Generation of Heterogeneous Synthetic Electronic Health Records using GANs

Generation of Heterogeneous Synthetic Electronic Health Records using GANs

Kieran Chin-Cheong
Computer Science Dept.
ETH Zürich, Switzerland
kchincheong@ethz.ch

Thomas Sutter
Computer Science Dept.
ETH Zürich, Switzerland
suttetho@inf.ethz.ch

Julia E. Vogt
Computer Science Dept.
ETH Zürich, Switzerland
julia.vogt@inf.ethz.ch

Abstract

Electronic Health Records (EHRs) are commonly used by the machine learning community for research on problems specifically related to health care and medicine. EHRs have the advantages that they can be easily distributed and contain many features useful for e.g. classification problems. What makes EHR data sets different from typical machine learning data sets is that they are often very sparse, due to their high dimensionality, and often contain heterogeneous data types. Furthermore, the data sets deal with sensitive information, which limits the distribution of any models learned using them, due to privacy concerns. In this work, we explore using Generative Adversarial Networks to generate synthetic, *heterogeneous* EHRs with the goal of using these synthetic records in place of existing data sets. We will further explore applying differential privacy (DP) preserving optimization in order to produce differentially private synthetic EHR data sets, which provide rigorous privacy guarantees, and are therefore more easily shareable. The performance of our model's synthetic, heterogeneous data is very close to the original data set (within 6.4%) for the non-DP model. Although around 20% worse, the DP synthetic data is still usable for machine learning tasks.

1 Introduction

Data sets used by the machine learning community for research on problems related to health care and medicine are often based on Electronic Health Records (EHRs). These records contain private details about patients visits to hospitals or health-care facilities, and in particular usually consist of heterogeneous administrative data (such as patient age, weight, or length of stay), and diagnostic data (such as associated ICD codes for both diagnoses made and procedures carried out). The administrative data is typically dense, while the diagnostic data is typically very sparse. Together, this results in a data set that is heterogeneous, high dimensional, and sparse. Examples of such data sets are the freely accessible Mimic-III data set [1] and the New Zealand National Minimal data set [2].

A typical use case for medical data sets is to perform binary classification, using a particular data set feature as a label. One such task is to try to use EHRs to predict the risk of a patient being unexpectedly readmitted to hospital after a discharge. Such unexpected readmissions are in many countries financially penalized, through programs such as the Hospital Readmission Reduction Program in the United States [3], which has cost hospitals $1.9 billion in penalties as of 2016 [4].

Using EHR data to train machine learning models requires some care, however, as the data being used is inherently sensitive. For instance, it is possible to recover training data from models [5]. Furthermore, even though in many cases the training data has been de-identified, it has also been shown that data re-identification is possible via linkage to external data sources [6]. Therefore, there exists a risk to the privacy of patients whose data is present in training data sets.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

One way to avoid these pitfalls is to release models trained using synthetic data sets, which are based on the original training data. Such models might provide better privacy to patients, because training data extraction attacks would then have to further re-identify a synthetic training record. However, to truly provide rigorous privacy guarantees, the synthetic data needs to be generated in a differentially private [7] manner. In this work, we investigate the generation of heterogeneous EHR data for use in returning hospital patient classification. We further investigate the impact of applying differential privacy to EHR generation.

2 Proposed Model

In our proposed model, synthetic EHR data is generated using a GAN [8]. A GAN is a type of adversarial training system, where two competing models are trained against each other. The generator model attempts to transform random input noise into samples mirroring those of the training data distribution, and the discriminator model attempts to distinguish between real training data samples and generated samples. In our implementation, both the generator and discriminator are implemented as feed-forward neural networks. GANs have enjoyed much recent success in many generative tasks, including the generation of EHRs. However, to the best of our knowledge, only homogeneous data sets have been considered so far, for example in the MedGAN framework introduced by Choi et al. in [9].

In our model, we utilize a different GAN loss formulation based on the Earth Mover or Wasserstein distance, proposed by Arjovsky et al. [10], and improved by Gulrajani et al [11]. The so-called Wasserstein GANs (WGANs) do not suffer from mode collapse, and have the further benefit that the loss metric provides a meaningful indicator of the progress of the training [10]. WGANs have been used as an improvement to original GANs in several applications, and it is our hypothesis that the additional power provided by this model will allow for training a model capable of capturing and generating heterogeneous EHR data. See fig. 1 for an illustration of our model.

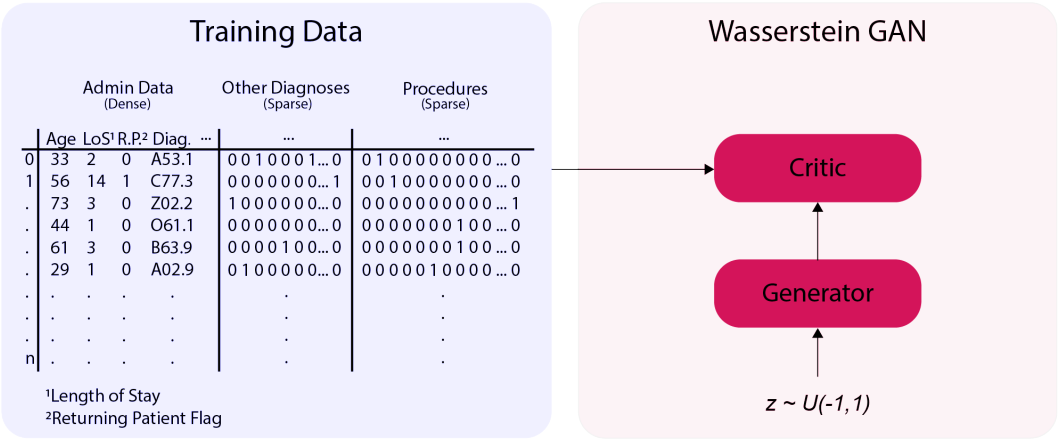

Figure 1: Our proposed model's architecture. The EHR training data is fed to the critic network of the WGAN model, along with the output of the generator network. The goal of the critic network is to learn a function maximizing the distance between the real and generated data distributions.

2.1 Differentially private variant

As discussed in the introduction, simply using synthetic data is not enough to guarantee privacy. However, rigorous statistical privacy guarantees can be given by using (ϵ, δ) differential privacy (DP) [7]. (ϵ, δ) DP is a framework which places theoretical limits on the likelihood of being able to determine the presence of a record in a data set. Without going into a theoretical introduction of DP, the general idea is that by adding strategic noise during the training process, the impact of any single training example is minimized. We additionally train our model using a DP-aware SGD optimizer provided by TensorFlow [12], with the privacy accounting performed as in [13], and the GAN training procedure modified as in [14]. The performance will be compared with the non-private model.

Table 1: Experiment Results. Data Fidelity metrics are the average divergence for each random variable type in the input space, as well as the Frobenius norm divergence. Data Utility metrics are the AUROC and AUPRC values, as well as the classification accuracy. All metrics are reported as the 95% confidence interval based on three full experiment repetitions. The differentially private configuration uses $\epsilon = 10$ and $\delta = 10^{-5}$.

	Data Fidelity		
Model	Bernoulli	Categorical	Frobenius
Baseline	-	-	-
WGAN Model	0.00057 ± 0.00009	0.000563 ± 0.00019	0.9811 ± 1.1272
WGAN Model w/ DP	0.0039 ± 0.0008	0.00294 ± 0.0008	14.24 ± 11.7778

	Data Utility		
Model	AUROC	AUPRC	Acc.
Baseline	0.8003	0.8245	0.7171
WGAN Model	0.7536 ± 0.0187	0.7747 ± 0.0155	0.6897 ± 0.0099
WGAN Model w/ DP	0.6427 ± 0.0088	0.6776 ± 0.0223	0.5844 ± 0.0215

3 Experiments

The model described in section 2, was trained using the New Zealand National Minimal data set [2]. This data set consists of de-identified EHRs from the New Zealand health care system, which includes dense administrative features such as patient age, gender, and length of stay, as well as sparse features such as diagnosis codes. These features are distributed according to several different distributions: diagnosis and procedure codes can be considered to be individual Bernoulli random variables, features such as admission month or department are categorical, age and length of stay are non-standard. This is therefore a heterogeneous data set. In this work we consider only the years 2012 to 2017. After balancing the data set and coarsening the diagnosis and procedure codes, we have a data set consisting of 2,873,466 rows and 795 columns. This data is split into the following two sections: training data (2,388,060 rows), and test data (485,406 rows).

We consider both the fidelity of the generated data to the training data set, as well as the utility of the generated data when evaluating the quality of the generated EHRs. In order to evaluate the generated data's fidelity, we compare the features between the two data sets depending on their distribution. Features given by Bernoulli random variables are compared according to the maximum likelihood estimate of the p parameter in each data set. The categorically distributed random variables are similarly compared using their estimated p_i parameters, where $i \in K$, the set of possible categories for each random variable. Finally, in order to compare the overall data sets (including non-standard features such as age or length of stay), we compare the Frobenius norm of samples from both sets. The closer the generated data matches the training data, the closer these parameter values will be.

To evaluate the utility of the generated data, we train binary classifiers using the training and generated data sets. The goal of these classifiers is to predict the risk of a patient's early readmission to a hospital, as described in the introduction. The efficacy of the classifiers is evaluated using the AUROC, AUPRC, and accuracy metrics. If the generative model has learned the training distribution well, the performance of both classifiers should be similar. Evaluation is performed on the following model configurations: the model proposed in section 2, and a differentially private variant. The results of these experiments are given in table 1.

4 Discussion

The results in table 1 show that our WGAN based model is capable of generating high-quality heterogeneous EHR data, both in terms of data fidelity and data utility. The Bernoulli random variables in the generated data match to within an average divergence of 0.00057, and the categorically

distributed random variables match to within an average divergence of 0.000563. As a further sanity check, for each categorical random variable, the probability of each category p_i is summed, and the average divergence from 1 is 0.000127. Finally, the Frobenius norm divergence between distribution samples is 0.9811, although this metric is somewhat noisy. This clearly demonstrates that our model learns to produce synthetic EHR records whose characteristics closely match those of the training data set.

Our model is also capable of producing EHRs with high utility. Compared to the training data baseline, when using the generated training data to classify real early readmitted patient data from the test set, the AUROC, AUPRC and accuracy metrics are only marginally worse, 0.0467 (5.8%), 0.0498 (6.4%), and 0.0274 (3.8%), respectively. This demonstrates that for this studied classification task, the synthetic data could viably be used in place of the real training data with little impact on performance.

The DP variant of our model does not produce data with as high fidelity and utility. In terms of data fidelity, the Bernoulli features have an average divergence of 0.0039 from the training data, and the categorical features match to an average divergence of 0.00294. For each categorical variable, the average divergence of the category probabilities from 1 is 0.02. The Frobenius norm divergence between the distribution samples is 14.24. These results, though less accurate than the non-DP variant, still match the training distribution fairly well. While the data fidelity results are reasonable, the drop in data utility performance is more pronounced. The AUROC is 0.1576 (19.7%) lower than the baseline, and 0.11 (14.7%) lower than the non-DP variant. Similarly, the AUPRC is 0.1469 (17.8%) lower than the baseline, and 0.0971 (12.5%) lower than the non-DP variant. Finally, the accuracy is 0.1327 (18.5%) lower than the baseline, and 0.1053 (15.3%) lower than the non-DP variant. See fig. 2 for AUROC and AUPRC curves illustrating the performance of the binary classifiers trained.

However, the obtained AUROC value of 0.6427 demonstrates that returning patient classification using differentially private synthetic heterogeneous EHR data is possible to some extent using our model. The (AUROC) performance penalty of 19.7% can therefore be seen as the cost of providing a reasonable $(10, 10^{-5})$ differential privacy guarantee to patients. Presumably, increasing the (ϵ, δ) privacy budget would lead to better results, as less noise would need to be added during the training process. If the lesser level of privacy provided by simply using synthetic EHR data suffices, our model allows for generation of very high-quality synthetic heterogeneous EHR data which can be used to train machine learning models, with only minimal performance loss.

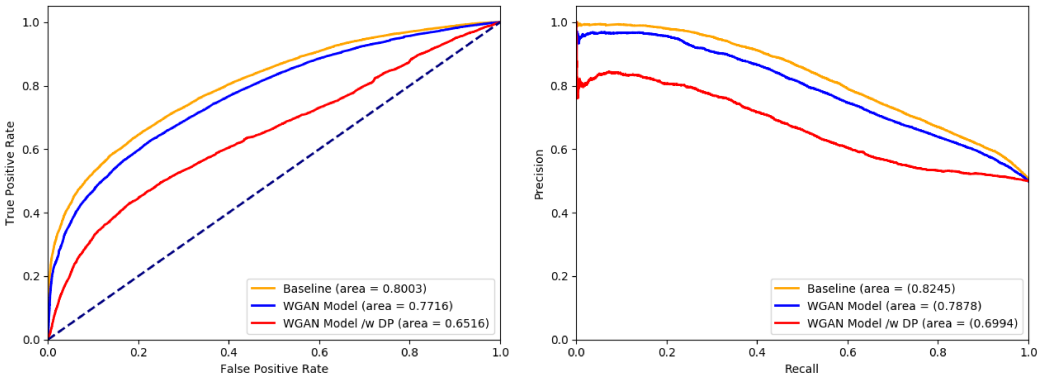

Figure 2: Left: AUROC curves for unexpected returning classifiers using training data (baseline), generated data, and generated data with DP. Right: AUPRC curves for unexpected returning classifiers using training data (baseline), generated data, and generated data with DP. These curves were generated using data from the first experiment repetition.

References

[1] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

[2] National minimum dataset (hospital events) | ministry of health nz. https://www.health.govt.nz/nz-health-statistics/national-collections-and-surveys/collections/national-minimum-dataset-hospital-events. Accessed: 2019-03-09.

[3] Readmissions reduction program (hrrp). https://www.cms.gov/medicare/medicare-fee-for-service-payment/acuteinpatientpps/readmissions-reduction-program.html. Accessed: 2019-02-09.

[4] Aha fact sheet: Hospital readmissions reduction program. https://www.aha.org/factsheet/2016-01-18-aha-fact-sheet-hospital-readmissions-reduction-program. Accessed: 2019-02-09.

[5] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22Nd ACM SIGSAC Conference on Computer and Communications Security*, CCS '15, pages 1322–1333, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3832-5. doi: 10.1145/2810103.2813677. URL http://doi.acm.org/10.1145/2810103.2813677.

[6] Khaled El Emam, Elizabeth Jonker, Luk Arbuckle, and Bradley Malin. A systematic review of re-identification attacks on health data. *PLOS ONE*, 6(12):1–12, 12 2011. doi: 10.1371/journal.pone.0028071. URL https://doi.org/10.1371/journal.pone.0028071.

[7] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, August 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL http://dx.doi.org/10.1561/0400000042.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *In NIPS*, pages 2672–2680, 2014.

[9] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F Stewart, and Jimeng Sun. Generating multi-label discrete patient records using generative adversarial networks. *Proceedings of Machine Learning for Healthcare*, 68, 2017.

[10] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. 2017. URL https://arxiv.org/abs/1701.07875.

[11] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein gans. *CoRR*, abs/1704.00028, 2017. URL http://arxiv.org/abs/1704.00028.

[12] tensorflow/privacy: Library for training machine learning models with privacy for training data. https://github.com/tensorflow/privacy. Accessed: 2019-10-09.

[13] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, pages 308–318, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4139-4. doi: 10.1145/2976749.2978318. URL http://doi.acm.org/10.1145/2976749.2978318.

[14] Xinyang Zhang, Shouling Ji, and Ting Wang. Differentially private releasing via deep generative model. *CoRR*, abs/1801.01594, 2018. URL http://arxiv.org/abs/1801.01594.