Using deep autoencoders for denoising and reduction of Single-Cell RNA-Seq data

Sajal Kumar
Department of Computer Science,
New Mexico State University
Las Cruces, New Mexico
sajal49@nmsu.edu

Jiandong Wang
Department of Computer Science,
New Mexico State University
Las Cruces, New Mexico
wangjd24@nmsu.edu

Xiaonan Zhu
Department of Mathematical Sciences,
New Mexico State University
Las Cruces, New Mexico
xzhu@nmsu.edu

ABSTRACT

Single-cell RNA sequencing (scRNA-seq) technology is un-parallel in providing high resolution gene expression data at cellular levels. However, it is also notoriously known for being noisy due to amplification and dropout, making it a challenge for analytical methods to differentiate noise from patterns, so a scalable imputation and denoising mechanism is required. Here we apply DCA – a deep neural network based denoising auto-encoders for scRNA-seq on a brain cell dataset from 10x genomics, which profiles 1.3 M cells from embryonic mice brains. We use this data set to see if gene expression can be used to classify various brain-cell types and illustrate the performance of the classical decision tree method before and after encoding.

CCS CONCEPTS

• **Computational biology → Dimensionality reduction**; **Denoising**; *Single cell RNA-Seq*; *Deep autoencoders*;

KEYWORDS

Dimensionality reduction, Denoising, Single cell RNA-Seq, Deep autoencoders

ACM Reference Format:
Sajal Kumar, Jiandong Wang, and Xiaonan Zhu. 2018. Using deep autoencoders for denoising and reduction of Single-Cell RNA-Seq data. In *Proceedings of ACM conference (XXXX'18),* xxxxx and xxxxx (Eds.). ACM, New York, NY, USA, Article xx, 1 page. https://doi.org/xx.xxx/xxxx

INTRODUCTION

Motivation: Single-cell RNA-seq is a promising technique that profiles transcriptomes over a plethora of cell types. By measuring individual cells, rather than averaged tissue types, scRNA-seq can uncover biological mechanisms that is not observed by the average behaviors of a bulk of cells [4]. However, despite improvements in measuring technologies, factors including amplification bias, cell cycle effects [1], etc. lead to substantial noise in scRNA-seq experiments. The low RNA capture rate leads to failure of detection of an expressed gene resulting in a "false" zero count observation, defined as dropout event; which potentially corrupts the underlying biological signal [2]. One such scRNA-seq data set is from the 10x genomics project that profiles 1.3 M embroynic brain cells of mice. It represents cells from cortex, hippocampus and subventricular zone of two mice at 18 days post conception which is an invaluable resource to study gene dynamics during an important development stage, however, the corruption induced by dropout events compromises our ability to use analytical techniques to extract meaningful patterns.

Current approaches for scRNA-seq denoising / imputation rely on using the correlation structure of single-cell gene expression data, leveraging similarities between cells and/or genes that does not scale well; or has explicit / implicit linearity assumption that may miss complex non-linear patterns [2]. Deep-learning auto-encoders can learn the underlying complex manifold, that represents the biological processes and/or cellular states and utilize that to reduce the high dimensional data space to significantly lower dimensions [3]. However, using an auto-encoder as is may fail due to the noise model not being adapted to typical scRNA-seq noise. One such autoencoder DCA – deep count autoencoder facilitates non-linear embedding of cells and uses scRNA-seq data specific loss function based on negative binomial count distributions which offers adaptive learning and encoding of the data manifold [2]. Thus, we utilize DCA to denoise and reduce the dataset and then apply the classical decision tree algorithms to learn if gene expression dynamics can infer different cell types.

Problem definition: Given expression of X genes across m cell types, we denoise and reduce the data set to X' gene expression of m cell types, expecting $n(X') \leq n(X)$ by using the DCA autoencoder. We then apply the decision tree classifier on X' (genes expressions) to learn the m classes (cell types). We also apply the decision tree algorithm on the raw data to see if the encoded X' actually results in improved accuracy.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
XXXX'18, xxx 2018, xxxx, xx USA
© 2018 Copyright held by the owner/author(s).
ACM ISBN xxxx-xxxx-xxxx-xxxx.
https://doi.org/xx.xxx/xxxx

REFERENCES

[1] Florian Buettner, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J Theis, Sarah A Teichmann, John C Marioni, and Oliver Stegle. 2015. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature biotechnology* 33, 2 (2015), 155.

[2] Gökcen Eraslan, Lukas M Simon, Maria Mircea, Nikola S Mueller, and Fabian J Theis. 2018. Single cell RNA-seq denoising using a deep count autoencoder. *bioRxiv* (2018), 300681.

[3] Kevin R Moon, Jay Stanley, Daniel Burkhardt, David van Dijk, Guy Wolf, and Smita Krishnaswamy. 2017. Manifold learning-based methods for analyzing single-cell RNA-sequencing data. *Current Opinion in Systems Biology* (2017).

[4] Dongfang Wang and Jin Gu. 2017. VASC: dimension reduction and visualization of single cell RNA sequencing data by deep variational autoencoder. *bioRxiv* (2017), 199315.