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, 3 pages. 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 [6]. 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 [4]. 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.

SOLUTION

Data source

We are using Single Cell Gene Expression Dataset by Cell Ranger 1.3.0 from the 10x Genomics project that provides 1.3 Million Brain Cells from E18 Mice which consists of cells from cortex, hippocampus and subventricular zone of two E18 mice. The data was retrieved from support.10xgenomics.com.

Proposed Pipeline

We will proceed with two phases of analysis as shown in figure 1. The first phase would include using the single cell data 'as-is' (X, Y) and learn if there exists a non-linear combination of gene expression

XXXX'18, xxx 2018, xxxx, xx USA

ACM ISBN xxxx-xxxx-xxxx-xxxx.
https://doi.org/xx.xxx/xxxx

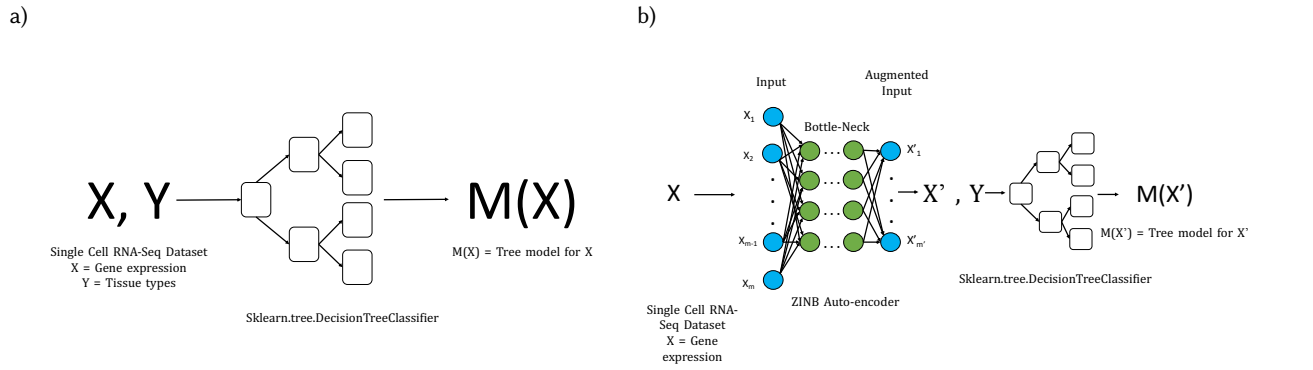

Figure 1: a) In the first analysis without any denoising, the single cell data will be supplied as is to a sklearn decision tree algorithm. b) In the second analysis, the single cell data will be denoised using the ZINB DCA autoencoder and then the learned manifold $X^{'}$ will be supplied to the sklearn decision tree algorithm.

$M(X)$ that can differentiate between the three class labels (Y) : *cortex*, *hippo-campus* and *sub-ventricular zone*. We will be utilizing the decision tree classifier – *sklearn.tree.DecisionTreeClassifier* provided by sci-kit learn package in Python. We will use sklearn.model_selection to perform the 10-fold cross-validation strategy and evaluate the performance of the decision tree model.

Next, we will use the zero-inflated negative binomial (ZINB) DCA auto-encoder [2] to first denoise the single cell data (X) and see if we can learn the manifold of the data, as claimed by the authors of this method. The augmented data X' will then be given to a similar setup of decision-tree learning with 10-fold cross validation to learn the three cell types. Our final objective would be to report any improvement in performance, followed by deeper insight on to why/why didn't the auto-encoding worked. Our expectation is that the augmented data X' should work better as it uses zero-inflated negative binomial distribution that can model highly sparse and over-dispersed data [2] which makes it very suitable for single-cell RNA-seq.

Preliminary results

Before applying the ZINB DCA auto-encoder, we established the performance of normalized single-cell gene expression profiles to predict cell types. For this purpose we used head and neck cancer single-cell data-set (GSE103322) [5] that profiled 23686 genes across 5902 single cells obtained from head and neck region of 18 patients. This data-set contains both normal and cancer samples, where the normal samples come from a patient's non-metastasized areas. We only used normal samples, reducing our sample size to 3363 containing : B-cell(138), Dendritic(51), Endothelial(260), Fibroblast(1440), Macrophage(98), Mast(120), Myocyte(19) and T cell(1237).

1	2	3	4	5	6	7	8	9	10
0.97	0.96	0.98	0.98	0.97	0.98	0.97	0.97	0.89	0.87

Table 1: Accuracy over 10-fold cross validation.

The task at hand was to model the 8 cell-types as a non-linear combination of gene-expression using decision tree learning. We utilized *sklearn.tree.DecisionTreeClassifier* for the task using default values of all parameters except `min_samples_split` which was set to 20 and `min_samples_leaf` which was set to 5. The single-cell data was transformed using the *sklearn.preprocessing.StandardScaler* transformation. We utilized the 10-fold cross-validation strategy, using stratified partition of cell-types, to measure the accuracy of our decision tree model. Table 1 shows accuracy for all 10 runs using 9/10th data as training and the rest 1/10th as testing, using different subset every-time.

Interestingly, this single-cell data can already do very well. Also, the number of decision tree nodes used to accomplish such accuracy was a modest 61 nodes, which means only a handful genes out of 23K genes can reliably predict the 8 cell-types. We rationaled that the high accuracy could be because of the imbalance in class distribution, however, Fibroblast and T-cells are quite balanced and the decision-tree was able to reliably distinguish between the two cell types.

While genes being able to distinguish cell types is not surprising, recently, Karaiskos et al. [3] used ISH (In-situ hybridization) stains of 84 genes to perfectly classify over 3000 cell locale in Drosophila embryo. However, single-cell, due to its noisy nature has not been thought to have the ability to be good at classifying cell-types. It would be interesting to see how other single-cell data-sets do in this scenario and how much can the performance improve after denoising. Unfortunately, the head-and-neck cancer single cell data-set does not offer raw data for denoising, but our final steps include considering methods to un-normalize the single-cell data and try other single-cell data with availability of the raw samples. Another noticeable peculiarity was a major drop in accuracy for the last two runs. We are yet to investigate on that issue.

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] Nikos Karaiskos, Philipp Wahle, Jonathan Alles, Anastasiya Boltengagen, Salah Ayoub, Claudia Kipar, Christine Kocks, Nikolaus Rajewsky, and Robert P Zinzen. 2017. The Drosophila embryo at single-cell transcriptome resolution. *Science* 358, 6360 (2017), 194–199.

[4] 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).

[5] Sidharth V Puram, Itay Tirosh, Anuraag S Parikh, Anoop P Patel, Keren Yizhak, Shawn Gillespie, Christopher Rodman, Christina L Luo, Edmund A Mroz, Kevin S Emerick, et al. 2017. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell* 171, 7 (2017), 1611–1624.

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