, Chiang Mai University, Chiang Mai, Thailand

wachiranun_sirikul@cmu.ac.th

Abstract. Thailand is a middle-income country where the road traffic injury crisis has been one of the most serious public health concerns. Currently, the machine learning (ML) algorithms are widely used for public health predictive analytics. Therefore, we developed the Multi-layer perceptron (MLP) classifier from the road traffic accident driver data in Thailand that aim to classify a high-risk driver who had severe injuries from road traffic accidents. However, the imbalanced data was a typical problem in public health data and also caused an "*accuracy paradox*" that the model intended to predict a majority class. Accurately detecting minority class was important especially in the public health data because it was associated with high impact events and serious adverse outcomes. Since the imbalanced data is unavoidable according to the nature of public health data. The rebalanced strategies or other data approaches were applied to encounter this problem. Subsequently, the oversampling techniques were significantly improved discrimination performances of models comparing with under-sampling or without rebalancing approach.

Keywords: alcohol-related road traffic injury; classification model; Imbalanced data.

1 Introduction

Public health is the medical discipline involved with the prevention and control of disease through the population and community levels **[1]**. In general, the imbalanced data has commonly appeared in several degrees but was more often observed in health data where the target endpoint or outcomes were almost always a minority part of datasets. Moreover, these minority class are usually associated with high impact events or serious adverse patient outcomes. Classification models using imbalanced data typically

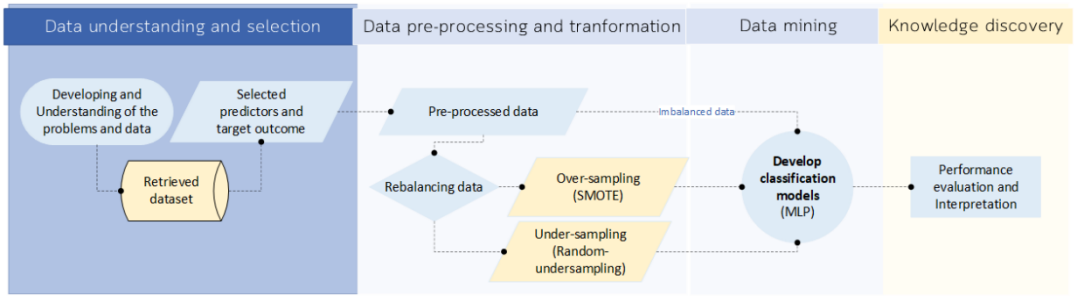
cause an "*accuracy paradox*" in that these classifiers tend to predict the majority class rather than the minority class. Therefore, the effect of imbalanced learning was often observed in the development of prediction model study **[2, 3]**. Consequently, using imbalanced data in statistical models and machine learning raised the concern of the model performances in a real-world implementation. Rebalanced Strategies, which were data level approaches (Oversampling and Under Sampling techniques) and algorithm approaches (Cost-Sensitive Learning) **[2, 4-7]**, were purposed to resolve the imbalanced problems. This study was conducted to demonstrate the effect of imbalanced data in classification model development and the effectiveness of rebalanced strategies.

2 Method

2.1 Data mining process

This study was applied Knowledge Discovery in Databases (KDD) for data mining approaches consist of developing an understanding of the problems and data, selecting the target data set, data cleaning and preprocessing, data transformation, choosing data mining tasks and algorithms to solve your problems, Interpreting the patterns, and consolidating the discovered knowledge. The overview of study process was showed in Figure 1.

Fig. 1. The Flow of Data Mining Processes



SMOTE: Synthetic Minority Oversampling Technique, MLP: Multi-Layer Perceptron

2.2 Derived data, predictors, and target outcome

The road traffic accident victims and inpatients at provincial hospitals were retrieved from Thai Governmental Road Safety Evaluation project in response to road safety planning during 7 dangerous days in 2002–2004. 4875 driver data were used for data exploration and analysis included patient demographics (e.g., age, gender), place of accidents, time of accidents, type of vehicles, helmet used, safety belt used, blood alcohol concentration, and injury outcomes. The target outcome was divided into high-severity injury and low-severity injury. The high-severity injury was the composite

outcome including brain injury, spinal cord/cervical spine injury, cardiac injury, pulmonary injury, and ophthalmic injury.

2.3 Data exploration and analysis

Regards imbalanced data problem, rebalanced strategies were operated using Synthetic Minority Oversampling Technique (SMOTE) and Random Under-sampling function via Sci-Kit Learn packages. The principal component analysis (PCA) was done to decompose the multidimensional data into lower-dimensional spaces and exploring the effect of each rebalanced strategy. The classification model development was done using Multi-Layer Perceptron classifier (MLP) from Sci-Kit Learn packages. All derived classifiers were trained and evaluated by Area under the ROC Curve (AUC), Sensitivity, Specificity, Predictive values, and Likelihood ratio from 10-fold cross-validation.

3 Result

3.1 Variance distribution of the derived datasets

To explore the data distribution, the imbalanced and rebalanced datasets were decomposed into lower-dimensional data by PCA. From Figure 2, the variance distributions of the derived datasets were similar. Moreover, PC1 and PC2 in 3 datasets were explained over 50% of the dataset variance.

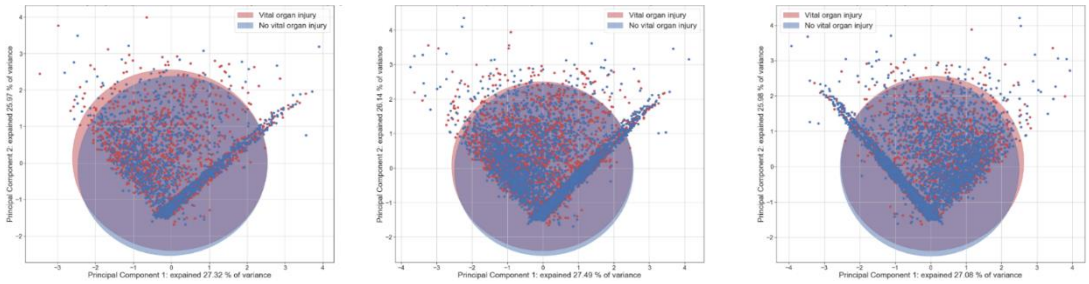
3.2 The discrimination performance of MLP classifiers

the MLP classifier from imbalanced data was performed with poor discrimination performance (AUC 0.61). Although the rebalanced data by the under-sampling approach slightly improved AUC (AUC 0.66) comparing with using imbalanced data. The over-sampling approach using SMOTE (AUC 0.72) was superiorly improved AUC contrasting with the others.

Table 1. The discrimination performance of MLP classifiers

Model	AUC		Likelihood ratio		Sens.	Spec.	Predictive value	
	mean	95% CI	Positive	Negative			Positive	Negative
Imbalanced	0.61	0.54 - 0.69	1.45	0.92	22.67	84.36	39.93	70.41
SMOTE	0.72	0.63 - 0.80	1.93	0.46	70.59	63.49	65.91	88.34
RUS	0.66	0.62 - 0.70	1.51	0.62	64.16	57.58	60.2	61.64

AUC: Area under the receiver operating curve, MLP: Multi-Layer Perceptron, RUS: Random over sampling, SMOTE: Synthetic Minority Oversampling Technique, Sens. : Sensitivity, Spec. : Specificity



2a. Imbalanced dataset 2b. Rebalanced dataset by SMOTE 2c. Rebalanced dataset by Random Under Sampling

Fig. 2. The scatter plot of the first two principal components from PCA analysis

4 Discussion

Imbalanced learning was an important factor that affected the discrimination performance of the models. Even though rebalanced strategies augmented the data using the original dataset, these should not cause a substantial difference in the variance distribution. We confirmed by PCA analysis that the rebalancing methods did not cause a significant alteration of the major components of data and the data distribution. The discrimination performance of the derived models was significantly improved by rebalanced techniques particularly using over-sampling by SMOTE. The models using SMOTE had a significantly improved model sensitivity (true-positive rate) while balanced the specificity (true-negative rate) as a trade-off. Even though the Random Under Sampling method increased the model sensitivity by reducing the specificity of the models, the overall improvement was inferior when comparing to SMOTE. Our results from imbalanced learning models were consistent with the previous study that used imbalanced medical or public health data for developing the prediction model **[4-7]**. As same as our results, the over-sampling by various method (e.g., Random Over Sampling, SMOTE, combined with Bagging method, and ADASYN) were evidently improved the discrimination performance, overwhelmed the under-sampling approaches, especially SMOTE and the over-sampling combined with Bagging method.

5 Conclusion

Our study was demonstrated that imbalanced learning affected the classification model development in public health data. In our study, the over-sampling technique by SMOTE has significantly increased the discrimination model performance consistent with the previous studies. We suggested that rebalanced strategies or other imbalanced data management approaches should be considered in the prediction model development.

References

1. Kindig DA. Understanding population health terminology. Milbank Q. 2007;85(1):139-61.
2. Awan SE, Bennamoun M, Sohel F, Sanfilippo FM, Dwivedi G. Machine learning-based prediction of heart failure readmission or death: implications of choosing the right model and the right metrics. ESC Heart Fail. 2019;6(2):428-35.
3. Belarouci S, Chikh M. Medical imbalanced data classification. Advances in Science, Technology and Engineering Systems Journal. 2017;2:116-24.
4. Zhao Y, Wong ZS-Y, Tsui KL. A Framework of Rebalancing Imbalanced Healthcare Data for Rare Events' Classification: A Case of Look-Alike Sound-Alike Mix-Up Incident Detection. Journal of Healthcare Engineering. 2018;2018:6275435.
5. Goswami T, Roy UB, editors. Classification Accuracy Comparison for Imbalanced Datasets with Its Balanced Counterparts Obtained by Different Sampling Techniques2021; Singapore: Springer Singapore.
6. Beryl Princess PJ, Silas S, Rajsingh EB, editors. Performance Comparison of Machine Learning Models for Classification of Traffic Injury Severity from Imbalanced Accident Dataset2021; Singapore: Springer Singapore.
7. Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews). 2012;42(4):463-84.