딥러닝 모델을 이용한 유전자 발현 데이터 결측치 예측에 관한 조사

김재윤, 석준희*

고려대학교, *고려대학교

jyoonkim@korea.ac.kr, *jseok14@korea.ac.kr

A Survey of Missing Value Imputation for Gene Expression Data Using Deep Learning Models

Jaeyoon Kim, Junhee Seok*

Department of Electrical and Computer Engineering, Korea University

요 약

유전자 발현 데이터는 전사체학 연구에 중요한 필수적인 데이터이다. 하지만, 장치에 의한 의도치 않은 노이즈 발생 또는 불충분한 mRNA 분자 수로 인해 거짓 제로 값인 결측치의 비율이 높다. 따라서, 이러한 결측치 값을 실제 값에 가깝게 예측하려는 딥러닝 기반 연구가 많이 진행되고 있다. 본 논문에서는, 최신 딥러닝 기반 모델인 DeepImpute, AutoImpute, Deep Count Autoencoder Network (DCA), scGAIN, scIGANs에 대해 자세히 살펴보도록 한다.

Abstraction

Gene expression analysis is essential for transcriptomics studies. However, due to unintended noise generation by the device or an insufficient number of mRNA molecules, the rate of missing values, which are false zero, is high. Therefore, a number of deep learning-based studies have been conducted to impute missing values. In this paper, we will overview the state-of-the-art deep learning-based models: DeepImpute, AutoImpute, Deep Count Autoencoder Network (DCA), scGAIN, scIGANs.

Keywords : Deep learning, gene expression data, missing value imputation, DeepImpute, AutoImpute, DCA, scGAIN, scIGANs

Ⅰ. 서 론

최근 딥러닝과 머신러닝에 대한 연구가 다양한 분야에서 활발히 진행되고 있다. 더욱 정교하고 정확한 결과를 얻기 위해서는 많은 양의 데이터와 완전한 데이터가 필요하다. 하지만, 현실에서는 장치 손상, 데이터 수집 실패 및 기록 손실과 같은 여러 이유로 완전한 데이터를 얻는 것은 거의 불가능하다. 따라서, 결측치 예측을 통해 데이터를 채우는 여러 연구가 진행되고 있다 [1].

본 연구에서는 생물정보학 연구에 사용되고 있는 유전자 발현 데이터의 결측치 예측 딥러닝 모델들에 대해 살펴보고자 한다. 유전자 발현 데이터는 행과 열이 유전자와 세포 또는 샘플로 이루어져 있고, 숫자는 해당 세포 또는 샘플에서의 유전자의 발현된 정도를 의미한다. 즉, 숫자가 높을수록 유전자의 발현이 높다는 것이다. 하지만, 실제로 유전자 발현 데이터 파일을 열어보면 0의 비중이 매우 높다는 것을 확인할 수가 있다. 0이 존재하는 이유는 크게 두 가지다. 첫 번째는, 해당 유전자가 실제로 발현이 되지 않은 경우이다. 두 번째는, 실험 도중에 장치 오류 또는 발현의 정도가 크지 않아서 0으로 나타나진 경우이다. 두 번째를 의도하지 않은 데이터 손실, 즉 dropout이라고 한다. Single-cell RNA sequencing (scRNA-seq) 데이터의 경우에는 dropout이 70% 이상인 데이터도 많이 찾아볼 수 있다. 그림 1은 쥐 배아 줄기세포에 해당되는 GSE65525 데이터의 일부를 히트맵으로 나타낸 것이다. 0인 부분들이 빨갛게 표시되었다. 실제로 데이터의 70%가 0을 나타낸다. 이렇게 dropout이 많은 경우 데이터를 통해 정확한 연구 결과를 얻기가 쉽지 않다. 이를 해결하기 위해 여러 딥러닝 모델들이 제기되었는데, 본 논문에서는 최신 모델들인 DeepImpute, AutoImpute, Deep Count Autoencoder Network (DCA), scGAIN, scIGANs에 대해 자세히 살펴보도록 한다.

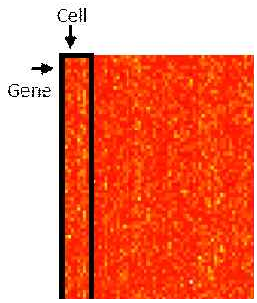

그림 1. Part of heatmap from GSE65525, scRNA-seq data. Areas with zeros are marked in red (70% of data)

Ⅱ. 본론

DeepImpute [2]는 심층 신경망 (Deep Neural Network, DNN) 기반 결측치 예측 알고리즘이다. 이 모델은 데이터의 패턴을 더 정확하게 파악하기 위해 드롭아웃 레이어와 손실 함수를 사용하였다. 상관관계가 높은 유전자의 관찰된 값을 고려하여 세포의 누락된 유전자 발현을 예측한다. 이

모델은 평균 제곱 오차와 피어슨 상관계수를 모델 성능 평가에 사용하였다. 모델의 학습과 성능 평가에는 scRNA-seq 데이터 셋인 Jurkat, 293T, neuron9k, Mouse1M, FISH, GSE99330, GSE67602와 GSE102827을 사용하였다.

AutoImpute [3]는 오토인코더 (Autoencoder) 기반 모델이다. 이 모델은 raw 유전자 발현 데이터 매트릭스에서 불필요해 보이는 유전자들을 걸러내는 유전자 선택 과정과 정규화를 거치면서 오직 중요한 유전자에만 초점을 맞춘다. 이 모델의 한계는 오직 상위 1,000개의 유전자에 대해서만 결측치 예측이 이루어지고, 나머지 대부분의 유전자들은 예측이 되지 않은 채 그대로 남겨진다는 것이다. 평균 제곱근 편차, 정규화 된 평균 제곱근 편차와 평균 절대 오차를 사용하였다. Blakeley, Jurkat-293T, Kolodziejczyk, PBMC, Preimplantation, Quake, Tapnell, Usoskin, Zeisel 데이터 셋을 사용하였다.

DCA [4] 모델 또한 AutoImpute와 마찬가지로 오토인코더 기반 모델이다. 이 모델은 overdispersion 하고 sparse 한 count 유전자 발현 데이터를 zero-inflated negative binomial 분포를 사용한 오토인코더로 학습시킨다. 이 모델은 세포 수에 따라 선형적으로 확장되므로 수백만 개의 세포 데이터 세트에 적용될 수 있다. DCA는 드롭아웃의 분포가 미리 결정된 노이즈 분포를 따른다고 가정한다. 따라서, 시뮬레이션 데이터에는 잘 작동하지만 실제 데이터에서도 잘 작동하는지는 명확하지 않다는 한계가 있다. C.elegans embryos, GSE75748, GSE100866, 68k PBMC 데이터 셋을 사용하였다.

scGAIIN [5]은 결측치의 비율이 매우 높은 scRNA-seq 데이터를 generative adversarial networks (GAN)을 사용하여 그 값을 예측하는 모델이다. scGAIN 모델은 이미지 데이터의 손실 부분을 채우기 위해 연구된 모델인 GAIN을 scRNA-seq에 적용 가능하도록 구현되었다. scGAIN은 일반적인 GAN 과는 다른 구조이다. 우선, generator 부분에는 특정 셀의 유전자 값이 dropout 부분인지를 분별할 수 있도록 mask를 사용한다. 따라서, dropout 되어있는 부분에 새로 채워진 값과 dropout 부분이 아닌 원래의 값들을 합쳐서 새로운 유전자 발현 매트릭스를 형성한다. Discriminator 부분에는 mask에 대한 정보를 제공하는 힌트 메커니즘을 사용한다. 이 모델은 Splatter와 SimData60k를 이용한 시뮬레이션 데이터 셋과 실제 데이터 셋인 PBMC를 사용하였다.

scIGANs [6]도 scGAIN과 마찬가지로 GAN 기반 모델이다. 다른 모델들과 가장 다른 점은, 유전자 발현 데이터를 이미지 데이터처럼 만들어서 Convolutional Neural Networks (CNN)을 사용했다는 점이다. scIGANs의 유전자 발현 데이터는 행이 유전자, 열이 세포를 의미한다. 하나의 세포가 정사각형 모양의 하나의 이미지처럼 만들어진 후, GAN 모델의 인풋으로 들어간다. 또한, 세포 타입을 나타내주는 label을 generator 부분에 넣어준다. Discriminator는 encoder와 decoder로 구성되어 있다. GAN을 통해 생성된 데이터에서 KNNImputer를 사용하여 0인 부분을 채워준다. 이렇게 생성된 가짜 데이터와 원래의 raw data를 비교하여 raw data의 0인 부분만 새로운 값으로 채우고 0이 아니었던 부분은 그대로 값을 유지한다. 이런 방식은 오버 피팅을 막기 위해서이다. scIGANs 모델은 CIDR, Splatter 시뮬레이션 데이터와 GSE67835, E-MTAB-2804, E-MTAB-2805, GSE65525, GSE75748 실제 데이터 셋이 사용되었다.

아래 표 1은 위에 설명하였던 모델들에 대한 비교를 간단하게 표로 정리한 것이다. DeepImpute는 순방향 신경망 (feed-forward neural network, FFNN), AutoImpute와 DCA는 오토인코더, scGAIN과 scIGANs는 GAN을 기반으로 한 딥러닝 모델이다.

Method	Structure	reference
DeepImpute	FFNN	[2]
AutoImpute	Autoencoder	[3]
DCA	Autoencoder	[4]
scGAIN	GAN	[5]
scIGANs	GAN	[6]

표 1. Gene expression missing value imputation models comparison

Ⅲ. 결론

딥러닝 모델들을 사용하여 결측치를 예측하는 연구는 여러 다양한 데이터에 적용하여 진행되고 있다. 그중, 이미지 데이터를 이용한 연구가 가장 활발하게 이루어지고 있고 numeric 데이터에 대한 연구도 활발히 이루어지고 있다. 유전자 발현 데이터도 숫자로 이루어진 매트릭스이긴 하나, 다른 데이터에 비해 dropout 비율이 매우 높다. 특히 scRNA-seq 데이터들은 dropout 비율이 70% 이상인 경우도 많아서 그 값을 예측하기가 더 어렵다.

본 논문에서는 DeepImpute, AutoImpute, DCA, scGAIN 그리고 scIGANs 모델들을 소개하였다. 모든 모델들에서 결측치를 채운 유전자 발현 데이터가 그렇지 않은 데이터에 비해 클러스터링 또는 분류 문제에서 더 높은 정확도를 보임을 증명하였다.

이러한 여러 모델들의 장점을 잘 활용한다면, 보다 정확하고 빠르게 유전자 발현 데이터의 결측치 예측을 할 수 있는 방법론을 구축할 수 있을 것이라 생각한다.

ACKNOWLEDGMENT

이 논문은 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2019R1A2C1084778).

참 고 문 헌

[1] J Kim, J Seok: A Survey of Missing Data Imputation Using Generative Adversarial Networks, ICAIIC (2020)

[2] Arisdakessian, C., Poirion, O., Yunits, B., Zhu, X., Garmire, L.X.: Deepimpute: an accurate, fast, and scalable deep neural network method to impute single-cell rna-seq data. Genome Biology 20(1), 1 - 14 (2019)

[3] Talwar, D., Mongia, A., Sengupta, D., Majumdar, A.: Autoimpute: Autoencoder based imputation of singlecell rna-seq data. Scientific reports 8(1), 16329 (2018)

[4] Eraslan, G., Simon, L.M., Mircea, M., Mueller, N.S., Theis, F.J.: Single-cell rna-seq denoising using a deep count autoencoder. Nature communications 10(1), 390 (2019)

[5] MK Gunady, J Kancherla, HC Bravo, S Feizi: scGAIN: Single Cell RNA-seq Data Imputation using Generative Adversarial Networks. BioRxiv (2019) https://doi.org/10.1101/837302

[6] Xu, Y. et al. scIGANs: single-cell RNA-seq imputation using generative adversarial networks. Nucleic Acids Res. 48, e85 (2020).