10th MALAYSIA STATISTICS CONFERENCE 2023
Looking Beyond GDP: Toward Social Well-being and Environmental Sustainability

26th September 2023
Sasana Kijang, Bank Negara Malaysia

STATISTICAL COMMUNICATION

Diabetes Prediction using Synthetic Oversampling Approaches for Imbalanced Classification

Nuryasmin Wahida Binti Hamil[1,2*]; Adilah Abdul Ghapor[1]; Yong Zulina Zubairi[3]

[1] Department of Decision Science, Faculty of Business and Economics, Universiti Malaya, 50603 Kuala Lumpur, Malaysia
[2] Faculty of Business and Communications, INTI International University,71800 Negeri Sembilan, Malaysia
[3] Institute of Advanced Studies, Universiti Malaya, 50603 Kuala Lumpur, Malaysia

Abstract:

Diabetes has become a global health concern, and accurate prediction of diabetes risk plays a vital role in early intervention and effective management. However, imbalanced datasets, where one class (e.g., diabetes-positive cases) is significantly outnumbered by another (e.g., diabetes-negative cases) also known as the Class Imbalance Problem (CIP), pose a challenge for traditional machine learning models. The CIP triggers the data scientist community for decades as one of the major problems in data mining classification process. This condition leads to degradation of the minority data thus severely affects the prediction accuracy of the data classification process. This problem is often tackled using various oversampling or under sampling techniques. Synthetic Minority Over-sampling Technique (SMOTE) is the pioneer oversampling method in research community for imbalance classification and widely applied to handle the CIP due its superior performance as a more powerful technique. This study applied synthetic oversampling to address this issue. These approaches show an improvement on the classification performance on prediction of diabetes patient. It can be concluded that based on the sensitivity value, ADASYN is the best synthetic OS approach in predicting diabetic cases since sensitivity is looking on the proportion of the total number of correctly classified under minority (positive) class. This study demonstrated the potential to make a meaningful contributionto good health and well-being, which is one of the keys focuses of the SDGs.

Keywords:

Human and health; CIP; SMOTE; ADASYN; Borderline-SMOTE

1. Introduction:

Diabetes mellitus, a chronic metabolic disorder characterized by elevated blood glucose levels, has become a global health concern with a significant impact on public

health systems. Timely and accurate prediction of diabetes onset plays a crucial role in effective management and prevention of complications. Machine learning techniques have garnered attention as potential tools to aid healthcare professionals in this endeavour. However, the imbalanced distribution of diabetes cases within datasets, where the positive class (diabetes cases) is significantly lower by the negative class (non-diabetes cases), poses a challenge to the development of robust prediction models.

In classification tasks, a data set is imbalanced when the class proportion are substantially different. Generally, the instances of majority class outnumber the amount of minority class instances. The ration between the majority and minority classes may be 10:1, 100:1, 1000: 1 or can be more vary than this. Usually, negative examples or most prevalent class are defined as majority class and positive example or rarest class are called minority class (Huang et al., 2016). Chawla et al. (2002) defined the dataset is imbalanced if the classification categories are not approximately equally represented. Batista et al. (2004) mention that imbalance datasets means that one class might be represented by many examples, while the other is represented by only few.

In recent years, researchers have been exploring innovative methods to address the challenges posed by imbalanced datasets in diabetes prediction. Among these methods, the combination of Synthetic Minority Over-sampling Technique (SMOTE) has emerged as a promising approach (Chawla et al 2004b, Kirui 2013, Yavuz 2019). SMOTE helps mitigate the class imbalance by generating synthetic instances of the minority class, effectively increasing its representation in the dataset. This study will address the problem of diabetes imbalanced dataset. It will aim to improve the classifier performance in predicting diabetes cases.

2. Methodology:

2.1 Research Process

Figure 1 depicts the framework of classification phases for this study. Dataset will split into two subsets: a training set with a ratio of 80% and a testing set with a ratio of 20%. The training set is used to build the classification model, while the testing set is used to evaluate its performance. The imbalanced training set will be balanced via synthetic oversampling (OS) approaches and being trained to develop a model. Random Forest classifier will be used as the classification algorithm for this study. The process continues by applying the trained model to the testing set, which contains data that the model has not seen during training. The model's predictions on the testing set are compared to the actual class labels. The final stage is to calculate various performance measures to assess how well the model is performing.

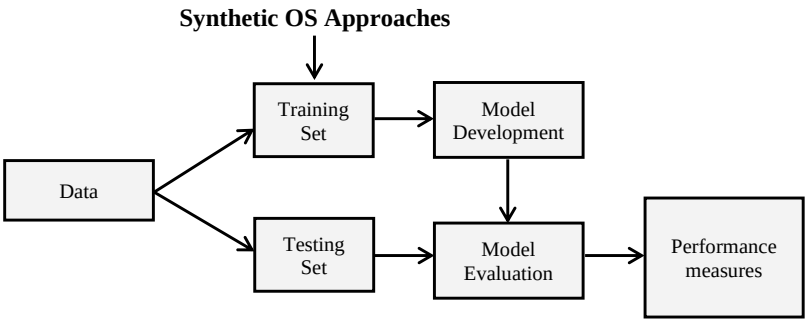

Figure 1. Classification Phases

2.2 Oversampling Technique

Oversampling (OS) approach is a widely used method to solve the imbalanced data. According to Batista et al (2004) in general, oversampling methods give better results than under sampling methods. When data is highly imbalance, significant differences between majority dan minority class can be handled by oversampling methods. In oversampling, the instances of minority class are duplicated randomly until a required sample size is obtained. However, these methods tend to remove important information of the sample or can lead to the introduction of meaningless new objects (Chawla et al., 2002). This have motivated the researchers to modify the existing/basic approach to improve its classification performance. The most popular modify techniques in data-level approach is Synthetic Minority Over-sampling Technique (SMOTE) proposed by Chawla in the year of 2002. Chawla et al. 2002, Chawla et al. 2003, Batista 2004, Han et al. 2005, Kirui 2013, Yavuz 2019 and Turlapati 2020 proved that SMOTE shown improved performance in handling CIP. Unlike OS, SMOTE create synthetic data in minority class to have a balanced distribution. Chawla mentioned that this technique is inspired by a technique that proved successful in handwritten character recognition by Ha & Bunke (1997) by creating extra training data by performing certain operations on real data.

In general, SMOTE generates synthetic data based on the distance between the minority data and the closest minority data therefore the new synthetic data will be formed between the two minority data. Formula to generate synthetic sample by SMOTE can be expressed as

$$D_{new} = D_i + (\widehat{D_l} - D_i) \times \delta$$

Where D_{new} = synthetic data, D_i=examples from minority, $\widehat{D_l}$=one of k-nearest neighbor from D_i, δ=random number between 0 and 1.

Synthetic samples are generated by taking the difference between the feature vector (sample) under consideration and its nearest neighbor up to five nearest neighbors. Then, multiply this difference by a random number between 0 and 1, and add it to the feature vector under consideration.

Although SMOTE is quite effective in improving the classification accuracy of the minority data, there is still room for improvement for this method. SMOTE may lead to the occurrence of overgeneralization. Synthetic data created by SMOTE is still possible to spread on both minority and majority data, hence it will reduce the performance of classification.

This has motivated other researchers to modify this method to create more effective techniques to improve classification performance. Han et al. (2005) develop Borderline-SMOTE which is only the minority examples near the borderline are over-sampled. This method focuses on the borderline area located on the boundary between the minority and majority of data. Borderline-SMOTE proved that it can achieve better true positive (TP) rate and F-value than SMOTE and random oversampling methods.

Adaptive Synthetic (ADASYN) sampling approach was proposed by He et al. (2008) and works similar to SMOTE in that it generates synthetic observations for the minority class. It is however based on generating more synthetic data for observations that are harder to learn than those that are easier to learn for a given model. As with SMOTE, ADASYN generates synthetic observations along a straight line between a minority class

observation and its k-nearest minority class neighbors. As with SMOTE, the number of k-nearest neighbors is set to five. However, ADASYN generates more synthetic observations for minority class observations which have more majority class observations inside the k-nearest neighbors' region.

2.3 Performance Measures

This section will provide a discussion on how to evaluate the performance of the different approaches on the testing set. This procedure will assist us to judge the specific objective of the reduction of imbalanced ratio in CIP dataset and the improvement the classifier performance for CIP dataset.

In this study, the classification performance will be evaluated by using few measurements which is accuracy, specificity, sensitivity, and precision.

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

$$Sensitivity = \frac{TP}{FN + TP} \qquad Precision = \frac{TP}{TP + FP}$$

All these performance measures are indicated from the amount of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). These values are based on the confusion matrix in binary class as presented in Table 1.

Table 1. Confusion Matrix for Binary Class

	Predicted as positive	**Predicted as negative**
Positive Class	True Positive (TP)	False Negative (FN)
Negative Class	False Positive (FP)	True Negative (TN)

2.4 Data and Tools

In this study, a study will be carried out by using imbalanced diabetes datasets. This dataset originated from National Diabetes and Digestive and Kidney Diseases Institute and based on data collected during 1980s. The dataset's purpose is to predict a diabetic patient based on some diagnostic measure. This dataset is an open-source dataset from Kaggle named Pima Indians Diabetes Database contains 768 samples, with 268 diabetic patients and 500 non- diabetic patients. All patients in this dataset are females who are at least 21 years old. This imbalanced data set will split into two groups. 80% of these data that will be selected by stratified random sampling and became the training data while the remaining 20% of the dataset is used to validate the efficiency of the proposed method as a testing set. This study was conducted by using Python and SPSS to access the performance of synthetic OS approaches.

3. Result:

The diabetic dataset used in this study contain 768 samples with 9 attributes. Table 2 presents the statistics description of the dataset attributes: Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, Body Mass Index (BMI), Diabetes Pedigree Function, Age and Outcome as indication of diabetic and non-diabetic. Table 3 shows the frequency

for each class in Outcome as Positive referring to diabetic patients and Negative referring to non-diabetic patients with negative as majority group and positive as minority group.

Table 2. Descriptive Summary

	N	Minimum	Maximum	Mean	Std. Deviation
Pregnancies	768	0	17	3.85	3.37
Glucose	768	0	199	120.89	31.97
Blood Pressure	768	0	122	69.11	19.36
Skin Thickness	768	0	99	20.54	15.95
Insulin	768	0	846	79.80	115.24
BMI	768	0	67	31.99	7.88
Diabetes Pedigree Function	768	0	2	0.47	0.33
Age	768	21	81	33.24	11.76
Outcome	768				

Table 3. Outcome Frequency

	Frequency	Percent
Positive	268	34.9
Negative	500	65.1
Total	768	100.0

After the data have been split into 2 sets (training and testing set), class distribution for training set before and after applying synthetic OS are presented in Table 4. Class distribution without SMOTE is maintain as imbalanced while class distribution with SMOTE, ADASYN and Borderline-SMOTE are balanced with equal and almost equal ratio for both classes.

Table 4. Class Distribution on Training Set

	Without SMOTE	SMOTE	ADASYN	Borderline- SMOTE
Positive	401	401	401	401
Negative	213	401	380	401

The values of the statistics obtained on performance measures with and without synthetic OS approaches are given in Table 5. The classifier shows less accuracy performance on imbalanced dataset compared to balanced dataset. ADASYN and SMOTE show the highest value of accuracy which is 74.7% each. Specificity of the classifier performance without SMOTE have the highest value of 77.8%. ADASYN performed well on sensitivity compared to other approaches with a value of 76.4%. The classifier shows high precision performance on balanced dataset via SMOTE with a value of 62.1%.

Table 5. Results of Performance Measures

	Accuracy	Specificity	Sensitivity	Precision
Without SMOTE	0.721	0.778	0.618	0.607
SMOTE	0.747	0.747	0.745	0.621
ADASYN	0.747	0.737	0.764	0.618
Borderline- SMOTE	0.727	0.717	0.745	0.594

4. Discussion and Conclusion:

As overall, the classifier performed well in balanced dataset with the help of synthetic OS approaches. By focusing on the sensitivity value, ADASYN is the best synthetic OS approach in predicting diabetic cases since sensitivity is looking on the proportion of the total number of correctly classified under minority (positive) class. In conclusion, Synthetic OS approaches has shown promise in addressing the challenges posed by imbalanced datasets in diabetes prediction. The outcome of this study has contributed to one of the main objectives of SDG 3 which is to ensure healthy lives and promote well-being for all ages. This diabetes prediction can contribute to this goal by identifying individuals at risk of developing diabetes at an early stage. Early intervention and management can lead to better health outcomes, reducing the burden of diabetes-related complications and improving the overall quality of life.

References:

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD explorations newsletter, 6(1), 20-29.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004b). Editorial : Special Issue on Learning from Imbalanced Data Sets. ACM SIGKDD Explorations Newsletter, 6(1), 1–6.

Chawla, N. V., Lazarevic, A., Hall, L. O., & Bowyer, K. W. (2003). SMOTEBoost: Improving prediction of the minority class in boosting. In European conference on principles of data mining and knowledge discovery. Springer, Berlin, Heidelberg, 107-119.

Han, H., Wang, W. Y., & Mao, B. H. (2005, August). Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In International conference on intelligent computing (pp. 878-887). Springer, Berlin, Heidelberg.

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008, June). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)* (pp. 1322-1328). Ieee.

Huang, C., Li, Y., Loy, C. C., & Tang, X. (2016). Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5375-5384).

Kirui, C., Hong, L., & Kirui, E. (2013). Handling class imbalance in mobile telecoms customer churn prediction. International Journal of Computer Applications, 72(23), 7-13.

Turlapati, V. P. K., & Prusty, M. R. (2020). Outlier-SMOTE: A refined oversampling technique for improved detection of COVID-19. Intelligence-based medicine, 3, 100023.

Yavuz, Ü. N. A. L., Sağlam, A., & Kayhan, O. (2019). Improving classification performance for an imbalanced educational dataset example using SMOTE. *Avrupa Bilim ve Teknoloji Dergisi*, 485-489.