[FA-0033] [Basic]

Synthetic Clinical Text Generation : A Generative Adversarial Networks Approach

Dao Tan Tri Nguyen[1], **Sungyoung Lee**[1]
[1]*Kyung Hee University, Computer Science & Engineer Department, Kyung Hee University, Korea*

Object: De-identifying structured and unstructured patient data has been a major focus of research in recent years. Traditional de-identification methods focus on randomizing, suppressing, or generalizing potentially identifiable patient demographic factors such as names, addresses, IDs, and contact information. However, such efforts are not foolproof; patient records that have been scrubbed of PHI may still be vulnerable to re-identification based on residual clinical information included in symptoms, diagnosis, prescriptions, or test results. The difficulty of detecting potentially sensitive information from free-text data has a significant impact on structured data de-identification. Researchers have proposed many methods for producing synthetic data that replicate clinical patterns in medical records as a solution to the risk of re-identification based on clinical information (i.e., data augmentation for cardiovascular abnormality classification application). Previous works created data that was not realistic enough for machine learning.
Methods: GANs (Generative Adversarial Networks) is a deep learning technology with a high potential for enhancing synthetic data generation. GAN algorithms are implemented using a system of two neural networks. The generator neural network seeks to produce synthetic data, whereas the discriminator neural network seeks to distinguish between synthetic and real data. The generator network effectively creates bogus data that the discriminator cannot identify when these networks are trained. The earliest GAN models were developed to mimic real-world data. As a result, they've been employed to create high-quality categories and image databases. In healthcare, GAN models have been used to generate numerical clinical data that is statistically equivalent to real data. Thanks to recent developments, GAN algorithms can now generate synthetic free-text data. In this work, using heart failure clinical records, we trained GAN models and determined the best accurate models for constructing synthetic datasets that replicate the informative features of the original dataset.
Results: Our GAN-based clinical text generation approach produces synthetic unstructured free-text medical data that closely mimics real-world data characteristics
Conclusions: We investigate the potential of using GAN models to produce synthetic unstructured free-text medical data, which can then be used to construct machine learning models that mirror similar models generated with real data
Keyword: Clinical Text Generation, GANs, Cardiovascular
Acknowledgement: This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2017-0-01629) supervised by the IITP(Institute for Information & communications Technology Promotion)", and No.2017-0-00655(Lean UX core technology and platform for any digital artifacts UX evaluation) and IITP-2020-0-01489 (2020 Grand ICT Research Center)and 2022-0-00078,(Explainable Logical Reasoning for Medical Knowledge Generation) by Institute for Information & communications Technology Promotion(IITP).