Cluster Aware Deep Dictionary Learning for Single Cell Analysis

Priyadarshini Rai[1], Angshul Majumdar[1(✉)], and Debarka Sengupta[1,2]

[1] Indraprastha Institute of Information Technology, New Delhi 110020, India
{priyadarshinir,angshul,debarka}@iiitd.ac.in
[2] Institute of Health and Biomedical Innovation,
Queensland University of Technology, Brisbane, Australia
https://iiitd.ac.in/

Abstract. The importance of clustering the single-cell RNA sequence is well known. Traditional clustering techniques (GiniClust, Seurat, etc.) have mostly been used to address this problem. This is the first work that develops a deep dictionary learning-based solution for the same. Our work builds on the framework of deep dictionary learning. We make the framework clustering friendly by incorporating a cluster-aware loss (K-means and sparse subspace) into the learning problem. Comparison with tailored clustering techniques for single-cell RNA and with generic deep learning-based clustering techniques shows the promise of our approach.

Keywords: Single cell clustering · Deep dictionary learning · Single cell analysis

1 Introduction

The problem of clustering is well known; there are many reviews (such as [1]) on this topic. The general topic of clustering studies the formation of naturally occurring groups within the data. The simplest (and still the most popular) approach for the same is perhaps K-means [2]. K-means segments the data by relative distances; samples near each other (pre-defined by some distance metric) are assumed to belong to the same cluster. Owing to the linear nature of the distance, the K-means was not able to capture non-linearly occurring groups. This issue was partially addressed by the introduction of kernel K-means [3]. Instead of defining the distance between the samples, a kernel distance was defined (Gaussian, Laplacian, polynomial, etc.) for clustering. Closely related to the kernel K-means is spectral clustering [3]. The later generalizes kernel distances to any affinity measure and applies graph cuts to segment the clusters.

K-means, kernel K-means, and spectral clustering are inter-related. A completely different approach is subspace clustering [4]. In the later, it is assumed that samples belonging to the same group/cluster will lie in the same subspace. There are several variants of subspace clustering, but the most popular one

H. Yang et al. (Eds.): ICONIP 2020, CCIS 1333, pp. 62–69, 2020.
https://doi.org/10.1007/978-3-030-63823-8_8

among them is the sparse subspace clustering (SSC) [5]. In SSC it is assumed that the clusters only occupy a few subspaces (from all possibilities) and hence the epithet “sparse”.

So far, we have discussed generic clustering techniques. In the single-cell analysis, cell type identification is important for the downstream analysis. Therefore, clustering forms a crucial step in single-cell RNA expression analysis. Single-cell RNA sequencing (scRNA-seq) measures the transcription level of genes. But, the amount of RNA present in a single cell is very low due to which some genes did not get detect even though they are present and this results in zero-inated data. This data further gets compounded by trivial biological noise such as variability in the cell cycle specic genes. Also, a large number of genes are assayed during an experiment but only a handful of them are used for cell-type identification. This leads to high feature-dimensionality and high feature-redundancy in single-cell data. Applying clustering techniques directly on the high-dimensional data will cause suboptimal partitioning of cells.

This triggers the need for customized techniques. The existing state-of-the-art clustering techniques for single-cell data do not propose new algorithms for clustering per se but apply existing algorithms on extracted/reduced feature sets. One popular technique Seurat [6], instead of applying a distance-based clustering technique on all the genes, selects highly variable genes from which a shared nearest neighbor graph is constructed for segmentation. GiniClust [7] is similar to the former and only differs in the use of the Gini coefficient for measuring differentiating genes. Single-cell consensus clustering (SC3) [8] algorithm uses principal component analysis (PCA) to reduce the dimensions and then applies a cluster-based similarity partitioning algorithm for segmentation.

The success of deep learning is well known in every field today. What is interesting to note is that success has been largely driven by supervised tasks; there are only a handful of fundamental papers on deep dictionary learning-based clustering [9]. Deep dictionary learning is a new framework for deep learning. In the past, it has been used for unsupervised feature extraction [10], supervised classification [11], and even for domain adaptation [12]. However, it has never been used for clustering. This would be the first work on that topic. The advantage of deep dictionary learning is that it is mathematically flexible and can easily accommodate different cost functions. In this work, we propose to incorporate K-means clustering and sparse subspace clustering as losses to the unsupervised framework of deep dictionary learning.

2 Proposed Formulation

There are three pillars of deep learning - convolutional neural network (CNN), stacked autoencoder (SAE), and deep belief network (DBN). The discussion on CNN is not relevant here since it can only handle naturally occurring signals with local correlations. Moreover, they cannot operate in an unsupervised fashion, and hence is not a candidate for our topic of interest. Stacked autoencoders have been used for our purpose (deep learning-based clustering); the main issue with SAE is that it tends to overfit since one needs to learn twice the number of parameters (encoder and decoder) compared to other standard neural networks. However, SAE's are operationally easy to handle with good mathematical flexibility. DBN on the other hand learns the optimal number of parameters and hence does not overfit. However, the cost function DBN is not amenable to mathematical manipulations.

Deep dictionary learning keeps the best of both worlds. It learns the optimal number of parameters like a DBN and has a mathematically flexible cost function making it amenable to handle different types of penalties. This is the primary reason for building our clustering on top of the deep dictionary learning (DDL) framework. In our proposed formulation, we will regularize the DDL cost function with clustering penalties, where X is the given data (X – in our case single cells are along the columns and genes are along the rows), D is the dictionary learned to synthesize the data from the learned coefficients Z.

$$\min_{D_1,\ldots D_N,Z} \|X - D_1\varphi\left(D_2\varphi(\ldots\varphi(D_N Z))\right)\|_F^2 \tag{1}$$

The first clustering penalty will be with K-means.

$$\min_{D_1,D_2,D_3,Z,H} \underbrace{\|X - D_1D_2D_3Z\|_F^2 \,\text{s.t.} D_2D_3Z \geq 0, D_3Z \geq 0, Z \geq 0}_{\text{DictionaryLearning}} \\ + \underbrace{\left\|Z - ZH^T\left(HH^T\right)^{-1}H\right\|_F^2 \text{s.t.} h_{ij} \in \{0,1\} \text{ and } \sum_j h_{ij} = 1}_{\text{K}-\text{means}} \tag{2}$$

Note that we have changed the cost function for dictionary learning. Instead of having activation functions like sigmoid or tanh, we are using the ReLU type cost function by incorporating positivity constraints. The reason for using ReLU over others is better function approximation capability [13]. The notations in the K-means clustering penalty has been changed appropriately.

In this work, we will follow the greedy approach for solving (2). In the dictionary learning part, we substitute $Z_1 = D_2D_3Z$. This leads to the greedy solution of the first layer of deep dictionary learning.

$$\min_{D_1,Z_1} |X - D_1Z_1|\,|_F^2 \; s.t. Z_1 \geq 0 \tag{3}$$

The input for the second layer of dictionary learning uses the output from the first layer (Z1). The substitution is $Z_2 = D_3Z$. This leads to the following problem

$$\min_{D_2,Z_2} \|Z_1 - D_2Z_2\|_F^2 \,\text{s.t.} Z_2 \geq 0 \tag{4}$$

For the third (and final) layer no substitution is necessary; only the output from the second layer is fed into it.

$$\min_{D_3,Z} \|Z_2 - D_3Z\|_F^2 \,\text{s.t.} Z \geq 0 \tag{5}$$

All the problems (3)–(5) can be solved by non-negative matrix factorization techniques; in particular, we have used the multiplicative updates [14]. Although shown here for three layers, it can be extended to any number.

The input to K-means clustering is the coefficients from the final layer (Z). This is shown as

$$\min_{H} \left\| Z - ZH^T \left(HH^T \right)^{-1} H \right\|_F^2 \text{s.t.} h_{ij} \in \{0, 1\} \text{ and } \sum_j h_{ij} = 1 \tag{6}$$

The standard K-means clustering algorithm is used to solve it.

This concludes our algorithm to solve for the K-means embedded deep dictionary learning algorithm. Owing to the greedy nature of the solution, we cannot claim this to be optimal (owing to lack of feedback from deeper to shallower layers); however, each of the problems we need to solve (3)–(6) have well-known solutions.

Next, we show how the sparse subspace clustering algorithm can be embedded in the deep dictionary learning framework.

$$\begin{aligned} \min_{D_1, D_2, D_3, Z, C} & \underbrace{\|X - D_1 D_2 D_3 Z\|_F^2 \text{s.t.} D_2 D_3 Z \geq 0, D_3 Z \geq 0, Z \geq 0}_{\text{DictionaryLearning}} \\ & + \underbrace{\sum_i \|z_i - Z_{i^c} c_i\|_2^2 + \|c_i\|_1, \forall i \text{ in } \{1, ..., n\}}_{\text{SparseSubspaceClustering}} \end{aligned} \tag{7}$$

The solution to the deep dictionary learning remains the same as before; it can be solved greedily using (3)–(5). Once the coefficients from the deepest layer are obtained (Z), it is fed into the sparse subspace clustering. This is given by

$$\min_{c_i{}'s} \sum_i \|z_i - Z_{i^c} c_i\|_2^2 + \|c_i\|_1, \forall i \text{ in } \{1, ..., n\} \tag{8}$$

Once (8) is solved, the affinity matrix is created and is further used for segmenting the data using Normalized Cuts.

3 Experimental Evaluation

3.1 Datasets

To evaluate the performance of the proposed method we used seven single-cell datasets from different studies.

Blakeley: The dataset consists of three cell lineages of the human blastocyst which are obtained using single-cell RNA sequencing (scRNA-seq). This scRNA-seq data of the human embryo gives an insight into early human development and was validated using protein levels. The study consists of 30 transcriptomes from three cell lines, namely, human pluripotent epiblast (EPI) cells, extraembryonic trophectoderm cells, and primitive endoderm cells [15].

Cell Line: Microfluidic technology-based protocol, Fluidigm, was used to perform scRNA-seq of 630 single-cells acquired from 7 cell lines. Each cell line was sequenced separately. Therefore, the original annotations were directly used. The sequencing results in 9 different cell lines, namely, A549, GM12878 B1, GM12878 B2, H1 B1, H1 B2, H1437, HCT116, IMR90, and K562. The cell lines GM12878 and H1 had two different batches [16].

Jurkat-293T: This dataset consists of 3,300 transcriptomes from two different cell lines - Jurkat and 293 T cells. The transcriptomes are combined in vitro at equal proportions (50:50). All transcriptomes are labeled according to the mutations and expressions of cell-type-specific markers, CD3D, and XIST [17].

Kolodziejczyk: This study reports the scRNA-seq of $\sim$704 mouse embryonic stem cells (mESCs) which are cultured in three different conditions, namely, serum, 2i, and alternative ground state a2i. The different culture condition of the cells results in different cellular mRNA expression [18].

PBMC: This dataset constitutes $\sim$68,000 peripheral blood mononuclear cell (PBMC) transcriptomes from healthy donors. They are annotated into 11 common PBMC subtypes depending on correlation with fluorescence activated cell sorting (FACS)-based purified bulk RNA-Seq data of common PBMC subtypes. For this study, we randomly sampled 100 cells from each annotated subtype and retained the complete cluster in case the number of cells in it was less than 100 [17].

Usoskin: The data consists of 799 transcriptomes from mouse lumbar dorsal root ganglion (DRG). The authors used an unsupervised approach to cluster the cells. Out of 799 cells, 622 cells were classified as neurons, 68 cells had an ambiguous assignment and 109 cells were non-neuronal. The 622 mouse neuron cells were further classified into four major groups, namely, neurofilament containing (NF), non-peptidergic nociceptors (NP), peptidergic nociceptors (PEP), and tyrosine hydroxylase containing (TH), based on well-known markers [19].

Zygote: The RNA-sequencing data consists of 265 single cells of mouse preimplantation embryos. It contains expression profiles of cells from zygote, early 2-cell stage, middle 2-cell stage, late 2-cell stage, 4-cell stage, 8-cell stage, 16-cell stage, early blastocyst, middle blastocyst, and late blastocyst stages [20].

3.2 Numerical Results

In the first set of experiments, we have compared the proposed algorithm with the two state-of-the-art deep learning techniques. The first technique is a stacked autoencoder (SAE) which comprises two hidden layers. The number of neurons in the first hidden layer of SAE is 20 and the nodes in the second layer are the same as the number of cell types in the single-cell data. The second method used as a benchmark is a deep belief network (DBN). Like SAE, DBN also has two hidden layers with 100 nodes in the first layer and the number of nodes in the second layer is the same as the number of clusters in the given dataset. For our proposed deep dictionary learning (DDL) the number of nodes in the first layer was 20 and those in the second one are the same as the number of cell types (similar to the configuration of SAE). These configurations yielded the best results. Both state-of-the-art techniques along with the proposed method use the K-means algorithm on the deepest layer of features to determine the clusters in the data.

To determine how SAE, DBN, and the proposed method can segregate different cell types using the respective deepest layer of features we employed two clustering metrics: adjusted rand index (ARI) and normalized mutual information (NMI), since the ground truth annotation (class) of each sample or cell is known apriori (Table 1).

Table 1. Clustering accuracy of the proposed method and existing deep learning techniques on single-cell datasets.

Algo	Metric	Blakeley	Cell line	Jurkat	Kolodziejczyk	PBMC	Usoskin	Zygote
DBN	**NMI**	.190	.567	.001	.032	.273	.015	.385
	ARI	.056	.430	.001	.171	.103	.007	.296
SAE	**NMI**	.181	.099	.925	.170	**.573**	.040	.107
	ARI	.011	.007	.958	.215	**.377**	.001	.006
Proposed method	**NMI**	**.933**	**.873**	**.974**	**.694**	**.546**	**.647**	**.639**
	ARI	**.891**	**.801**	**.989**	**.645**	**.359**	**.642**	**.359**

We see that the proposed method improves over existing deep learning tools by a large margin. Only in the case of PBMC are the results from SAE a close second.

In the next set of experiments, we used two well-known single-cell clustering methods, namely, GiniClust [7] and Seurat [6] as benchmark techniques. For both of our proposed methods (K-means and SSC) the configuration remains the same as before.

GiniClust could not yield any clustering results for the Cell Line dataset. It performs clustering by utilizing genes with a high Gini coefficient value. But, for this particular dataset, the technique could not identify any highly variable gene and hence could not cluster. Overall GiniClust almost always yields the worst results.

Among the proposed techniques (K-means and SSC), we find that K-means is more stable and consistently yields good results. Results from SSC fluctuate, yielding perfect clustering for Blakely to poor results in Kolodziejczyk, PBMC, and Usoskin. Only for the Kolodziejczyk and PBMC datasets does Seurat yield results comparable to Proposed + K-means; for the rest, Seurat is considerably worse than either of our techniques (Table 2).

Table 2. Clustering accuracy of the proposed method and single-cell clustering algorithms on single-cell datasets

Algo	Metric	Blakeley	Cell line	Jurkat	Kolodziejczyk	PBMC	Usoskin	Zygote
GiniClust	**NMI**	.277	–	.007	.214	.153	.061	.282
	ARI	.037	–	.000	.055	.030	.006	.025
Seurat	**NMI**	0	.717	.946	**.695**	**.585**	.447	.453
	ARI	0	.533	.974	**.710**	**.296**	.382	.123
Proposed + Kmeans	**NMI**	.933	.873	**.974**	**.694**	**.545**	**.647**	**.639**
	ARI	.891	.801	**.989**	**.645**	**.359**	**.642**	**.359**
Proposed + SSC	**NMI**	**1**	**.879**	.889	.522	.481	.492	.623
	ARI	**1**	**.814**	.821	.510	.303	.453	.317

4 Conclusion

This work proposes a deep dictionary learning-based clustering framework. Given the input (where samples/cells are in rows and features/genes are in columns) it generates a low-dimensional embedding of the data which feeds into a clustering algorithm. The low dimensional embedding represents each transcriptome; it is learned in such a manner that the final output is naturally clustered.

To evaluate the proposed method, we have compared against state-of-the-art deep learning techniques (SAE and DBN) and tailored single-cell RNA clustering techniques (GiniClust and Seurat). Our method yields the best overall results.

The current approach is greedy and hence sub-optimal; there is no feedback between the deeper and shallower layers. In the future, we would like to jointly solve the complete formulations (2) and (7) using state-of-the-art optimization tools.

References

1. Saxena, A., et al.: A review of clustering techniques and developments. Neurocomputing **267**, 664–681 (2017)
2. Jain, A.K.: Data clustering: 50 years beyond K-means. Pattern Recogn. Lett. **31**(8), 651–666 (2010)
3. Dhillon, I.S., Guan, Y., Kulis, B.: Kernel k-means: spectral clustering and normalized cuts. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 551–556, August 2004
4. Vidal, R.: Subspace clustering. IEEE Signal Process. Mag. **28**(2), 52–68 (2011)
5. Elhamifar, E., Vidal, R.: Sparse subspace clustering. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2790–2797, June 2009
6. Waltman, L., van Eck, N.J.: A smart local moving algorithm for large-scale modularity-based community detection. Eur. Phys. J. B **86**(11), 1–14 (2013). https://doi.org/10.1140/epjb/e2013-40829-0
7. Jiang, L., Chen, H., Pinello, L., Yuan, G.C.: GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. Genome Biol. **17**(1), 144 (2016)
8. Kiselev, V.Y., et al.: SC3: consensus clustering of single-cell RNA-seq data. Nat. Methods **14**(5), 483–486 (2017)

9. Peng, X., Xiao, S., Feng, J., Yau, W.Y., Yi, Z.: Deep subspace clustering with sparsity prior. In: IJCAI, pp. 1925–1931, July 2016
10. Tariyal, S., Majumdar, A., Singh, R., Vatsa, M.: Deep dictionary learning. IEEE Access **4**, 10096–10109 (2016)
11. Mahdizadehaghdam, S., Panahi, A., Krim, H., Dai, L.: Deep dictionary learning: a parametric network approach. IEEE Trans. Image Process. **28**(10), 4790–4802 (2019)
12. Singhal, V., Majumdar, A.: Majorization minimization technique for optimally solving deep dictionary learning. Neural Process. Lett. **47**(3), 799–814 (2018)
13. Yarotsky, D.: Optimal approximation of continuous functions by very deep ReLU networks. arXiv preprint arXiv:1802.03620 (2018)
14. Lin, C.J.: On the convergence of multiplicative update algorithms for nonnegative matrix factorization. IEEE Trans. Neural Netw. **18**(6), 1589–1596 (2007)
15. Blakeley, P., et al.: Defining the three cell lineages of the human blastocyst by single-cell RNA-seq. Development **142**(18), 3151–3165 (2015)
16. Li, H., et al.: Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. Nat. Genet. **49**(5), 708 (2017)
17. Zheng, G.X., et al.: Massively parallel digital transcriptional profiling of single cells. Nat. Commun. **8**(1), 1–12 (2017)
18. Kolodziejczyk, A.A., et al.: Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. Cell Stem Cell **17**(4), 471–485 (2015)
19. Usoskin, D., et al.: Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. Nat. Neurosci. **18**(1), 145 (2015)
20. Yan, L., et al.: Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nat. Struct. Mol. Biol. **20**(9), 1131 (2013)