MSR 73

Synthetic Data Generation for Augmenting Small Samples

Dan Liu PhD, Samer El Kababji PhD, Nicholas Mitsakakis PhD, Lisa Pilgram MD, Thomas Walters MS, Mark Clemons MD, Greg Pond PhD, Alaa El-Hussuna PhD, Khaled El Emam PhD,
Children's Hospital of Eastern Ontario Research Institute, Ottawa, ON, Canada.

BACKGROUND

- Training machine learning (ML) models on small datasets may lead to suboptimal model performance.
- Data augmentation has been used as an effective tool to address the data scarcity problem for imaging and time series data.
- However, there is no evaluation of data augmentation on tabular health data.

STUDY OBJECTIVE

- To evaluate data augmentation using generative models on tabular health data n_0.
- To assess the impact of diversity versus increasing the sample size using data augmentation.

METHODS

- Simulations were performed using 13 large health datasets to evaluate the impact of data augmentation on the prediction performance (measured by AUC) on binary classification gradient boosted decision tree models.
- Four synthetic data generation models, including sequential decision trees, Bayesian networks, conditional GAN and tabular VAE were evaluated.
- A decision support model was established to help analysts determine if augmentation would improve model performance.
- Seven case studies were conducted to illustrate the application of the decision model for augmentation on small datasets and compare the data augmentation with the resampling approach.

DECISION SUPPORT TOOL RESULTS

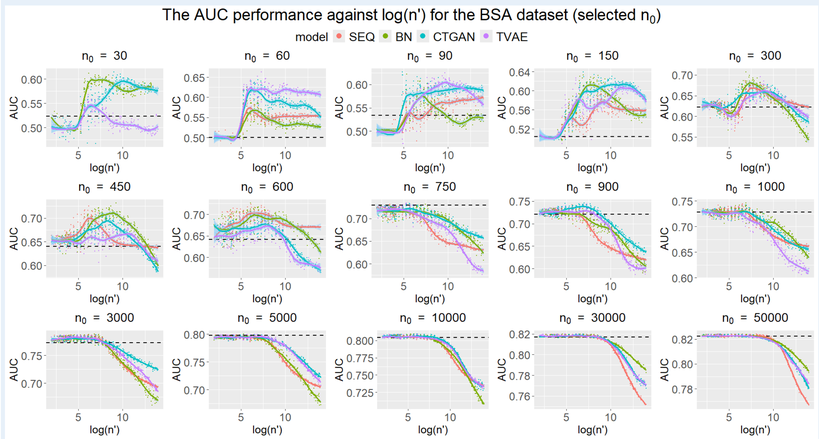
- Logistic regression was chosen as the model to establish the decision support tool. The association between augmentation indicator and relevant data characteristics is summarized as follows.

$$I(\text{augmentation}) \sim 6.75 - 4.79\times10^{-5} n_0 - 4.94\times10^{-2}\,\text{imbalance factor} + 5.12\times10^{-4}\,\text{degrees of freedom} - 7.63\,\text{baseline AUC},$$

where $I(\cdot)$ is the indicator function, n_0 denotes the original sample size, imbalance factor is defined to measure the outcome distribution, degrees of freedom indicates the degrees of freedom for the predictors, and baseline AUC is the AUC obtained from model training on the original data.

- The AUC from the original data was found to have the biggest impact on whether to recommend augmentation.
- Moreover, the datasets that are smaller, more balanced, and more complex with higher cardinality are more likely to benefit from augmentation.

AUGMENTATION RESULTS



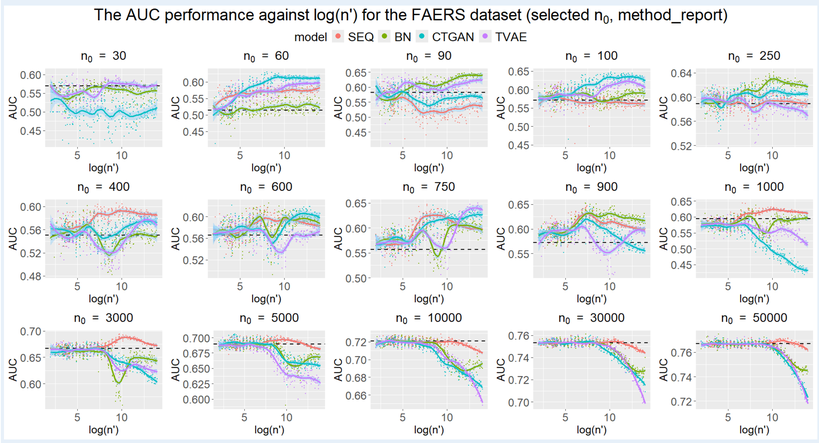


Figure 1. Augmentation performance of AUC against log(*n'*) for the two representative datasets (simple and complex), BSA and FAERS datasets, for a subset of the baseline data sizes. The black dotted line is the baseline AUC for the base dataset of size n_0. n_0: the size of the original data. n': the size of additional data simulated from generative models.

- As more data are incorporated, data augmentation led to instrumental improvements of model performance, especially for small and medium n_0.
- However, for large n_0, augmentation has little benefits and even become detrimental to the model performance.
- FAERS, a more complex dataset than BSA dataset, is more benefited from augmentation in terms of model performance.

CASE STUDY RESULTS

- Diversity is defined as a threshold-based measure to quantify the amount of synthetic data records that are significantly different from the original sample.

Table 1. Analysis results of augmentation performance for the seven small datasets. n'_{max}: n' that leads to maximum AUC. Baseline AUC: baseline AUC from the base data. Augmented AUC: maximum AUC from the augmented data. Resampled AUC: AUC from the augmented data with a size of n'_{max} using resampling with replacement method. Diversity generative: diversity of data augmented using a generative model. Diversity resample: diversity of data augmented using the bootstrap method.

			AUC Results				Diversity Results	
Dataset	**Model**	**n'_{max}**	**Baseline AUC**	**Augmented AUC**	**Relative AUC (%)**	**Resampled AUC**	**Diversity generative**	**Diversity resample**
Hot Flashes	CTGAN	720	0.7161	0.7668	7.08	0.6477	0.0023	0.0013
Danish Colorectal Cancer Group	TVAE	720	0.7171	0.778	8.50	0.7077	0.0000	0.0008
Breast Cancer Coimbra	BN	53	0.7392	0.8722	18.00	0.8291	0.0061	0.0019
Breast Cancer	CTGAN	25	0.7143	0.7451	4.31	0.6729	0.0017	0.0008
Colposcopy/Schiller	CTGAN	2,205	0.5125	0.7341	43.23	0.6116	0.0883	0.0004
Diabetic Retinopathy	BN	11,534	0.74	0.7974	7.75	0.7299	0.1177	0.0002
Thoracic Surgery	TVAE	6,602	0.5584	0.67	19.98	0.6914	0.0000	0.0003

- In general, the model performance has been greatly enhanced after augmentation, with relative AUC ranging from 4.31% to 43.23% (baseline AUC vs augmented AUC: p=0.0078).
- Resampling augmentation does not contribute to the improvement of model performance as much as the other synthetic data generative models (augmented AUC vs resampled AUC: p=0.016).
- Generative models increase the diversity of the datasets compared to a simple increase in the sample size by resampling the original data (generative diversity vs. resampled diversity: p=0.046).

CONCLUSION

- Data augmentation using generative models was demonstrated to provide significant improvements of prognostic model performance of ML models.
- The datasets that are smaller, more balanced, and more complex with higher cardinality are more likely to benefit from augmentation.
- A decision support tool was developed to aid users determine the necessity of the data augmentation for a given dataset.
- Increasing diversity of the datasets is more beneficial than simply increasing the sample size without making it diversified.

REFERENCES AND CONTACT

- Access our preprint via the following QR code →
- Questions? Email: kelemam@ehealthinformation.ca