Synthetic Data Generation in Small Datasets to Improve Classification Performance for Chronic Heart Failure Prediction

Roy S Zawadzki, Saman Parvaneh

Edwards Lifesciences, Irvine, USA

Abstract

In cardiology, we frequently develop machine learning models to predict events such as heart failure. Oftentimes, these events occur at low incidence in the available data, especially for under-represented subpopulations, which limits classifier performance due to class imbalance. To mitigate these issues, we investigate the use of synthetic data generation, or algorithms trained to mimic realistic patient data. In particular, we use synthetic data to augment training data for Catboost in classifying chronic heart failure using the University of California, Irvine myocardial infarction complications dataset (n = 1,700). Our primary metrics of interest are the mean and the variability of AUC and F1-Score across five-fold cross-validation. Overall, we find modest gains in performance over the baseline classifier with no augmented data. Nevertheless, the more sophisticated generators, both with and without hyperparameter tuning, did not confer better performance than simpler methods. Furthermore, all methods were subject to large variability in classification metrics across folds. While synthetic data generation is a promising tool for class imbalance, more investigation is needed to find optimal sample sizes and settings for the stability of results.

1. Introduction

A common application of machine learning (ML) in cardiology is the detection and prediction of events (e.g., arrhythmia and heart failure). Often, it is the case that the available datasets for training ML models contain a low incidence rate of these events, which is commonly referred to as “class imbalance.” Class imbalance creates a scenario where a classifier’s performance can be biased toward the majority class even if the dataset is reasonably sized. In turn, this results in poor performance for the minority class, or those who are truly of interest in an adverse outcome prediction model [1, 2]. Furthermore, class imbalance can exacerbate disparities in ML performance among certain underrepresented subpopulations in medical datasets, such as women and non-whites.

Popular methods to mitigate class imbalance include either undersampling the majority class or oversampling the minority class via the synthetic minority oversampling technique (SMOTE). Undersampling is usually undesirable for classification due to losing potentially valuable data points [3]. Meanwhile, SMOTE may not produce a diverse sample across the entire support of the distribution of features due to its local “nearest-neighbors” approach [3].

A new class of methods, called synthetic data generation (SDG), offers another potential solution to class imbalance. SDG refers to models, often neural networks, trained to create “realistic” patient records that preserve the original dataset’s complex multivariate distributions between the features [4]. The generators can then be used to produce additional data points to augment original training sets for ML algorithms. Compared to traditional methods, if SDG can produce higher quality and more diverse data, downstream ML performance in the presence of class imbalance should theoretically improve.

We investigate the use of SDG, both with and without hyperparameter tuning, for two aims: (1) to increase overall ML performance by augmenting the training set with generated synthetic data and (2) to mitigate ML performance disparities by augmenting only generated data for certain underrepresented subpopulations. We utilize the University of California, Irvine Myocardial Infarction (MI) complications dataset as a case study [5]. Specifically, using features available upon hospitalization, we use a Catboost classifier to predict chronic heart failure (CHF), a binary outcome with an imbalance in both CHF incidence and the proportion of females in the dataset.

2. Methods

2.1. Dataset and Pre-processing

Our dataset contains 1,700 patients who were hospitalized due to an MI. 394 (23.2%) patients were ultimately diagnosed with CHF and 635 (37.3%) were female. The male and female CHF rates were 20% and 29%, respectively. After excluding features not available upon hospitalization, severely imbalanced binary features (less than 1% incidence), highly missing features (over 70% missing), and highly correlated features (over 70% pairwise Pearson correlation), we were left with 60 features. To further decrease the feature count to prevent overfitting, we selected 30 clinically meaningful features to achieve the rule of thumb of ten events per feature [6, 7]. These features include a mix of numerical, categorical,

ISSN: 2325-887X DOI: 10.22489/CinC.2023.081

and binary features. Please refer to the MI complications dataset reference for more details [5].

2.2. Synthetic Data Generators

We utilize the synthetic data generators available in the open-source "Synthetic Data Vault (SDV)" Python library: Gaussian Copula (GC), Conditional Tabular Generative Adversarial Network (CT GAN), Copula GAN, and Tabular Variational Autoencoder (TVAE) [8]. Let d represent the number of features; for the GC method, a copula is a unit cube $[0,1]^d$ constructed from the marginal cumulative distribution functions of a multivariate Gaussian distribution. The latter three algorithms are based on generative adversarial networks (GANs), which consist of two submodels: the "generator" that produces synthetic data and the "discriminator" that attempts to discern whether a data point is real or synthetic [9]. The sub-models are trained together in a zero-sum manner such that one model improves when the other fails. For a simple method to compare to SDG, we sample with replacement, or bootstrap sample, the original dataset as our "synthetic data" augmentation.

2.2. Hyperparameter Tuning

Within each aim, we examined whether the SDG model hyperparameter tuning affects the downstream classifier performance. Only the GAN models in SDV (CT GAN, Copula GAN, and TVAE) were tunable. Using each fold's training set, we searched the following grid:

- Epochs: 100, 200, 300 (SDV default: 300)
- Dimension of generator and discriminator: 64, 128, 256 (SDV default: 256)
- Discriminator steps (for CT GAN and Copula GAN only): 1, 3, 5 (SDV default: 1)

These values were chosen to ideally minimize the overfitting of the generators to the small dataset.

The metric optimized for choosing the final hyperparameters was the SDV "quality score," a score between 0 and 1 that quantifies how similar the generated data's distributions are to the real data the generator was trained on. For each column in the generated data, a similarity score between the real and synthetic data was calculated: for continuous columns, one minus the Kolmogorov-Smirnov statistic and, for non-continuous columns, one minus the total variation distance. Then, these scores were averaged across all columns to calculate the final quality score.

Ultimately, for each fold, the combination of hyperparameters with the best quality score will be used in the final generator model. We only computed quality scores on data generated for the entire population (i.e., not the female-only data) and used the best hyperparameters for the models in both aims one and two. The implicit assumption in the proposed hyperparameter tuning framework is that higher-quality data will improve downstream ML performance. Quality scores were reported for all methods.

2.3. Classification of CHF

We aim to predict the presence of CHF given the 30 features extracted. We examine whether augmenting data produced by SDG to the training set will improve the area under the receiver operating curve (AUC) and F1-score, both overall (aim one) and sex-specific (aim two) when compared to a baseline model trained on the original training. To smooth over variability, we utilize five-fold cross-validation, stratified on CHF and sex, and report the average and standard deviation of the classification metrics across each fold. In each fold, the following is executed:

1. Use random forest imputation to impute missing values in the training and testing set separately [10].
2. Using the training set and Catboost, find the optimal classification probability threshold based on the F1-score.
3. With the optimal threshold, re-train a Catboost model on the training data and record the baseline (i.e., original data with no augmentation) classification metrics on the test set.
4. For each generator examined: (a) train the generator on the training data, (b) generate the pre-determined number of synthetic data observations (see below), (c) augment the synthetic data to the original training set, (d) train the Catboost model with the augmented data, and (e) record test set performance.

For our first study aim, overall performance, we double the training dataset by augmenting 1,300 synthetic data points, and for our second aim, sex-specific performance, we conditional sample females only and augment 500 females, doubling the number of females in the training set.

3. Results

Compared to the default hyperparameter settings, only CT GAN and Copula GAN improved the quality score with custom hyperparameters; thus, there is no hyperparameter-tuned TVAE model. For Copula GAN, across all folds, the best hyperparameters were an epoch count of 300, a dimension of 64, and a discriminator step of 5. In our results, we call this model "Copula GAN + HT" where HT stands for "hyperparameter tuned." For CT GAN, the best hyperparameters varied across folds, with the epoch count ranging from 100 to 300 and dimensions ranging from 64 to 128. However, the discriminator step was always 5. We call these models "CT GAN + HT"

Bootstrap sampling had the highest quality score averaged across folds, followed by the hyperparameter-tuned CT GAN (Table 1). The lowest average quality score came from the GC method.

Table 1: Mean Quality Scores by Method where HT stands for "hyperparameter tuned"

Method	
GC	0.72
TVAE*	0.74
CT GAN	0.77
Copula GAN	0.77
Copula GAN + HT	0.84
CT GAN + HT	0.85
Bootstrap	0.97

*TVAE + HT have the same parameters as TVAE

Figure 1 presents the overall AUC and F1-score metrics for each generator examined. Each method has a dot representing the average metric across folds, while the bars are a result of adding and subtracting the standard deviation across the folds. The wider the bar, the more variable the results were. The red dotted line represents the average baseline performance of the Catboost classifier trained on the original data.

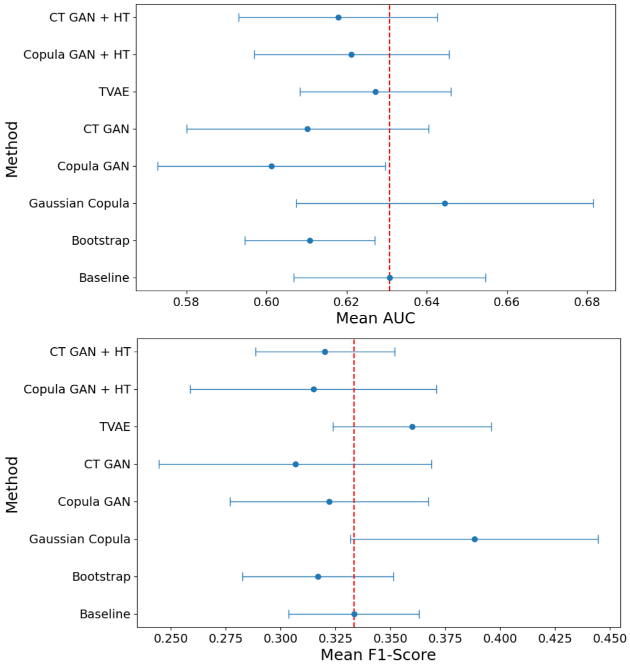

Figure 1: Forest plots of ML performance for AUC and F1-score score across all folds for different SDG methods.

For AUC, on average, only GC showed an improvement but, nevertheless, had the highest variability among all methods that ranged into a considerable decrease in performance. For F1-score, on average, the highest gains were from GC followed by TVAE. CT GAN had the highest variability while TVAE had the lowest.

Figure 2 shows the results from augmenting female-only data. The key performance indicator is an increase in the female-specific metrics without a loss in the male-specific metrics. The baseline AUC for males was better than that of females, while the opposite was true for F1-score.

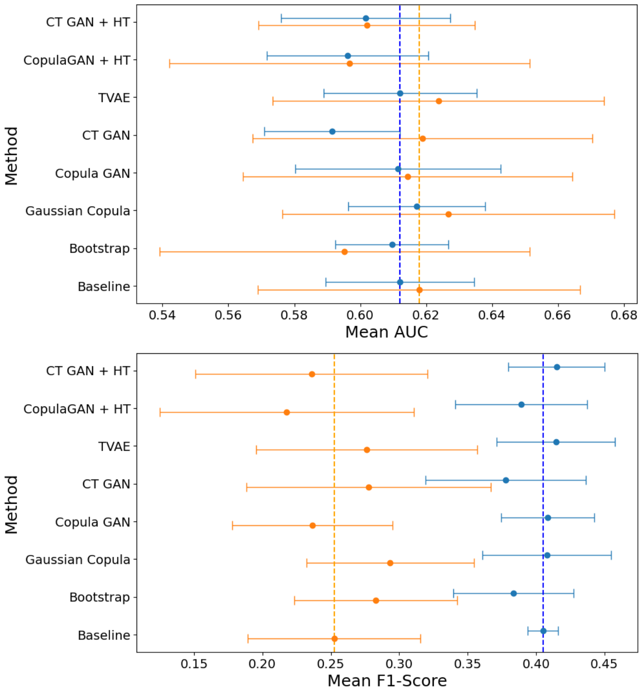

Figure 2: A forest plot of sex-specific ML performance across all folds for AUC and F1-score. Blue and orange lines represent the female and male metrics, respectively. The dotted lines are the baseline performance.

For AUC, GC boosted performance for both females and males. The next best algorithm, TVAE, could only provide gains for males but not females. The results, however, were quite variable. All other algorithms resulted in a performance decrease in at least one of the sexes. Variability was generally large for these results. For F1-score, both TVAE and GC showed notable improvements in both male and female performance. Other methods did indeed increase female performance but at the cost of male performance. Female variability was comparable to the baseline classifier, but male variability was much higher.

4. Discussion

In this case study, we found that utilizing SDG modestly improved classification performance compared to a baseline model. Overall, the complexity of the method (i.e., GAN vs. non-GAN) had little bearing on the final results. Both GC and TVAE improved overall and sex-specific metrics, especially for F1-score – a key metric where there is class imbalance. Interestingly, these two methods had the lowest average data quality (Table 1), demonstrating that data quality is perhaps not directly indicative of downstream ML performance. This is further supported by the poorer performance of the hyperparameter-tuned models, which had the highest data quality scores.

One notable issue with SDG is the variability across the folds, with no methods in Figures 2 and 3 having an unequivocal gain over the baseline classifier in any metric. In some cases, this variability was much larger than the baseline classifier's variability. For example, the F1-score for males in CT GAN ranges from roughly 0.175 to 0.375 (Figure 3). Interestingly, although we did not generate any male observations, the variability across folds increased compared to the baseline classifier. This variability could be ameliorated with larger sample sizes, but this fact motivates a discussion of the potential problems surrounding the use of SDG. Because SDG is ultimately an ML procedure, the generators benefit from larger sample sizes. Nevertheless, as the sample sizes rise, the baseline classifier's performance also increases, and thus, the utility of synthetic data in the use case decreases. One may face a situation where the number of samples required to decrease the variability in the generated data to acceptable levels may be far above what is required for adequate ML performance. Even more importantly, SDG is most useful when the original sample size is low, but the generators themselves require a minimum sample size for acceptable performance. Once again, one could ask whether this sample size is more or less than the sample size required for adequate performance of the original classifier.

Our work has limitations, which motivate future work. Firstly, this case study is only a single example, limiting the scope of conclusions. This same study could be replicated in datasets of different sizes with a varying number of features and types of targets (e.g., continuous). Due to the large computational time to run these generators, we were limited in the number of folds and grid search for meaningful hyperparameters. Furthermore, as opposed to using data quality, other metrics for choosing hyperparameters can be explored. Lastly, we did not compare the SDGs to other methods, such as SMOTE.

Besides the directions already mentioned, an important avenue of future work is the stability of these algorithms under different sample sizes. Intuitively, sample sizes that are too low are a non-starter for training SDG algorithms and the variability of the resulting estimates will be large. Thus, the key is to establish rules of thumb for analysts that clarify the appropriate sample size needed to train SDGs themselves or pooling procedure such the data generated are both accurate and within a certain acceptable variability. Work in the area of multiple imputation of missing data could be relevant, particularly looking at "between dataset" variability [11].

To conclude, SDG has the potential to be a powerful tool in the ML arsenal, allowing a data scientist to both increase the size of training datasets and mitigate class imbalance. Nevertheless, while common sense tells us more data is usually better, not all data is created equal. The use of synthetic data may carry with it a set of nuanced challenges, some of which we have highlighted in this paper.

References

[1] C. Bellinger, S. Sharma, and N. Japkowicz, "One-class Versus Binary Classification: Which and When?." pp. 102-106.

[2] T. van der Ploeg, P. C. Austin, and E. W. Steyerberg, "Modern Modelling Techniques Are Data Hungry: A Simulation Study For Predicting Dichotomous Endpoints," *BMC Medical Research Methodology,* vol. 14, no. 1, pp. 1-13, 2014.

[3] A. F. Hilario, S. G. López, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, "Learning From Imbalanced Data Sets," *Artificial Intelligence. Springer, Cham*, 2018.

[4] M. Hernandez, G. Epelde, A. Alberdi, R. Cilla, and D. Rankin, "Synthetic Data Generation For Tabular Health Records: A Systematic Review," *Neurocomputing,* vol. 493, pp. 28-45, 2022.

[5] S. E. Golovenkin, V. A. Shulman, D. A. Rossiev, P. A. Shesternya, S. Y. Nikulina, Y. V. Orlova, and V. F. Voino-Yasenetsky. "Myocardial Infarction Complications," https://archive.ics.uci.edu/dataset/579/myocardial+infarction+complications.

[6] K. G. Moons, D. G. Altman, J. B. Reitsma, J. P. Ioannidis, P. Macaskill, E. W. Steyerberg, A. J. Vickers, D. F. Ransohoff, and G. S. Collins, "Transparent Reporting of A Multivariable Prediction Model For Individual Prognosis or Diagnosis (TRIPOD): Explanation And Elaboration," *Annals of Internal Medicine,* vol. 162, no. 1, pp. W1-W73, 2015.

[7] D. K. Plati, E. E. Tripoliti, A. Bechlioulis, A. Rammos, I. Dimou, L. Lakkas, C. Watson, K. McDonald, M. Ledwidge, and R. Pharithi, "A Machine Learning Approach For Chronic Heart Failure Diagnosis," *Diagnostics,* vol. 11, no. 10, pp. 1863, 2021.

[8] N. Patki, R. Wedge, and K. Veeramachaneni, "The Synthetic Data Vault." pp. 399-410.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Networks," *Communications of the ACM,* vol. 63, no. 11, pp. 139-144, 2020.

[10] D. J. Stekhoven, and P. Bühlmann, "MissForest—Non-parametric Missing Value Imputation for Mixed-/Type Data," *Bioinformatics,* vol. 28, no. 1, pp. 112-118, 2012.

[11] J. W. Graham, A. E. Olchowski, and T. D. Gilreath, "How Many Imputations Are Really Needed? Some Practical Clarifications Of Multiple Imputation Theory," *Prevention Science,* vol. 8, pp. 206-213, 2007.

Address for correspondence:
Saman Parvaneh
1 Edwards Way, Irvine, CA, USA
parvaneh@ieee.org