Generation and Utilization of Synthetic Data to Enhance Type I and Type II Error Resistance for Small Imbalanced Datasets

Jason Orender, Matthew McCombs
Frontier Technology, Inc.
Chesapeake, VA
jorender@FTI-net.com, mmccombs@FTI-net.com

Ralitsa Maduro, Brock Spencer
Naval Safety Command
Norfolk, VA
ralitsa.s.maduro.civ@us.navy.mil,
brock.a.spencer.civ@us.navy.mil

ABSTRACT

Predictive analytics for Naval safety data often involves small and imbalanced sets. To cope with categorical data sets of this nature, the usual remedies often include replicating or oversampling the minority class to bring the numbers into balance, trimming the larger categories to achieve the desired balance or simply choosing a modeling method that is less sensitive to the imbalance. However, by creating a synthetic version of the data based on the characteristics of the original set, the new problems created by oversampling or using less data can be avoided. A more diverse set of data can ultimately be used in model training that will allow use of a broader range of modeling options and marginally increase the generalizability of the resultant solution. This paper will walk through a method for constructing a synthetic data set using an adversarial generative technique which generates a data set of arbitrary size with the same apparent modeling characteristics of the original set. This generated data can also be used for other purposes aside from directly modeling, including as a highly effective anonymization technique for the purpose of collaboration with other researchers while using sensitive data sets.

ABOUT THE AUTHORS

Jason Orender is a Senior Advisory Data Scientist at Frontier Technology, Inc., and he works on the Naval Safety Command's data science team as a technical lead. He retired in 2014 from the US Navy as a Nuclear Power Program Designated Surface Warfare Officer. He has since acquired his Master of Science in Computer Science and is a PhD candidate at Old Dominion University.

Ralitsa Maduro is an Operations Research Analyst/Data Scientist at the Naval Safety Command with prior work as an Enterprise Analytics Consultant and a Biostatistician for Sentara. Additionally, she is an Adjunct in Virginia Wesleyan University's Department of Psychology. Dr. Maduro holds a PhD and Master of Science in Applied Experimental Psychology from Old Dominion University, a Master of Science in Clinical Psychology from Francis Marion University, and a Bachelor of Arts in Psychology from Stockton University.

Matthew McCombs is a Senior Data Scientist with Frontier Technology Inc., supporting the Naval Safety Command, Department of the Navy. He was the contract team's first data scientist in 2018. He led the development of the initial data modeling pipeline for the project and has since supported the development of aviation in-flight mishap models. He graduated with a Bachelor of Science in Pure and Applied Mathematics from Virginia Commonwealth University in 2014 and will be graduating from Georgia Institute of Technology in May 2023 with a Master of Science in Analytics.

Brock Spencer is a Data Scientist/Operations Research Analyst with over a decade of experience in data analytics. Throughout his career, he has developed a strong foundation in data analysis, statistical modeling, and machine learning, utilizing these skills to tackle complex problems in the industry.

Generation and Utilization of Synthetic Data to Enhance Type I and Type II Error Resistance for Small Datasets

Jason Orender, Matthew McCombs
Frontier Technology, Inc.
Chesapeake, VA
jorender@FTI-net.com, mmccombs@FTI-net.com

Ralitsa Maduro, Brock Spencer
Naval Safety Command
Norfolk, VA
ralitsa.s.maduro.civ@us.navy.mil,
brock.a.spencer.civ@us.navy.mil

INTRODUCTION

The problem of too little data is not uncommon in the data science space. While a small data set may be an incomplete picture of the data, some utility can still be extracted while maintaining the integrity of the training process. Much as the halftone printing process (see Fig. 1) is used by the print industry to approximate a full resolution picture (Fresener, 2018), the model fitting process can benefit from filling in the gaps between sparse data and approximating a model with marginally greater generalization powers based on this new and expanded data set (Douzas et. al., 2022).
Since a specialized modeling technique does not have the interpretive powers of the human mind, the added step of filling in the problem space must necessarily occur beforehand. To accomplish that task in this work, we use a constrained Monte Carlo method to propose new examples and then test those examples against a discriminator model initially based on the original data. An example that meets the acceptance criteria is then added to the data set and the process continues, gradually filling in the problem space with consistent data. This is termed an adversarial process because one routine is used to generate the new examples and another algorithmically distinct routine is tasked with detecting whether or not the new examples are members of the original data set (Goodfellow et. al., 2020). Both routines start out from a base state constructed using the original data.

The problem is also defined by the following criteria:

1. The synthetic data produced must be able to generate a model which correctly classifies the original data.
2. The generation process must be reasonably computationally efficient for both positive and negative examples.
3. The process should be capable of producing large numbers of examples for all categories.

This paper will cover the process using a binary classification model for simplicity and ease of understanding, but the same strategy could be applied to other model types.

Related works.

Previous work in this area spans several decades (Niyogi and Poggio, 1998) and was known in the beginning as “Virtual Sample Generation” (Rubin, 1993) with the most common methods utilizing fuzzy set theory (Zimmerman, 2010) then later extended by diffusion theory (Chongfu, 1997) to application in diffusion-neural-networks (Huang and Moraga, 2004). The problem with these methods is that they view the independent variables in isolation and largely ignore the relationships between them. While many features may be “independent” in the sense that they are not correlated, certain combinations of feature

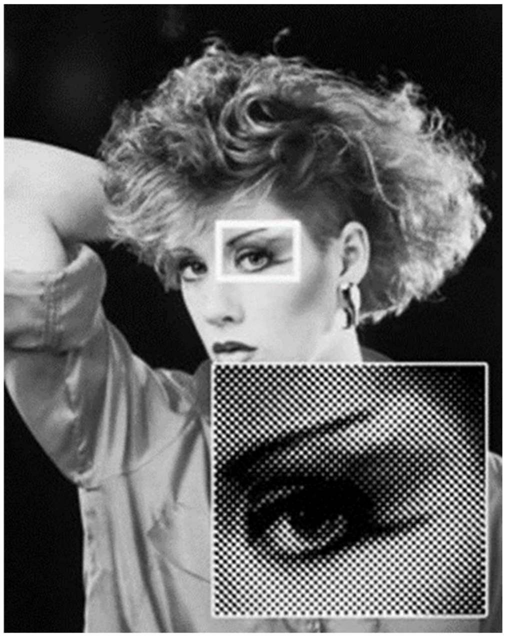


Figure 1. A halftone printing example. Since the picture of the eye is incomplete (about half of the dots are missing), the viewer’s brain must interpret what the picture is intended to represent, and it can in fact easily distinguish that this is an eye. The mind interprets the dots as a coherent picture even as half of the information is missing (Fresener, 2018).

values may be rare or impossible to occur within the system that generated the features and other combinations may be more common than others. Despite these shortcomings, and with the help of additional improvements such as employing genetic algorithms to mutate the most feasible proposed examples (Li and Wen, 2014), these methods as a group still performed better than cases in which no additional synthetic example generation was done (Lin and Li, 2011).
The motivation for many of the advances in this space has been the desire for dissemination of functionally equivalent data to data that is either proprietary or is beholden to privacy considerations that prevent its unrestricted distribution (Nowok et. al, 2016). However, the desired final result is still the same: a data set with the same modeling and inference characteristics as the original (Kinney, et. al., 2011).

The current state of the art for synthetic data generation extends the method such that the synthetic data mirrors a joint distribution which is defined in terms of a series of conditional distributions for each of the parameters in the data set (Nowok et. al., 2016), coupled with an objective function type which governs the modeling characteristics of the generated data. This is known as the “normrank” method.

Contributions.

The contribution of this paper is introduction of a novel adversarial technique which may improve upon the standard methods of synthetic data generation for certain data sets; this will be shown by the performance of controlled experiments on engineered data for which the characteristics are known.

Paper Organization.

This paper is organized into five parts. After the introduction section, a background section will review some of the necessary material required to understand the experiments and techniques as well as the context of the discussion. Following this, the implementation of the code used for the experiments will be reviewed. The results section discusses the significance of the experimental outcomes. Finally, the conclusion section will summarize the paper and consider a way forward for this research.

BACKGROUND

The purpose of this paper is to evaluate the adversarial process as an enhancement to synthetic data generation. To that end, this paper uses the well-known “synthpop” package to generate the synthetic data, and it in turn utilizes the state-of-the-art normrank method described in the introduction. According to the package documentation (Nowok et. al., 2016), normrank generates univariate synthetic data using linear regression analysis and preserves the marginal distribution. This method first generates synthetic values of normal deviates of ranks of the values in y using the spread around the fitted linear regression line of normal deviates of ranks given x. Then synthetic normal deviates of ranks are transformed back to get synthetic ranks which are used to assign values from y. For proper synthesis, first the regression coefficients are drawn from normal distribution with mean and variance from the fitted model.

The Least Absolute Shrinkage and Selection Operator (LASSO) is the analysis method used as the basis for comparison. It is a well-known regularized regression method which performs variable selection that reduces complexity and prevents overfitting (Tibshirani, 1996). It was originally designed for linear regression, but it can be expanded to a wide variety of other modeling problems. This includes logistic regression, which will be utilized for the binary classification problem presented in this paper; this is accomplished by estimating the parameters of the binomial generalized linear model (GLM) by optimizing the binomial likelihood (i.e. log odds) while imposing the LASSO penalty (λ) on parameter estimates (Shalizi, 2013). The cost function to be minimized is:

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{1}$$

The implementation in the “glmnet” R package will be used for application of the LASSO algorithm to the data presented in this paper. Since the objective of this paper is to evaluate the utility of the adversarial refinement method presented, a well-known implementation of a well-known algorithm was used as the basis for comparison.

The left-hand component of expression 1 is the sum of squared errors formula from traditional linear regression. The right-hand component is the penalty (regularization) term introduced by LASSO. It calculates the L1 norm (i.e. the sum of the absolute values) of the model coefficients and multiplies it by λ, which is a penalization constant which is typically modified as a tuning parameter. If $\lambda = 0$, then no penalty is applied, and the problem is equivalent to traditional linear regression. As λ increases, the penalty towards larger coefficients increases, which therefore shrinks (i.e. regularizes) the coefficients in the optimal solution. As λ increases, the bias increases and the variance decreases.

Adversarial networks are a class of machine learning frameworks where two neural networks contest with each other such that one network's gain is the other network's loss (Goodfellow et. al., 2020). The adversarial process described in this paper draws inspiration from these networks in the way that two functions are employed: a generator and a discriminator, which are locked in a battle in which each is trying to outdo the other. They pass a data set back and forth, which the other is obligated to use to either generate additional examples (the generator) or figure out which examples are real, and which are false (the discriminator). Using a simple set of functions for this works well and is significantly less computationally intensive than a generative adversarial network (GAN) which performs the analogous process using neural networks.

IMPLEMENTATION

Imbalanced data sets can significantly impact a model's accuracy, leading to a bias towards the majority class (Kotsiantis et. al., 2006). Algorithms can generally overcome this issue by using a data set balanced with synthesized data. As stated in the introduction, the below method will improve upon the standard methods of synthetic data generation for data sets, as shown by the performance of controlled experiments on engineered data.

The general method of the adversarial generative technique can be summarized as occurring in two steps. First, the generator tries to create examples that are indistinguishable from the original examples. Then on its turn, the discriminator tries to determine whether the example is original or synthetic using a model generated from the previous set of passed data.

The adversarial algorithm accomplishes these tasks by: 1) having a generator produce examples based on a data set of original examples and the selection of synthetic examples which previously made it through an iteration, 2) having the discriminator test the new examples against a model based on the previous cycle data set which the discriminator updates over each iteration, 3) If an original member of the real data set is misclassified, adding back in another copy of those examples back into the data set, thereby multiplying the number of copies of that example and reinforcing those particular examples in the data set. This makes those examples less likely to be misclassified again during the next iteration in addition to acting as a counteragent and anchor to prevent model drift.

This paper focuses on employing the adversarial method for a binary classification model with variations in the number of iterations, class weights, and sample sizes to evaluate the efficacy of synthetic data to improve the results beyond those that would be accomplished without synthetic data or with synthetic data but without the adversarial generation and discrimination process.

Once the binary classification model was trained, threshold shifting was used to get better evaluation metrics. Threshold shifting involved altering the decision threshold of the model and observing the effects on evaluation metrics such as Youden's Index, F1-Score, Accuracy, Balanced Accuracy, and Matthews Correlation Coefficient. Shifting the threshold can tune an algorithm to optimize the evaluation metric of interest. This helps to ensure that the model is making accurate and reliable predictions. The Youden's index was focused on in the results section as it is a widely utilized simple metric which is well understood and adequately highlights the difference in performance regardless of how balanced or imbalanced the data set is.

The performance is assessed by evaluating the results from each iteration. With each additional iteration of the algorithm, there is expected to be an improvement in the evaluation metric. However, the number of examples for which this is true declines with each additional iteration. In the other cases, the result stayed the same or even deteriorated.

At the conclusion of each iteration the effectiveness of the adversarial generative technique can be judged. If the result of each iteration is incrementally better than the previous one, then the algorithm is understood to be performing effectively. On the other hand, if the result of each iteration is not better than the previous one, then it is understood that the algorithm is not performing in a way that enhances the results.

This method was implemented in the R programming language using the “synthpop” and “glmnet” libraries. The “synthpop” library encapsulates the state-of-the-art methods for synthetic data generation described in the introduction and therefore provides a well-established starting point from which to assess improvement. The “glmnet” library is an industry standard which is used to fit several different model types, including using the LASSO. The adversarial generative technique is illustrated in Fig. 2.

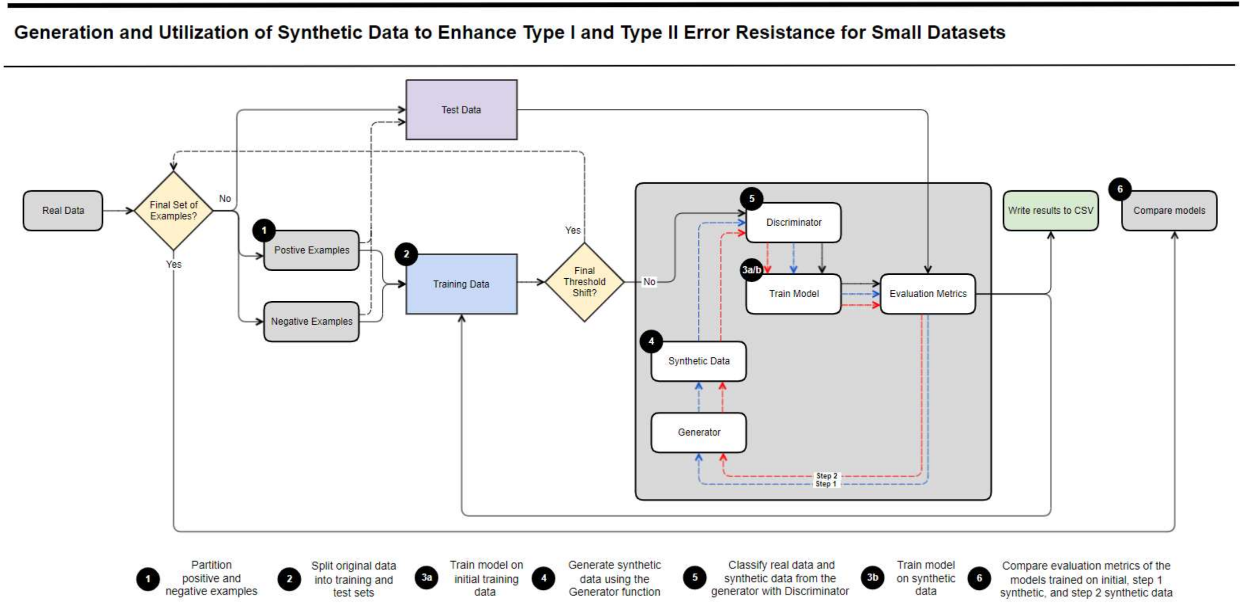

Generation and Utilization of Synthetic Data to Enhance Type I and Type II Error Resistance for Small Datasets
Test Data
Real Data
Final Set of Examples?
No
Yes
1
Postive Examples
Negative Examples
2
Training Data
Final Threshold Shift?
Yes
No
5
Discriminator
3a/b
Train Model
Evaluation Metrics
4
Synthetic Data
Generator
Step 2
Step 1
Write results to CSV
6
Compare models
1
Partition positive and negative examples
2
Split original data into training and test sets
3a
Train model on initial training data
4
Generate synthetic data using the Generator function
5
Classify real data and synthetic data from the generator with Discriminator
3b
Train model on synthetic data
6
Compare evaluation metrics of the models trained on initial, step 1 synthetic, and step 2 synthetic data

Figure 2. Illustration of adversarial generative technique

RESULTS

A figure representing the amount of execution time in seconds is presented in Fig. 3. The proportion of positive examples in the sample did not significantly impact the execution time, while number of iterations had a large impact, making each next iteration significantly slower; this is due to the increase in the number of synthetic examples. The growth rate measured was significantly greater than exponential, creating a significant increase in computational load for every new iteration. The execution times up to the third iteration were reasonable on consumer hardware, but more than three iterations is not likely feasible without modifications to the process or without access to computational power much greater than on most consumer devices. As mentioned previously, this increase in execution time is due primarily to the growth in the synthetic data size from iteration to iteration and therefore will probably be machine dependent; those with greater memory will likely perform better as the number of synthetic examples increases.

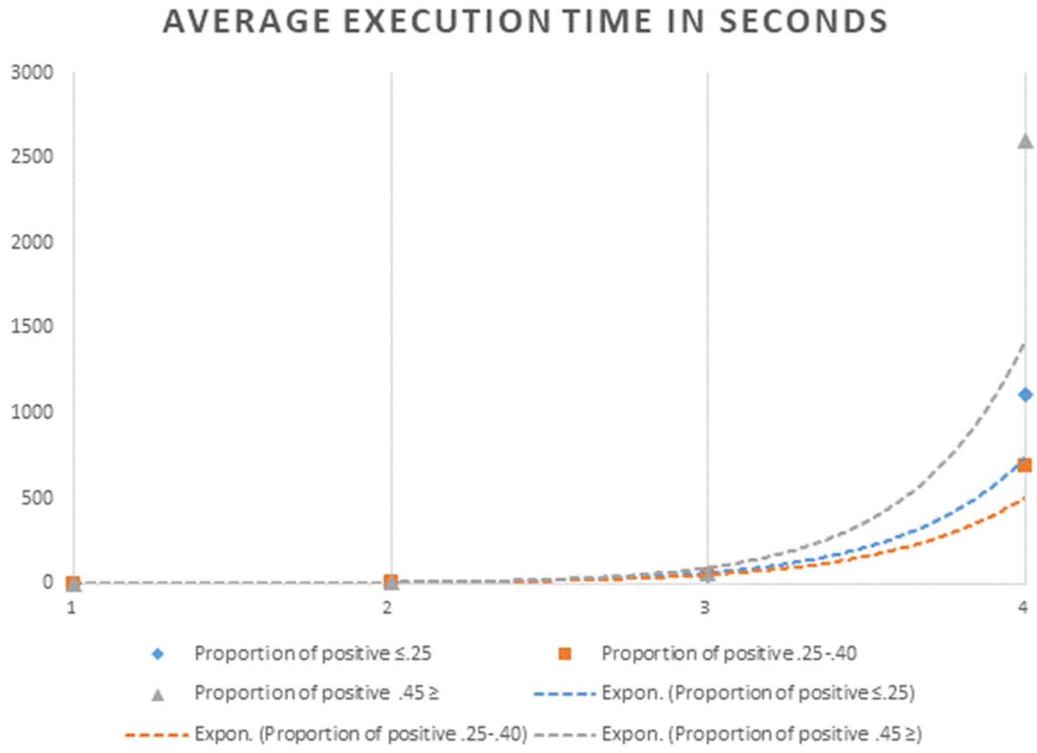


Figure 3. Algorithm execution time. A greater than exponential relationship exists with respect to increasing iterations due to the growth in the number of synthetic examples. The dotted curves show what an exponential fit would look like; note that the data points are universally above their respective curves. There are three data sets shown with each a different proportion of positive examples.

We examined the effect of various iterations and proportions of positive examples on Youden Index point estimates (referred to as the 'J' Index) from each iteration relative to its preceding iteration (the Youden Index scale ranges −1 to +1). Importantly, we also observed the change between each iteration in terms of Youden's index gain or loss. Our analysis indicates that the Youden Index scores were optimized in 27 of 30 best performing experiments after the first iteration. The index continues to improve for 15 out of 30 experiments after the second iteration. Lastly, 5 out of the 14 experiments ran at iteration 3 showed improvement. We used a Kruskal-Wallis non-parametric method for testing whether there was a statistically significant difference in Youden's Index between each iteration (Kruskal-Wallis chi-squared = 13.60, df = 3, p < .001). Pairwise comparisons support a recommendation that the improvement in Youden's Index *past the second iteration was not significant* (Fig. 4).

There are some other important broad trends as well. As number of data points increased, the likelihood of an increase in Youden's index from iteration to iteration dramatically increased as well; in conjunction with this, for a low number of initial data points the spread between the highly imbalanced to highly balanced Youden's index metrics increased, though this effect diminished to marginal amounts with just a few hundred added examples. In addition, the level of improvement in Youden's index was generally much greater the *more imbalanced* the data set was to begin with (Fig. 5). In fact, it is apparent that the synthetic data generation process creates a near parity between imbalanced sets with initially as little as 10% positive examples and balanced sets with 50% positive examples despite a much poorer initial performance without the added synthetic data.

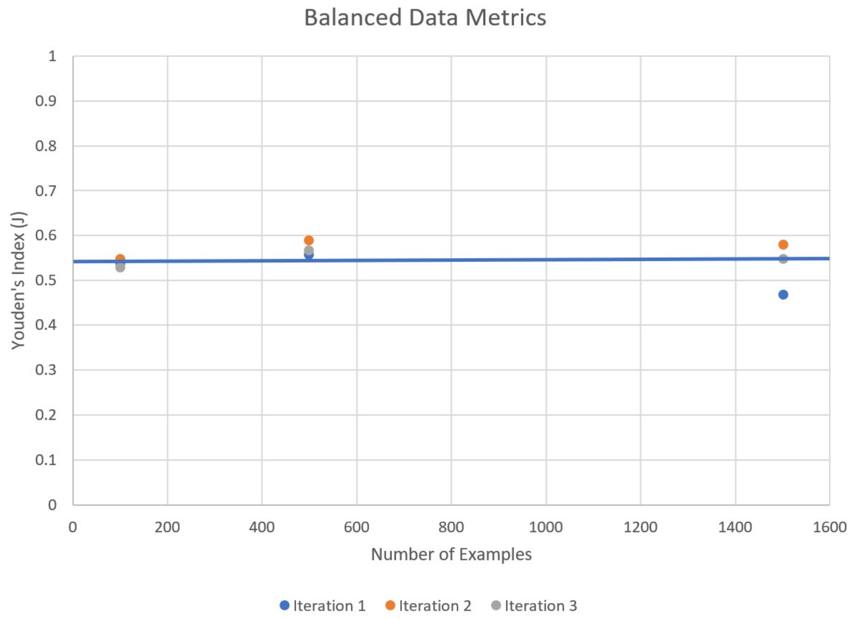


Figure 4. Youden's Index For all Iterations and Data Set Sizes for Balanced Sets. The balanced data Youden's Index (J) for this data set is about 0.55 at all training set sizes and throughout the iterative process, which is as expected since the synthetic data is not creating additional fidelity beyond what would be expected of the balanced data set. It is merely helping the modeling algorithms compensate for a lack of ability to cope with the data set imbalance.

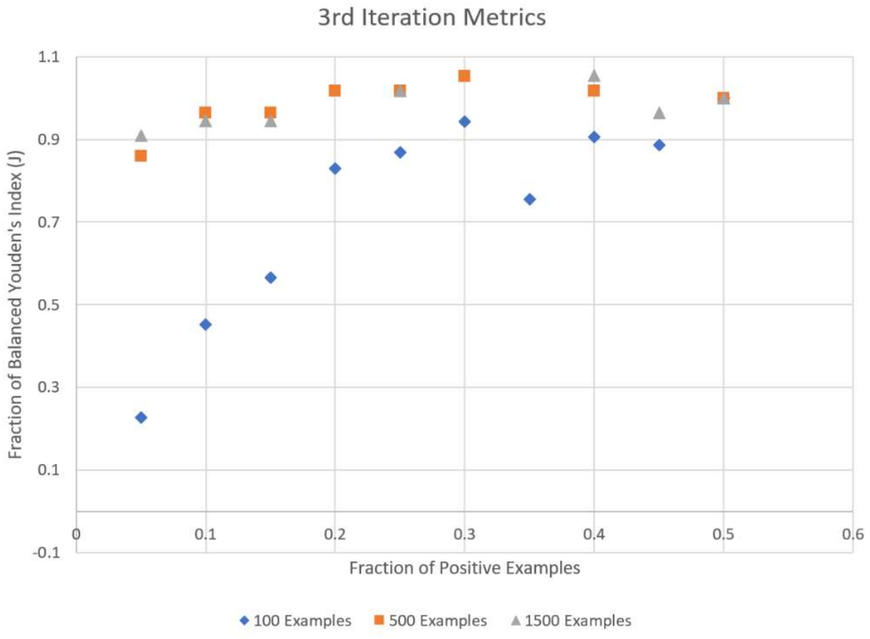
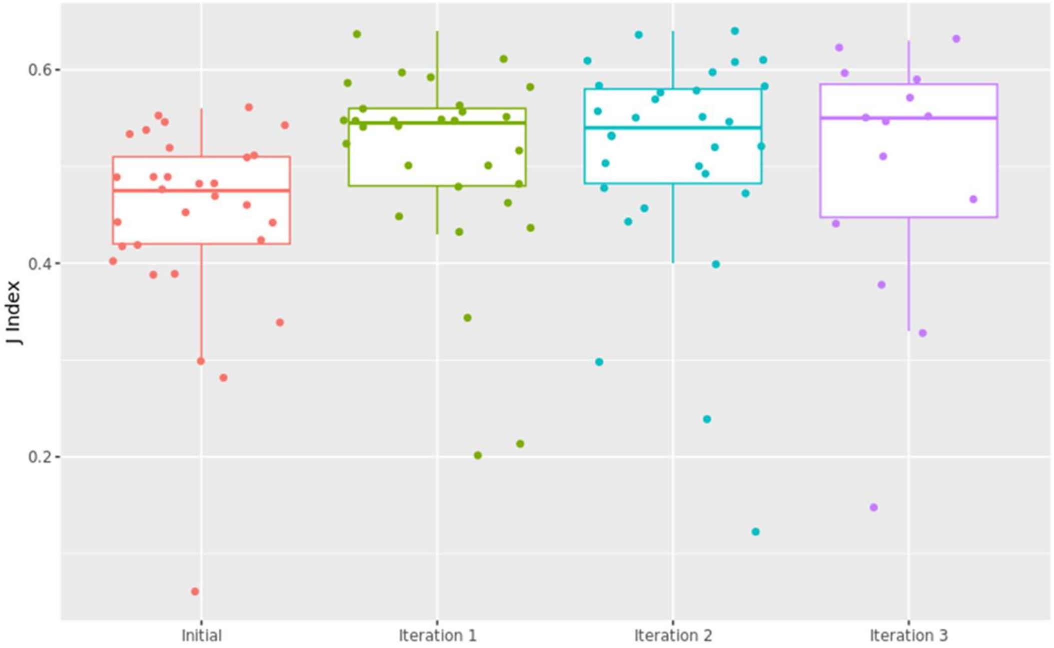


Figure 5. Expected Improvement as a Function of Imbalance. The performance gain from generation of synthetic data that can be expected is greatest when the imbalance is also the greatest. This is what would be anticipated since the greater imbalance would pose a greater difficulty for most modeling algorithms to cope with.

Notably, the variation in performance for balanced sets, regardless of data set size or number of iterations, was very small (see Fig. 4). This would be expected since the synthetic data does not provide any new information, it merely fills in the gaps between the existing examples such that the ability to apply a wider variety of modeling algorithms is enhanced. This allows data set imbalances to potentially become irrelevant, or nearly so, since any effect from the imbalance may be compensated for by generating the appropriate amount of synthetic data before modeling.

Figure 6. Kruskal-Wallis Rank Sum Test Comparing Youden's Index Values for each Iteration.



CONCLUSIONS

The adversarial process presented in this paper improves upon the already significant effect of synthetic data generation as a means to reduce both type I and type II errors in small imbalanced data sets. Further, several rules of thumb are proffered as takeaways:

- Increasing total data points available increases efficacy regardless of how balanced the data set is.
- Highly imbalanced data sets show much more improvement than balanced sets, making this process a way to potentially equalize the outcomes of imbalanced vs. balanced data sets.

In addition, this process almost always works for at least one complete generation and discrimination cycle. One complete iteration has a 90% success rate at improving model effectiveness overall and a 100% success rate for all experiments with 500 examples or greater.

Alternative applications.

Besides using the generated data as a means to refine model development, and because this data behaves like the original in aggregate but creates unique examples sufficiently different from the original data, it can be used in applications which require anonymity but are of such importance that a desire to share research resources with those outside the bounds of the organization that generated the proprietary or privacy sensitive data is desired. Using this adversarial process to improve upon the quality of the generated data to behave more like the original creates a greater likelihood that the shared generated data will lead to useful insights from third party researchers.

Way Forward

Limiting the size of the data set growth by being more selective regarding which examples are retained may result in increasing the number of iterations in which the metrics are enhanced as well as the quality of the synthetic data set as a whole. Devising an acceptable method or objective function to enable this extra selection step is the next logical step in improving this generation method.

ACKNOWLEDGEMENTS

Thanks to the Naval Safety Command leadership for encouraging this work and adopting a forward-looking attitude in which analyses such as this one are employed to preserve national assets and save lives.

REFERENCES

Chongfu, H. (1997). Principle of information diffusion. Fuzzy sets and Systems, 91(1), 69-90.

Douzas, G., Lechleitner, M., & Bacao, F. (2022). Improving the quality of predictive models in small data GSDOT: A new algorithm for generating synthetic data. Plos one, 17(4), e0265626.

Fresener, S. (2018, June 13). Halftone Dots Made Easy. T-Biz Network International. Retrieved January 11, 2023, from https://t-biznetwork.com/articles/screenprinting/halftone-dots-made-easy/

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2020). Generative adversarial networks. Commun. Acm, 63(11), 139-144.

Huang, C., & Moraga, C. (2004). A diffusion-neural-network for learning from small sam-ples. International Journal of Approximate Reasoning, 35(2), 137-161.

Kinney, S. K., Reiter, J. P., Reznek, A. P., Miranda, J., Jarmin, R. S., & Abowd, J. M. (2011). Towards unrestricted public use business microdata: The synthetic longitudinal business database. International Statistical Review, 79(3), 362-384

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets: A review. GESTS international transactions on computer science and engineering, 30(1), 25-36.

Li, D. C., & Wen, I. H. (2014). A genetic algorithm-based virtual sample generation tech-nique to improve small data set learning. Neurocomputing, 143, 222-230.

Lin, Y. S., & Li, D. C. (2011). The generalized-trend-diffusion modeling algorithm for small data sets in the early stages of manufacturing systems. Quality control and applied statistics, 56(5), 505-507.

Niyogi, P., Girosi, F., & Poggio, T. (1998). Incorporating prior information in machine learning by creating virtual examples. Proceedings of the IEEE, 86(11), 2196-2209.

Nowok, B., Raab, G. M., & Dibben, C. (2016). synthpop: Bespoke creation of synthetic da-ta in R. Journal of statistical software, 74, 1-26.

Rubin, D. B. (1993). Statistical disclosure limitation. Journal of official Statistics, 9(2), 461-468.

Shalizi, C. (2013). Advanced data analysis from an elementary point of view.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.

Zimmermann, H. J. (2010). Fuzzy set theory. Wiley interdisciplinary reviews: computational statistics, 2(3), 317-332.