SDGnE: A Synthetic Data Generation and Evaluation System for Rare Event Prediction

Wan D. Bae[1], Shayma Alkobaisi[2](✉), Sartaj Bhuvaji[1], and Siddheshwari Bankar[1]

[1] Computer Science, Seattle University, Seattle, WA, USA
{baew,sbhuvaji,sbankar}@seattleu.edu

[2] College of Information Technology, United Arab Emirates University, Al Ain, UAE
shayma.alkobaisi@uaeu.ac.ae

Abstract. Class imbalance in datasets creates a significant challenge for building efficient classifiers and results in poor prediction of rare events. This problem is more critical in applications where the size of the dataset is often small such as individual-based health risk prediction modeling and engineering problems heavily based on simulations. While several techniques have been proposed in this field, their performance with small size datasets requires improvement for practical use of the machine learning algorithms. This paper presents a system framework called "Synthetic Data Generation and Evaluation (SDGnE)" for the class imbalance problem by generating synthetic data using various techniques, analyzing data quality, and comparing the performance of the implemented techniques. We demonstrate the proposed system using a web-based user interface that includes methods for data generation, statistical analysis, and visual evaluation. The proposed system can help users have better understanding and insight of the generated data when using different techniques and can be straightforwardly extended to include new data generation techniques and evaluation tools.

Keywords: synthetic data generation · class imbalance · classification · SMOTE · generative adversarial network · autoencoder

1 Motivation

A dataset is imbalanced if the class of interest (minority class) has significantly fewer samples than the other classes (majority classes). The classification on imbalanced datasets has received widespread attention in many applications, particularly medical applications focusing on predicting rare events correctly rather than normal cases, such as cancer diagnosis and asthma exacerbation prediction. It has been an increasing need across disciplines to find ways of improving the accuracy of machine learning (ML) models in rare events prediction.

M. Onizuka et al. (Eds.): DASFAA 2024, LNCS 14856, pp. 508–512, 2024.
https://doi.org/10.1007/978-981-97-5575-2_49

Synthetic data generation techniques are under active research and development in ML community, which include Synthetic Minority Oversampling Techniques (SMOTE) [1,2,5–7], autoencoders [9], and generative adversarial networks (GAN) [4]. While many of the techniques have shown a great success in text, audio, and video related applications with the availability of large datasets, the performance of models trained on small sized medical data is not satisfactory [1]. Moreover, little exists in the literature regarding the quality of data generated by different techniques and their effect on the performance of models.

This paper proposes a web-based system that integrates various techniques to generate synthetic data and evaluation tools to measure the data quality. The system can help users build a framework of core infrastructures for rare event prediction models in medical applications. Our demonstration using real asthma patients' data includes the system overview, data generation techniques, visual evaluation tools and user interface.

2 System Framework

2.1 System Overview

We propose a system framework for the class imbalance problem called "Synthetic Data Generation and Evaluation (SDGnE)", which uses the standard module approach for the high-level architecture. Figure 1 presents three main components of the system: (1) user interface, (2) analytical processing engine including data generation and data evaluation modules, and (3) data management.

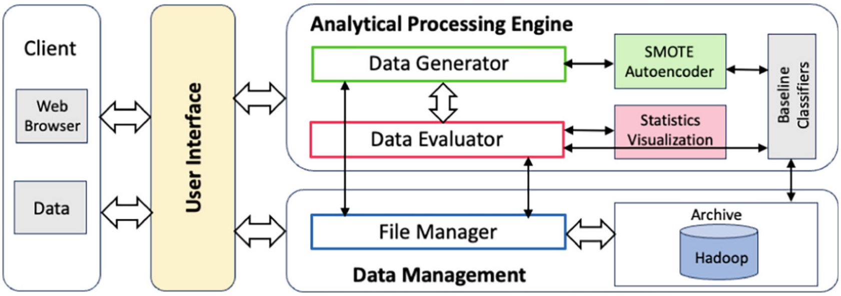


Fig. 1. System overview

The system takes an input file containing a set of data features and class label. Users can explore various techniques to generate the synthetic data and evaluate its quality. They can also download the generated data or augment them to the original data as well as download a summary of the evaluation.

The following technologies are used to implement SDGnE: Windows 11 to build the application but it is noted that the program can run on any operating systems; the frontend is a Python Web interface; the backend and evaluation engine were implemented using Python 3; SDGnE implements data repository using Hadoop file system. Our proposed system is evaluated using a case study of asthma risk prediction which serves as a prototype for generating and evaluating synthetic data.

2.2 Methods for Synthetic Data Generation and Evaluation

The proposed framework provides the following set of state-of-the-art data augmentation techniques:

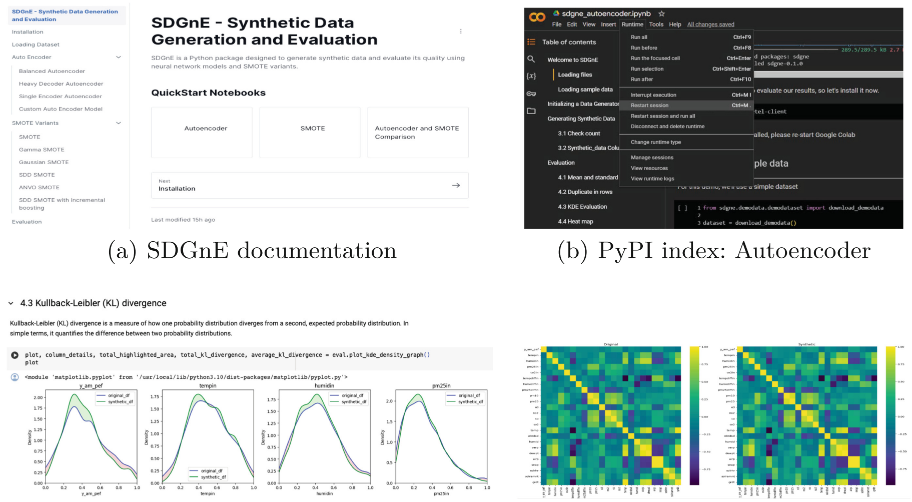
- Synthetic Minority Oversampling Techniques (SMOTE) generates synthetic samples similar to the minority class data to achieve a more balanced class distribution of samples. The methods implemented in the system are SMOTE [2], Gaussian SMOTE [6], Gamma SMOTE [5], Sample Density Distribution SMOTE [7], and Average Neighbor Vector Oversampling [1].
- Autoencoders are neural networks that encode input sample data into a smaller representation and re-generate similar data from this representation. We implemented three baseline autoencoders: single encoder, heavy decoder and balanced autoencoder. We also implemented a TVAE model that is based on the VAE-based Deep Learning data synthesizer [9].

Data evaluation is based on two aspects: while maintaining similar probability density functions of features in the real data, diversifying the data that is sufficiently representative of real data to reduce overfitting. Several metrics were implemented: (a) mean and standard deviation, (b) Gretel score [3], (c) Kullback-Leibler divergence, and (d) Kernel density estimation. The system also allows users to check the data quality by using baseline classifiers, and can be extended to integrate other data generation and evaluation techniques.

3 Demonstration

ML models can be used to estimate an asthma patient's health risk from the patient's accumulated exposure to environmental triggers, by answering to "what is the probability of an asthma patient having a peak expiratory flow rate to fall tomorrow below a certain threshold value, given his/her environmental exposure today?" and provide optimal, individual-specific recommendations on how to reduce health risks. We demonstrate the proposed system by generating synthetic minority class samples for asthma patients' health risk prediction and comparing the data quality through visual evaluation tools. Each patient's dataset includes 168 samples at average (minority class samples: 35, majority class samples: 132) with 27 features of the patient' physiological condition and their exposures to air pollutants and other environmental variables [1,8].

The system is demonstrated through the main functionalities: (1) upload an input file, (2) display the data in a table, (3) select data generation methods, (4) visualize a summary of data quality and comparisons of selected methods, and (5) download the generated data. A prototype of SDGnE packages are shown in Fig. 2. and a demo video provides several examples for the use of the packages. Users can also access SDGnE documentation and a GitHub repository for code.


SDGnE - Synthetic Data Generation and Evaluation
Installation
Loading Dataset
Auto Encoder
Balanced Autoencoder
Heavy Decoder Autoencoder
Single Encoder Autoencoder
Custom Auto Encoder Model
SMOTE Variants
SMOTE
Gamma SMOTE
Gaussian SMOTE
SDD SMOTE
ANVO SMOTE
SDD SMOTE with incremental boosting
Evaluation

SDGnE - Synthetic Data Generation and Evaluation

SDGnE is a Python package designed to generate synthetic data and evaluate its quality using neural network models and SMOTE variants.

QuickStart Notebooks

Last modified 15h ago

(a) SDGnE documentation

```
289.5/289.5 kB 2.7
d packages: sdgne
led sdgne-0.1.0
```

evaluate our results, so let's install it now.

```
tel-client
```

alled, please re-start Google Colab

ple data

For this demo, we'll use a simple dataset

```
[ ]  1 from sdgne.demodata.demodataset import download_demodata
     2
     3 dataset = download_demodata()
```

(b) PyPI index: Autoencoder

4.3 Kullback-Leibler (KL) divergence

Kullback-Leibler (KL) divergence is a measure of how one probability distribution diverges from a second, expected probability distribution. In simple terms, it quantifies the difference between two probability distributions.

```
plot, column_details, total_highlighted_area, total_kl_divergence, average_kl_divergence = eval.plot_kde_density_graph()
plot
```

```
<module 'matplotlib.pyplot' from '/usr/local/lib/python3.10/dist-packages/matplotlib/pyplot.py'>
```

(c) Kullback-Leibler divergence

(d) Heat map analysis

Fig. 2. SDGnE packages and evaluation tools

Acknowledgements. This work was supported by UAEU-NFRP grant number G00004281 and SU-Bannan Endowed Chair of Engineering Fund.

References

1. Bae, W.D., Alfonso, A., Stanko, D., Hao, L., Le, L., Horak, M.: Improving classification performance on rare events in data starved medical applications. In: 2023 IEEE International Symposium on Medical Measurements and Applications (MeMeA), pp. 1–6. IEEE (2023)
2. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
3. Gretel: Gretel. Accessed Jan 2024. https://gretel.ai/
4. Huang, Y., Fields, K.G., Ma, Y.: A tutorial on generative adversarial networks with application to classification of imbalanced data. The ASA Data Science Journal, Statistical Analysis and Data Mining (2021)
5. Kamalov, F., Denisov, D.: Gamma distribution-based sampling for imbalanced data. Knowl.-Based Syst. **207**, 106368 (2020)

6. Lee, H., Kim, J., Kim, S.: Gaussian-based smote algorithm for solving skewed class distributions. Int. J. Fuzzy Logic Intell. Syst. **17**(4), 229–234 (2017)
7. Wan, Q., Deng, X., Li, M., Yang, H.: Sddsmote: synthetic minority oversampling technique based on sample density distribution for enhanced classification on imbalanced microarray data. In: The 6th International Conference on Compute and Data Analysis, pp. 35–42 (2022)
8. Woo, J., Rudasingwa, G., Kim, S.: Assessment of daily personal pm2. 5 exposure level according to four major activities among children. Appl. Sci. **10**(1), 159 (2020)
9. Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling tabular data using conditional GAN. In: Advances in Neural Information Processing Systems, vol. 32 (2019)