Incremental SMOTE with Control Coefficient for Classifiers in Data Starved Medical Applications

Wan D. Bae[1], Shayma Alkobaisi[2](✉), Siddheshwari Bankar[1], Sartaj Bhuvaji[1], Jay Singhvi[1], Madhuroopa Irukulla[1], and William McDonnell[1]

[1] Computer Science, Seattle University, Seattle, WA, USA
{baew,sbankar,sbhuvaji,jsinghvi,mirukulla,mcdonn11}@seattleu.edu

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

Abstract. Prediction models for data-starved medical applications lag behind general machine learning solutions, despite their potential to improve early interventions. This is largely due to the assumption that optimization approaches are applied on a balanced distribution of events, yet medical data often has an imbalanced distribution within classes. The curse of dimensionality is further exacerbated by small samples and a high number of features in individual-based risk prediction models. In this paper, we propose a data augmentation system to gradually create synthetic minority samples with a control coefficient, which improves the quality of generated data over time and consequently boosts prediction model performance. This system incrementally adjusts to the data distribution, avoiding overfitting. We evaluate our approach using four synthetic oversampling techniques on real asthma patient data. Our results show that this system enhances classifiers' overall performance across all four techniques. Specifically, applying the incremental data augmentation approach to three oversampling methods led to an increase in sensitivity of 4.01% to 7.79% in deep transfer learning-based classifiers.

Keywords: class imbalance problem · synthetic minority oversampling technique · rare event prediction · data starved contexts · control coefficient

1 Introduction

Recent advances in machine learning play a pivotal role in clinical decision-making, enabling early disease detection and enhancing patient care across various medical conditions. For example, avoidable asthma exacerbations account for 63% of the total annual asthma costs. Preventive approaches to predict the likelihood of symptom changes and risk levels can improve healthcare quality and achieve significant economic savings [8].

R. Wrembel et al. (Eds.): DaWaK 2024, LNCS 14912, pp. 112–119, 2024.
https://doi.org/10.1007/978-3-031-68323-7_9

However, building robust predictive models in healthcare often faces the challenge of class imbalance within medical datasets. Class imbalance occurs when certain health conditions, demographics, or outcomes are rare or underrepresented, resulting in the minority class having significantly fewer samples than the majority class. This imbalance leads to biases machine learning algorithms and compromises performance, as accurately identifying minority class data is crucial for effective healthcare delivery.

Synthetic minority oversampling (SMOTE) techniques [1,4–6,11], are actively researched to improve prediction models. SMOTE variants and boosted SMOTE methods provide a wide range of solutions for addressing class imbalance. However, their success in data-starved contexts, common in medical applications with few daily observations per patient, is limited. In our individual asthma risk prediction modeling, the average dataset size is 168 and the imbalance ratio is 3.98, hence traditional SMOTE methods may not suffice.

Given the need to generate a substantial amount of synthetic data to rebalance classes in data deprived medical applications, we introduce a new meta-algorithm to improve the quality of synthetic data produced by SMOTE methods and develop a system that can be extended to other data generation techniques. Our research unfolds in several stages. First, we conduct data-level analysis to assess the quality of synthetic data generated by SMOTE variants within the incremental data generation system. Next, we assess the effectiveness of the proposed system on conventional classifiers. Finally, we explore deep transfer learning (TL) classifiers to enhance classifier performance.

2 Related Work

Synthetic Minority Oversampling Technique (SMOTE) is a widely used method for addressing class imbalance by generating synthetic samples similar to minority class data, thus achieving a more balanced class distribution. The first SMOTE method [1] generates synthetic samples using k-nearest neighbors and linear interpolation. This often results in data lying on the same line, motivating researchers to explore alternative data distributions.

Gaussian SMOTE (G-SMOTE) [6] generates synthetic samples similar to SMOTE [1] but incorporates a Gaussian distribution to produce more varied data, avoiding the duplication common with SMOTE. Gamma SMOTE (Gamma-SMOTE) [5] uses the Gamma distribution to create new samples in a non-linear manner. Due to the Gamma distribution's asymmetry, new minority samples are generated close to the existing minority data samples. Sample Density Distribution SMOTE (SDD-SMOTE) [11] considers the overall dataset distribution and local sample density to reduce fuzzy classification boundaries and control the randomness of the SMOTE method. It identifies the k-nearest neighbors of minority samples, measures their density, and generates synthetic samples with controlled coefficients to balance class distribution. Integrating AdaBoost with SMOTE, as in SMOTEBoost [2], is a natural way to enhance the SMOTE methods by combining oversampling with boosting algorithms. Another approach [9] adjusts the weights of synthetic data to mitigate data noise.

Generative Adversarial Networks (GANs) and autoencoders, typically used for synthetic image generation, have been applied to tabular data. Techniques like conditional tabular GAN [13] and the SMOTified GAN technique [10] show potential in augmenting tabular data and improving prediction models. Despite these advancements, they have not been successful with small training datasets.

3 Method

"Incremental boosting" typically refers to a technique where boosting algorithms are applied sequentially in multiple stages, each building on the results of the previous stage. The goal is to iteratively correct the mistakes made by previous learners and to improve the overall accuracy of the model. In our context, incremental boosting is applied in multiple iterations to enhance the performance of the SMOTE methods. After each iteration, new synthetic samples are generated using the augmented training data, and the boosting process is reapplied to further refine the model's ability to classify minority class instances.

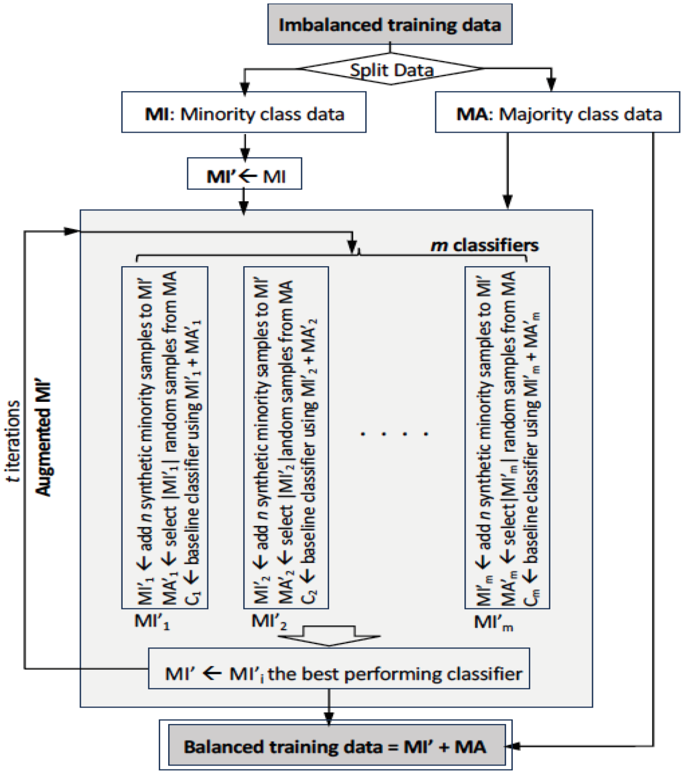
Imbalanced training data
Split Data
MI: Minority class data
MA: Majority class data
MI' ← MI
m classifiers
t iterations
Augmented MI'
MI'1 ← add n synthetic minority samples to MI'
MA'1 ← select |MI'1| random samples from MA
C1 ← baseline classifier using MI'1 + MA'1
MI'2 ← add n synthetic minority samples to MI'
MA'2 ← select |MI'2| random samples from MA
C2 ← baseline classifier using MI'2 + MA'2
. . . .
MI'm ← add n synthetic minority samples to MI'
MA'm ← select |MI'm| random samples from MA
Cm ← baseline classifier using MI'm + MA'm
MI'1
MI'2
MI'm
MI' ← MI'i the best performing classifier
Balanced training data = MI' + MA

Fig. 1. Overview of an incremental synthetic data generation system

3.1 An Incremental Synthetic Data Generation System

The incremental synthetic data generation system divides the SMOTE data generation process into several iterations, as illustrated in Fig. 1. In each iteration, a SMOTE method generates n synthetic minority class samples and adds them to the current minority training dataset. The system creates m such minority training datasets. Then m classifiers are trained and validated using these datasets. The system selects the subset of synthetic data that produces the best-performing classifier and adds it into the current training data.

To select the best subset, we utilize a weighted evaluation function that incorporates performance metrics, focusing on correctly predicting minority class events. In subsequent iterations, the method generates minority class samples by using the augmented training data. By including synthetic data from previous iterations, the SMOTE method can generate more diverse data, thus reducing overfitting in classification models.

For further improvement, we adopt the Control Coefficient (CC) from the SDD-SMOTE algorithm [11] to consider the distribution of synthetic data points relative to the dense area of minority samples. The CC value is calculated during data preprocessing and applied when the SMOTE_{incrCC} method generates synthetic data points. A random probability distribution function is used for SDD and SDD-SMOTE methods, Gaussian function for G-SMOTE, and Gamma function for Gamma-SMOTE.

4 Experiments

4.1 Datasets and Experiments Setup

Our experiments used datasets from 20 non-smoking asthma patients, collected through a case study [12]. Each data sample consists of 27 variables including the patient' peak expiratory flow rate (PEFR) and indoor/outdoor environmental exposure data, and a binary label indicating risk (0 or 1). The variables and measurements are summarized in Table 1.

Patients' exposures to environmental variables were estimated using a 24-h time window for PEFR measurement. The high-risk zone is defined as a PEFR below a patient's critical cutoff (PEFR_C), set at the 20% quantile PEFR value of the patient's dataset, with samples below PEFR_C being minority samples. Patients' datasets contain between 88 and 210 samples (average 168), with minority class samples ranging from 16 to 38 (average 35) and majority class samples from 72 to 172 (average 132). The class imbalance ratio ranged from 2.32 to 5.52 (average 3.98), higher than the ratios in the SDD-SMOTE [11].

All our data generation methods were implemented in Python 3.8 using the Keras framework, with data analysis and model performance evaluation conducted using scikit-learn. Model hyperparameters were selected through extended training and validation with k-fold cross validation (CV) to avoid overfitting and enhance performance. For TVAE, we used the Adam optimizer (learning rate = 0.001), 200 epochs, 16 batch size, and a validation split rate of 0.25.

Table 1. Variables and measurements in 20 asthma patients' datasets

Data Category	Major variables	Measurement
Physiological data	yesterday's PEFRs	twice a day
Indoor air quality	$PM_{2.5}$, CO_2, temperature, humidity	every 60 s.
Outdoor air quality	SO_2, CO, O_3 , NO_2, PM_{10}, temperature, humidity	every 30 min.
Home location	home distance from major roads	level 1–level 5
Life style	income level	level 1–level 9
Cooking habit	frequency of frying	level 1–level 7

Table 2. Statistical summary on the generated synthetic data

Method	Mean diff.	STD diff.	KL	KDE area	Gretel score
SMOTE	**0.27%**	13.08%	0.008074	1.02	92.63
SMOTE$_{incrCC}$	0.33%	**12.31%**	**0.006867**	**0.85**	**92.95**
G-SMOTE	1.33%	14.17%	**0.012217**	**1.32**	**91.88**
G-SMOTE$_{incrCC}$	**0.96%**	**10.32%**	0.015719	1.44	91.58
Gamma-SMOTE	0.69%	14.73%	0.011598	1.17	90.25
Gamma-SMOTE$_{incrCC}$	**0.56%**	**12.03%**	**0.004292**	**0.69**	**92.84**
SDD-SMOTE	2.24%	12.95%	**0.005129**	**0.82**	**93.74**
SDD-SMOTE$_{incrCC}$	**1.84%**	**11.03%**	0.006220	0.85	91.52
TVAE synthesizer	3.84%	33.21%	0.070124	2.84	77.34

Conventional classifiers used 3-fold CV, while TL classifiers used 5-fold CV for the source model and 3-fold CV for the target model, with the Adam optimizer (learning rate = 0.001) and 100–1,000 epochs for both models. Each dataset was split into 80% training/validation and 20% testing. The training/validation data was augmented with synthetically generated data, while the testing data retained the original imbalance ratio.

4.2 Statistical Analysis

Synthetic data generation for rebalancing classes focuses on three factors: (1) maintaining similar probability density functions of variables within the augmented training dataset, (2) preserving class boundaries, and (3) increasing data diversity to reduce overfitting.

Factor (1) is measured by the difference between the means of the original and generated datasets, with a smaller difference indicating better maintenance of data distribution. Factor (2) is assessed by the difference in standard deviations (STD), where a large increase suggests boundary distortions. Factor (3) is measured by probability density functions, such as Kullback-Leibler (KL) divergence and Kernel density estimation (KDE). We also used an open platform called "Gretel" [3] that assesses the overall quality of synthetically generated

Table 3. Conventional classifiers with SDD-SMOTE and incremental SDD-SMOTE

Classifier	Method	Accuracy	Sensitivity	Specificity	Precision	F_1 score	AUC ROC
DT	SDD-SMOTE	0.5955	0.3979	**0.7931**	0.5849	0.5772	0.5908
	SDD-SMOTE$_{incrCC}$	**0.5987** (+0.54%)	**0.4194**(+5.40%)	0.7780 (−1.90%)	**0.5884** (+0.60%)	**0.5783** (+0.19%)	**0.5974** (+0.11%)
KNN	SDD-SMOTE	0.5974	0.5255	0.6722	0.5783	0.5615	0.5851
	SDD-SMOTE$_{incrCC}$	**0.6177** (+3.40%)	**0.5287** (+0.61%)	**0.7127** (+6.02%)	**0.5904** (+2.09%)	**0.5710** (+1.69%)	**0.5976** (+2.13%)
LR	SDD-SMOTE	0.6180	0.5290	0.7089	0.5894	0.5805	0.5997
	SDD-SMOTE$_{incrCC}$	**0.6453** (+4.42%)	**0.5447** (+2.97%)	**0.7459** (+5.22%)	**0.6097** (+3.44%)	**0.6043** (+4.10%)	**0.6178** (+3.02%)
NB	SDD-SMOTE	0.6011	0.3880	**0.8141**	0.5881	0.5876	0.5992
	SDD-SMOTE$_{incrCC}$	**0.6080** (+1.15%)	**0.4095** (+5.54%)	0.8065 (−0.93%)	**0.5951** (+1.19%)	**0.5890** (+0.24%)	**0.6070** (+1.30%)

data. Gretel measures data quality by comparing the distributional distance between the principal components in the original and synthetic data, with closer principal components indicating higher data quality.

Table 2 summarizes the statistics of data generated by the incremental data generation system compared to existing SMOTE methods and the TVAE synthesizer [7]. The mean differences between the generated and original data using the incremental generation system are relatively small, ranging from 0.33% to 1.84%. G-SMOTE, Gamma-SMOTE, and SDD-SMOTE reduced the mean values, while SMOTE increased it. All SMOTE methods reduced the STD values: 5.9% in SMOTE, 27.2% in G-SMOTE, 18.3% in Gamma-SMOTE, and 14.8% in SDD-SMOTE. The proposed system decreased KL divergence and KDE area in SMOTE and Gamma-SMOTE but increased them in G-SMOTE and SDD-SMOTE, a trend also seen in Gretel scores. Further analysis is needed to relate these metrics to actual performance in risk prediction. In contrast, data generated by TVAE showed the highest mean and STD differences from the original data, 3.84% and 33.21%, respectively. This resulted in the highest KL and KDE values and the lowest Gretel score.

4.3 Performance Evaluation on Classifiers

We evaluated classifiers using common binary classification metrics: weighted accuracy, F_1-score average, and Receiver Operating Characteristic Area Under the Curve (ROCAUC). While all these metrics are important, we focused on improving sensitivity, representing the model's ability to predict health risks correctly. The effectiveness of the incremental data generation system on each SMOTE method was tested using conventional classifiers and TL classifiers. The conventional classifiers include: (1) Decision Tree (DT), (2) K-Nearest Neighbors (KNN), (3) Logistic Regression (LR), and (4) Naive Bayes (NB). TL classifiers train the source model with population data from 19 asthma patients' datasets and retrain the target model with a target patient's dataset, and then the target model predicts the patient's health risk.

Table 3 shows that the proposed system with SDD-SMOTE$_{incrCC}$ improved classifier performance on augmented data in all metrics except specificity, with improvements of 0.54%-4.42% in accuracy, 0.61%-5.54% in sensitivity, 0.60%-3.44% in precision, 0.19%-4.10% in F1 score, and 0.11%-3.02% in ROCAUC.

Table 4. TL classifiers with SMOTEs, incremental SMOTEs and TVAE

Method	Accuracy	Sensitivity	Specificity	Precision	F_1 score	AUC ROC
SMOTE	0.6697	**0.5449**	0.7846	0.6503	0.6489	0.6647
SMOTE$_{incrCC}$	**0.6762**(+0.97%)	0.5371 (−3.21%)	**0.8153** (+3.91%)	**0.6599** (+1.48%)	**0.6572** (+1.26%)	**0.6762** (+1.72%)
G-SMOTE	0.6592	0.5461	0.7723	0.6308	0.6282	0.6592
G-SMOTE$_{incrCC}$	**0.6973** (+5.79%)	**0.5886** (+7.79%)	**0.8060** (+4.37%)	**0.6549** (+3.81%)	**0.6597** (+5.01%)	**0.6973** (+5.79%)
Gamma-SMOTE	0.6737	0.5505	0.7969	0.6508	0.6501	0.6770
Gamma-SMOTE$_{incrCC}$	**0.6975** (+3.54%)	**0.5726** (+4.01%)	**0.8225**(+3.21%)	**0.6708** (+3.08%)	**0.6704** (+3.12%)	**0.6975** (+3.04%)
SDD-SMOTE	0.6853	0.5720	0.7986	0.6556	0.6543	0.6853
SDD-SMOTE$_{incrCC}$	**0.6982** (+1.88%)	**0.5953** (+4.07%)	**0.7998** (+0.15%)	**0.6591** (+0.54%)	**0.6598** (+0.84%)	**0.6975** (+1.78%)
TVAE Synthesizer	0.6653	0.5062	0.8233	0.6474	0.6461	0.6650

Due to space limitations, we present results using SDD-SMOTE but note similar performance was achieved with other SMOTE methods.

Table 4 presents the performance improvement in TL models trained using augmented datasets generated by the existing SMOTE methods, the incremental SMOTE and TVAE. The improvement in sensitivity for TL models, +4.01% with SDD-SMOTE$_{incrCC}$, +4.07% with Gamma-SMOTE$_{incrCC}$, and +7.79% with G-SMOTE$_{incrCC}$ while it decreased by 3.21% with SMOTE$_{incrCC}$. G-SMOTE$_{incrCC}$ showed the highest improvement across all metrics while sensitivity decreased by 3.21% with SMOTE$_{incrCC}$. Classifiers trained on data generated by SMOTE variants and incremental SMOTE methods outperformed those trained on data by TVAE, with sensitivity improvements ranging from 6.08% to 17.5%.

5 Conclusions

In medical applications where predicting rare disease events or exacerbations is the main concern, class imbalance can affect the accuracy of prediction models. This study systematically evaluated various SMOTE variants and proposed an incremental data generation system to enhance these variants, which generate better-quality training data. We compared our incremental SMOTE methods to the original SMOTE variants using real asthma patients' datasets with four conventional and TL-based classifiers. The findings demonstrate that the incremental SMOTE methods improved prediction accuracy by 4.01% to 7.79% in the sensitivity of TL models using three SMOTE variants. However, with very few minority samples, it may not generate enough non-duplicate data points to balance the dataset. Open challenges include developing flexible, scalable SMOTE variants robust to different imbalanced ratios and data sizes, and TL architectures with other classifiers.

Acknowledgement. This study received support from United Arab Emirates University under UAEU NFRP grant (Grant No. G00004281) and Seattle University.

References

1. 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)
2. Chawla, N.V., Lazarevic, A., Hall, L.O., Bowyer, K.W.: SMOTEBoost: improving prediction of the minority class in boosting. In: Lavrač, N., Gamberger, D., Todorovski, L., Blockeel, H. (eds.) PKDD 2003. LNCS (LNAI), vol. 2838, pp. 107–119. Springer, Heidelberg (2003). https://doi.org/10.1007/978-3-540-39804-2_12
3. Gretel: Gretel. https://gretel.ai/. Accessed 4 May 2024
4. Hoens, T.R., Chawla, N.V.: Imbalanced datasets: from sampling to classifiers. Foundations, Algorithms, and Applications. Wiley, Imbalanced Learning (2013)
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. MIT: The synthetic data vault. https://sdv.dev. Accessed 4 May 2024
8. Raghupathi, W., Raghupathi, V.: Big data analytics in healthcare: promise and potential. Health Inf. Sci. Syst. **2**(1), 1–10 (2014)
9. Sağlam, F., Cengiz, M.A.: A novel smote-based resampling technique trough noise detection and the boosting procedure. Expert Syst. Appl. **200**, 117023 (2022)
10. Sharma, A., Singh, P.K., Chandra, R.: SMOTified-GAN for class imbalanced pattern classification problems. Ieee Access **10**, 30655–30665 (2022)
11. 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)
12. 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)
13. 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)