MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190509

An Adversorial Approach to Enable Re-Use of Machine Learning Models and Collaborative Research Efforts Using Synthetic Unstructured Free-Text Medical Data

Suranga N. Kasthurirathne[a,b], Gregory Dexter[a], Shaun J. Grannis[a,c]

[a] *Center for Biomedical Informatics, Regenstrief Institute, Indianapolis, Indiana, USA*
[b] *Richard M. Fairbanks School of Public Health, Indiana University, Indianapolis, Indiana, USA*
[c] *School of Medicine, Indiana University, Indianapolis, IN, USA*

Abstract

We leverage Generative Adversarial Networks (GAN) to produce synthetic free-text medical data with low re-identification risk, and apply these to replicate machine learning solutions. We trained GAN models to generate free-text cancer pathology reports. Decision models were trained using synthetic datasets reported performance metrics that were statistically similar to models trained using original test data. Our results further the use of GANs to generate synthetic data for collaborative research and re-use of machine learning models.

***Keywords*:**

Neural Networks (Computer); Machine Learning; Dataset

Introduction

Large scale adoption of Health Information Systems (HIS), together with the rapid evolution of Artificial Intelligence (AI) and various analytical and machine learning toolkits have led to the widespread development of machine learning solutions to address various healthcare challenges using patient data. However, legislation on sharing of Patient Health Identifiers (PHI) restricts researchers from (a) re-using machine learning solutions across larger audiences, (b) fostering inter-organizational collaboration addressing various healthcare challenges, and (c) building generalized machine learning models targeting larger, diverse populations.

Current efforts to enable better data sharing focus on de-identification efforts, where PHI is scrubbed from patient data. However, de-identified free-text data may be vulnerable to re-identification based on clinical data elements. In contrast, alternate approaches to create synthetic data that mimic clinical patterns in data present considerably lower re-identification risk [1]. We leverage recent advances in Generative Adversarial Networks (GAN) [2] to produce synthetic unstructured free-text medical data, and assess (a) possibility of using these datasets to replicate machine learning results generated using original patient data, and (b) levels of re-identification risk posed by these synthetic datasets.

Methods

We leveraged a convenience sample of 7,000 free-text pathology reports on potential cancer cases from the Indiana Network for Patient Care (INPC) [3], a statewide Health Information Exchange (HIE) to build decision models capable of identifying positive cancer cases for public health reporting. Positive and negative report sets were extracted, and used to train SeqGAN [2] models of varying epoch sizes. We selected optimal GAN models for positive and negative cancer report sets by comparing synthetic data generated by these models with original test data using Bilingual Evaluation Understudy (BLEU) scores. We created vectors by counting presence of each stemmed feature in positive and negative contexts across each report in the report set. Next, we developed decision models to predict cancer cases using the Random Forest classification algorithm [4] and the top 5, 10, 20 and 50 features selected from the original test and synthetic feature sets using the Kullback-Leibler divergence (information gain) method [5].

We compared the performance of these models using sensitivity, specificity, F1-measure (harmonic mean between precision and recall) and area under the ROC curve values (AUC), together with their 95% confidence intervals. We assed re-identification risk for presence disclosure [6], where attackers in possession of a set of patient records can determine if any of them were used to train GAN models by comparing these records against the synthetic patient dataset using Hamming scores, a measure of variation between two binary strings.

Results

The 7,000 free-text cancer cases consisted of 1,950 (27.86%) positive reports and 5,050 (72.14%) negative reports [7]. Comparison of BLEU scores identified models trained for 70 epochs as the optimal synthetic data generation model for both positive and negative cancer reports. We extracted the top 50 features from each of the original and synthetic datasets using information gain scores (*Table 1*). Feature selection identified a 36% overlap between the top 50 features extracted from each dataset (*Table 2*).

Figure 1 presents variance of information gain scores across each of the top 50 feature sets. Decision models trained using the top 5, 10, 20 and 50 features extracted from the synthetic and original datasets reported performance metrics that were significantly high (sensitivity: 77-92%, specificity: 95.7-99.8%, F1-measure: 91-97%, AUC: 90-99%). Further, there was no statistically significant difference between many performance metrics reported by models trained using original or synthetic datasets. Presence disclosure tests performed using Hamming score comparisons indicated relatively low probability for re-identification across positive and negative synthetic reports.

Table 1 - List of top 20 stemmed features selected from the original and synthetic datasets using information gain scores.

Rank	Original dataset	Synthetic dataset
1	Tumor	consult
2	Carcinoma	slide
3	Invasion	node
4	Slide	lymph
5	Cell	malign
6	Metastat	grade
7	Lymph	right
8	Node	pathologist
9	Return	submit
10	adenocarcinoma	prostat
11	Margin	collect
12	involve	carcinoma
13	Consult	section
14	differenti	receiv
15	mass	left
16	cassett	specimen
17	phone	posterior
18	left	surgic
19	grade	identifi
20	microscop	tumor

Table 2 - Intersection of top 5, 10, 20 and 50 features selected from the original and synthetic datasets using information gain scores.

Feature subset size	# features overlap	List of features present in both datasets
5	1 (20%)	slide
10	3 (30%)	slide, lymph, node
20	8 (40%)	slide, consult, node, grade, tumor, lymph, left, carcinoma
50	18 (36%)	note, section, prostat, cassett, slide, malign, consult, node, grade, tumor, right, margin, phone, submit, case, lymph, carcinoma, left

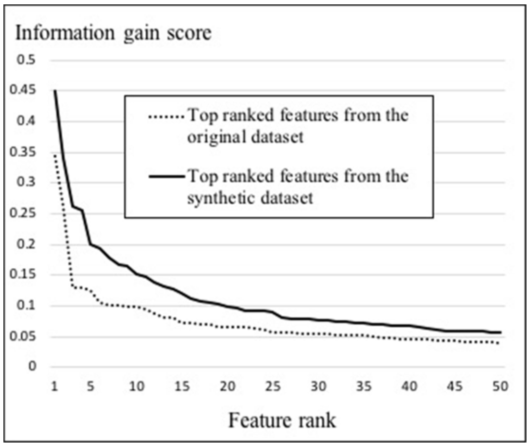

Figure 1 - Variance of information gain scores reported by the top 50 original and synthetic features.

Conclusions

Our results indicate that GAN methodologies can generate synthetic free-text medical data with limited re-identification risk, and that synthetic datasets can be used to develop machine learning models with statistically similar performance metrics to decision models trained using original test data. As such, they are of considerable importance for enabling cross-institutional collaboration and broader dissemination of machine learning models. Adoption of GAN models alone does not result in de-identified data. However, synthetic data generation reduces re-identification risk by creating new patient records with similar, but different data. It also removes any 1-to-1 mapping between test and synthetic reports. Our results demonstrate that synthetic datasets poise a significantly lower chance of re-identification based on clinical information. However, synthetic data produced by these efforts must undergo rigorous de-identification of PHI elements before they can be distributed for public use. Future research includes use of GAN models to create truly de-identified synthetic free-text data that does not require additional de-identification, and expansion of our work across other more challenging healthcare datasets.

We propose the following hypothetical scenario to demonstrate how our approach could be applied in a real-life setting. An organization that possesses rich free-text data sources, but lacks adequate machine learning expertise can leverage our approach to create synthetic data. They de-identify and share the synthetic data with experts who use it to build machine learning models. Once optimal models have been identified, they can be implemented across the original dataset with compatible performance measures.

References

[1] E. Choi, S. Biswal, B. Malin, J. Duke, W.F. Stewart, and J. Sun, Generating multi-label discrete patient records using generative adversarial networks, *arXiv preprint arXiv:1703.06490* (2017).

[2] L. Yu, W. Zhang, J. Wang, and Y. Yu, SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient, in: *AAAI*, 2017, pp. 2852-2858.

[3] C.J. McDonald, J.M. Overhage, M. Barnes, G. Schadow, L. Blevins, P.R. Dexter, B. Mamlin, and I.M. Committee, The Indiana network for patient care: a working local health information infrastructure, *Health affairs* **24** (2005), 1214-1220.

[4] L. Breiman, Random forests, *Machine learning* **45** (2001), 5-32.

[5] D. Polani, Kullback-leibler divergence, *Encyclopedia of Systems Biology* (2013), 1087-1088.

[6] M.E. Nergiz and C. Clifton, δ-presence without complete world knowledge, *IEEE Transactions on Knowledge and Data Engineering* **22** (2010), 868-883.

[7] S.N. Kasthurirathne, B.E. Dixon, J. Gichoya, H. Xu, Y. Xia, B. Mamlin, and S.J. Grannis, Toward better public health reporting using existing off the shelf approaches: A comparison of alternative cancer detection approaches using plaintext medical data and non-dictionary based feature selection, *Journal of biomedical informatics* **60** (2016), 145-152.

Address for correspondence

Suranga N. Kasthurirathne, PhD; E-mail: snkasthu@iu.edu