IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 23, 2020, accepted June 7, 2020, date of publication June 12, 2020, date of current version June 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3001986

DSCD: A Novel Deep Subspace Clustering Denoise Network for Single-Cell Clustering

ZHIYE WANG[1], YIWEN LU[1], CHANG YU[2], TAO ZHOU[2], RUIYI LI[1], AND SIYUN HOU[1]
[1]Department of Computer Science, Tongji University, Shanghai 201804, China
[2]Department of Mathematical Sciences, Tongji University, Shanghai 201804, China

Corresponding authors: Ruiyi Li (ryli@tongji.edu.cn) and Siyun Hou (housiyun@tongji.edu.cn)

ABSTRACT Single-cell RNA sequencing(scRNA-seq) technology has boomed in the past decade which makes it possible to study biological problems at the resolution of cellular-level. Currently, the research mainly focuses on exploring the cellular heterogeneity, involving studies about identifying cell type identification, cell lineage tracing, spatial model reconstruction of complex organizations, etc. Clustering analysis is always the most effective way in grouping single cells in previous studies. However, existing scRNA-seq clustering methods separate pre-processing and clustering tasks that complicated the problem. In addition, the emergence of big data further limits the traditional clustering algorithms' application on scRNA-seq data. Therefore, developing novel clustering methods and improving clustering accuracy for growing scRNA-seq data is a continuous task. In this paper, we propose a highly integrated **D**eep **S**ubspace **C**lustering **D**enoise Network named DSCD, which integrates denoise, dimension reduction and clustering in a unified framework. Based on the neural network architecture of autoencoder, DSCD discovers the low dimensional latent structure within scRNA-seq data from the compressed representation. Furthermore, we add a novel self-expressive denoise layer to learning the global relationships between single cells, which is the main innovation of DSCD. Experimental results on the synthetic data demonstrate the effectiveness of the novel denoise layer. From the clustering results on 5 real scRNA-seq datasets, we find that DSCD outperforms the related subspace clustering algorithms and state of the art methods. In conclusion, DSCD responds well to the rapidly increasing scRNA-seq data scale, greatly reduces human interference in dimension reduction and handles the noisy scRNA-seq data in proper way thus obtain a higher clustering accuracy.

INDEX TERMS Single cell RNA-seq data, auto-encoder, sparse self-express, spectral clustering.

I. INTRODUCTION

Traditional RNA sequencing (RNA-seq) technology measures the average expression of genes across many cells of different cell types, which may mask the real functional capacities of each cell type and thus hinder the study of cell heterogeneity. Fortunately, the emergence of scRNA-seq technology overcame the limitations in traditional bulk RNA-seq, which enables researchers to investigate the cellular heterogeneity from many aspects including determine cell types and predict cell fates, thus presenting more potential benefits for cell biology and clinical applications.

To mining valuable information from scRNA-seq data, researchers solve multiple tasks of single cell data from a computational perspective. Combined with the unique characteristics of scRNA-seq data, tailored approaches involved in feature selection, feature reduction, clustering, visualization an differentiated genes identification have been developed in recent years. Thereinto, clustering is the most effective way to study the cellular heterogeneity. It is an intuitive method to identify cell types from a large number of heterogeneous cells. Therefore, it is of great significance to improve the accuracy of clustering algorithm [1].

Formally, the process of clustering scRNA-seq data is as follows. Given a gene expression matrix, the row of the matrix represents the cell, the column represents the gene, each element in the matrix represents the expression value of the gene in the corresponding cell. The purpose of clustering is to group cells by measuring the similarity of genes. After clustering the cells, the cells would be divided into several categories, each type of cells has a specific biological meaning. There are many challenges to this task [2]: First, noise

The associate editor coordinating the review of this manuscript and approving it for publication was Shahzad Mumtaz.

which is caused by the cell cycle state [3], the difference between scale of gene pool [4] and the low RNA capture rate [5] will damage the potential biological signals and affect the single-cell sequencing data analysis [6]. Second, Batch effects, which are mainly due to daily changes in environmental conditions, such as temperature, machine calibration, or measurement efficiency [7]. Last, the increase in data scale is due to the continuous exponential growth of the number of cells that can be used for scRNA-seq analysis, which is resulted from the development of technology and the improvement in protocol. The number of cells collected can be increased to a million level by using the latest sequence technology called 'In situ barcoding' [8]. Therefore, new effective dimension reduction methods and methods that can deal with large data sets need to be developed.

II. LITERATURE REVIEW

At present, many research teams are studying the single-cell clustering method, as in [9], Justina Zurauskiene *et al.* proposed a new method, which combined principal component analysis and hierarchical clustering, and established a framework to describe the consistency of cell state. In SSC [10], Elhamifar *et al.* proposed a sparse subspace clustering algorithm to cluster data points on the union of low dimensional subspaces. However, in the sparse subspace clustering algorithm, the solution of the minimization problem of norm needs a lot of iterative process with high time complexity [11]. In Zheng's research [12], they used K-means to cluster droplet-seq data, K-means algorithm has high efficiency and scalability in dealing with massive data, but K-means is difficult to identify aspheric shape clusters [13]. Guo's research [14] is a processing flow of scRNA-seq data analysis. He firstly preprocessed the data, screened out the low expression genes according to preset standards (such as deleting genes expressed in less than n cells), and then used Z-SCORE to regularize the columns. Butler's research [15] is a tool for analyzing single cell transcriptome data, which provides a standard analysis process, including data normalization, finding differential genes, constructing SNN (shared nearest neighbors) graph and modular optimizer algorithm clustering. For the deep learning method, Gökcen Eraslan *et al.* combined the auto-encoder and zero expansion negative binomial distribution [16]. Through redesigning the loss function, they got a good "dropout" removal effect on the simulation data after training the distribution parameters. After getting "clean" data, they used K-means to cluster.

Through the introduction to previous studies, we can find that the basic structure of the former clustering framework is a two-step mode, including 'data preprocessing' and 'clustering'. The preprocessing methods include dimension reduction, feature extraction and gene normalization, while clustering methods are various, including density based, distance based, graph based and depth learning based methods [17]. There are four problems of these models. Firstly, these models lack integration and unity, preprocessing and clustering are not closely linked and integrated into a unified framework. Secondly, when dealing with a large number of high-dimensional sample, the limitations of traditional clustering algorithm are increasingly obvious, so it is urgent to introduce new methods to solve the clustering problem of scRNA-seq. Thirdly, previous studies have introduced too many subjective factors into the feature extraction which is based on experience, but experience is not always correct. Fourthly, related works lack the research of clustering itself, such as exploring the subspace structure hidden in cell groups.

In this paper, based on the assumption that similar data points are easier to be expressed by linear combination of similar data points, we proposed a deep neural network structure of unsupervised subspace clustering [18] for single cell clustering. This built on the deep auto-encoder structure, maps the input data onto a potential space, thus achieves the purpose of spontaneously dimension reduction. In order to simulate the effective "self-expressive" features in traditional subspace clustering and denoise the data, a new self-expressive denoise layer is introduced between encoder and decoder. The new self-expressive denoise layer can remove the noise and learn the pair affinities among all data points more accurately through a standard back propagation process.

III. METHOD

The algorithm flow is shown in Figure1. The original dataset is firstly divided into training set and validation set. Initialize pre-training network with random parameters. The training set is divided into mini-batches and then feed to the network. The loss value is calculated according to the loss function and the parameters of the pre-training network are updated. After the update, the validation set will be sent to the pre-training network to get the loss value, determining whether the loss value is the minimum. When the loss value achieved the minimum, stop the pre-training and save the current network parameters, then initialize the deep subspace denoise network parameters with the corresponding pre-training network location. Update parameters according to the newly designed loss function. When the training reaches up to 300 rounds, terminate the training and extract the self-expressive matrix located in the self-expressive layer. After constructing similarity matrix with the aid of the self-expressive matrix, clustering results is obtained by spectral clustering. The following parts will explain the components of the algorithm in detail, in which subsections 'self-expressive' and 'deep auto-encoders' are both components of the subsection 'deep subspace denoise network'.

A. SELF-EXPRESSIVE

Self-expressive means that, a sample can be expressed in terms of a few other complete samples from the same linear subspace [19]. That means, given multiple data points $\{S_i\}_{i=1,2,\ldots,K}$ extracted from linear subspace $\{X_i\}_{i=1,2,\ldots,N}$, any point in the subspace can be expressed as a linear combination of other points in the same subspace. If all the points are superposed into a column of data matrix X,

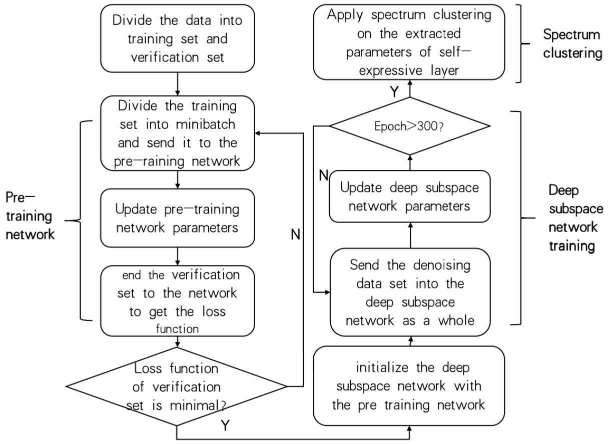
Divide the data into training set and verification set
Divide the training set into minibatch and send it to the pre−raining network
Update pre−training network parameters
end the verification set to the network to get the loss function
Loss function of verification set is minimal?
Y
N
Pre− training network
initialize the deep subspace network with the pre training network
Send the denoising data set into the deep subspace network as a whole
Update deep subspace network parameters
Epoch>300?
N
Y
Apply spectrum clustering on the extracted parameters of self− expressive layer
Spectrum clustering
Deep subspace network training

FIGURE 1. Flow chart of algorithm.

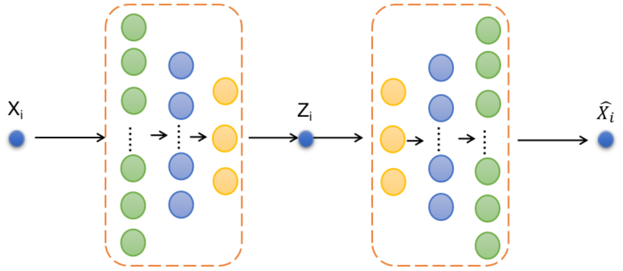

FIGURE 2. Architecture of deep auto-encoders.

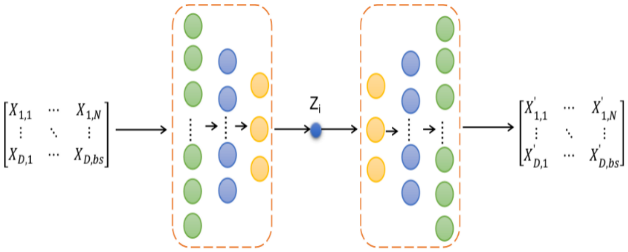

FIGURE 3. Architecture of pre-training network.

the self-expressive property can be expressed by $X = XC$, in which C is the self coefficient matrix. Generally speaking, C is not unique, so we look for the sparsest C to construct affinity matrix for spectral clustering.

$$\min_{C} \|C\|_p \; s.t. X = XC, \; (diag\,(C\,)= 0) \tag{1}$$

where $\| \bullet \|_p$ is the norm of any matrix, and the optional diagonal constraint on C prevents the trivial solution of sparse induced norm. In this paper, equality constraints are relaxed to regularization terms,like

$$\min_{C} \|C\|_p + \frac{\lambda}{2}\|X - XC\|_2^F \; s.t. diag\,(C\,)= 0) \tag{2}$$

However, self-expressive is only applicable to linear subspaces. In this paper, our goal is to learn an explicit mapping to facilitate the separation of subspaces. Therefore, in this paper, we built a neural network based on the deep auto-encoder.

B. DEEP AUTO-ENCODERS

Auto-encoder [20] is a data compression algorithm, in which the data compression and decompression functions are data-related, lossy, and learning from the samples automatically. The neural network of auto-encoder consists of two parts: encoder and decoder. The encoder compresses the input into a potential spatial representation, and the decoder aims to reconstruct the input from the hidden spatial representation. The goal of this part is to train a deep auto-encoder as shown in Figure2, so we introduce a new layer to present the concept of self-expressive. Loss function of auto-encoder is shown as follows.

$$L = \frac{1}{2}\left\|X' - X\right\|_F^2 \tag{3}$$

C. PRE-TRAINING NETWORK

In order to improve the training efficiency of the model, pre-training is introduced to this paper, and input the reconstructed single-cell data to the formal deep subspace model as the denoised data [21]. The structure of the pre-training network is the part except for the self-expressive layer in the deep subspace network, as shown in Figure 3. The network of pre-training includes two auto-encoding layers and two auto-decoding layers. The number of neurons in the encoding layer is the dimension of the genes, and the decoding layer is symmetrical with the encoding layer, and the latent layer has 256 neurons. The Adam optimizer is selected and the learning rate is set to 0.001. Different parameters have been tried, including increasing the layers and neurons in each layer. Basically, the model is supposed to make a balance between efficiency and precision. Additionally, due to the ’dropout’ phenomenon, which means that 70%-90% genes in each cell express zero. The ’dropout’ phenomenon contribute negative effect to the clustering, the remaining gene transform into protein which gathering in small group to function. Therefore, the number of latent layer(256) is corresponding to the functional numbers in the orders of magnitude.

In order to confirm the training destination and avoid over-fitting of network, we adopt the strategy of early termination in the process of pre-training. The data set is divided into training set and validation set. After each iteration, the respective error rate is calculated. When the error rate on the validation set reaches the minimum, the training would be stopped. Because if the training not be stopped, the error rate on the training set will continue to decrease, while the error rate on the validation set will increase. It means that the generalization ability of the proposed model will start to deteriorate. At this time, the whole data set can be sent into the model to get the denoised data for the deep subspace network. Then the parameters corresponding to the pre-training network and the deep subspace network are migrated as the

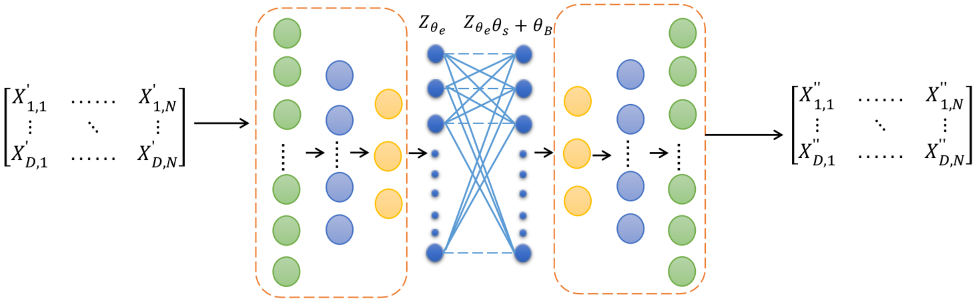
$\begin{bmatrix} X'_{1,1} & \cdots & X'_{1,N} \\ \vdots & \ddots & \vdots \\ X'_{D,1} & \cdots & X'_{D,N} \end{bmatrix}$
Z_{θ_e}
$Z_{\theta_e}\theta_s + \theta_B$
$\begin{bmatrix} X''_{1,1} & \cdots & X''_{1,N} \\ \vdots & \ddots & \vdots \\ X''_{D,1} & \cdots & X''_{D,N} \end{bmatrix}$

FIGURE 4. Architecture of deep subspace network.

initial parameters of the deep subspace network. This greatly increases the speed and efficiency of network training.

D. DEEP SUBSPACE DENOISE NETWORK

The network of deep subspace denoise network is shown as Figure4, which includes two auto-encoding layers, one self-expressive denoise layer and two auto-decoding layers. Since the auto-encoder is composed of encoder and decoder, the parameters of the auto-encoder Θ can also be decomposed into encoder parameter Θ_e and decoder parameter Θ_d. And let the decoder parameter Θ_d represent the output of the encoder $Z_{\Theta e}$.

Define the loss function as:

$$L(\Theta, C, B) = \left\| X - \widehat{X_\Theta} \right\|_F^2 + \lambda_1 \left\| C \right\|_p + \lambda_2 \left\| B \right\|_p + \lambda_3 \left\| Z_{\Theta_e} - (Z_{\Theta_e} C + B) \right\|_F^2 \quad (4)$$

where $\widehat{X_\Theta}$ represents the data reconstructed by auto-encoder, and the goal of auto-encoder reconstruction is to minimize the loss function. Since C and B can be regarded as the parameters of additional network layer, that is, the coefficient matrix in self expression and the noise bias, we can use the back propagation to solve Θ, C and B.

In formula 4, parameters λ_1, λ_2 and λ_3 are taken as the reciprocal of their dimension product to balance three terms in the loss function. The first term of the fraction aims to ensure the reconstruction accuracy of the auto-encoder, which achieves the dimension reduction of the encoder. The second term of the fraction guarantees the sparsity of the self-expressive matrix which is represented by C in the self-expressive denoise layer, and the norm constraint can achieve the sparsity when p is a positive integer, which has been proved in [10]. The third term is similar to the second term, which is sparse and aims at solving the noise in the data matrix, also makes sure that the similarity matrix reflects real relationship between data points. The forth term of fraction guarantees the validity of self-expressive matrix, because each data point can be represented by a linear combination of other points,this linear operation corresponds to a group of linear neurons without nonlinear activation.

In practical, the number of neurons in encoding layer is the dimension of genes(512) and the endecoding layer is symmetrical with the encoding layer. The middle self-expressive denoise layer maps the whole batch which means the whole dataset into the same dimension which called self-expressive and sends it to the decoder for decoding. The activation function of all layers in the network uses leaky corrected linear. The deep subspace network needs to be trained with the whole data set as a mini-batch. We choose the Adam optimizer and set the learning rate to 0.001 for the noise and similarity term, 0.0001 for the FC layers which have been trained in the pre-train process. It should be noted that since each batch is the whole data, it is necessary to increase the number of training rounds to ensure that the network is fully trained.

After getting the result of C, we use C to construct similarity matrix, then utilize spectral clustering to get the final clustering results.

E. SPECTRAL CLUSTERING

The spectral clustering algorithm is based on the spectral graph theory. Compared with the traditional clustering algorithm, it has the advantages of clustering on the sample space of any shape and converging to the global optimal solution.

The spectral clustering algorithm first defines an affinity matrix to describe the similarity of pairs of data points according to the given sample data set, and calculates the eigenvalues and eigenvectors of the matrix, then selects appropriate eigenvectors to cluster different data points. Algorithm 1 shows the algorithm flow chart of spectral clustering algorithm.

IV. RESULT

In this section, we firstly make a brief introduction about the evaluation criteria and datasets used in this paper. Then we generate a synthetic dataset as toy example to evaluate the effectiveness of the denoise layer. We consider two kinds of regularization method of C, (i) l_1 norm is the absolute addition of all values in the matrix, which will form DSCD-Net-l_1 network, (ii) l_2 norm is the square addition of all values in the matrix, which will form DSCD-Net-l_2 network. As a contrast, We use DSC-Net-l_1, DSC-Net-l_2 [22] and some other state-of-art methods as comparasion experiments.

Algorithm 1 Framework of spectral clustering

```
Input: Similarity matrix $S \in \mathbb{R}^{n\times n}$,number $k$ of clusters to
    constuct;
Output: Clusters $A_1, \dots, A_K$ with $A_i = j|y_j \in C_i$
  Construct a similarity graph. Let $W$ be its weighted
    adjacency matrix.
  Compute the first $K$ eigenvectors $v_1, \dots, v_k$ of $L$;
  Let $V \in \mathbb{R}_{n\times k}$ be the matrix containing the vectors
    $v_1, \dots, v_k$ as columns.
  For $i = 1, \dots, n$,Let $y_i \in \mathbb{R}_k$ be the vector correspond-
    ing to the $i$-th row of $V$.
  Cluster the points $(y_i)_{i=1,\dots,n}$ in $\mathbb{R}_k$ with the k-means
    algorithm into clusters $C_1, \dots, C_k = 0$.
```

A. EVALUATING CRITERIA

1) ADJUSTED RAND INDEX

The Rand coefficient requires a given actual category information C, assuming that K is a clustering result, a means that both C and K are elements of the same category, and b means that in C and K are elements of different categories, the RAND index is:

$$RI = \frac{a+b}{C_2^{n_{samples}}} \tag{5}$$

Numerator represent the number of samples with consistent attributes, they can belong to or do not belong to this class. a is the number of samples whose ground truth in the same class, and the prediction classification is also in the same class; b is the number of samples whose ground truth in different classes, and predicted classification are also in different classes. The denominator represents how many combinations that two samples have in a class, which is the total element that can be composed in the dataset. The range of RI is [0,1], and a larger value means that the clustering results match the real situation.

For random results, RI does not guarantee that the score is close to zero. In order to achieve "in the case of random clustering results, the indicator should be close to zero", the adjustment rand coefficient(ARI) was proposed, which has a higher degree of differentiation

$$ARI = \frac{RI - E[RI]}{max(RI) - E[RI]} \tag{6}$$

The value range of ARI is [-1,1]. A higher ARI value means that the clustering results match the ground truth. In a broad sense, ARI measures the degree to which the two data distributions fit.

2) SILHOUETTE COEFFICIENT

Silhouette coefficient(SC) is also an evaluation criterion of clustering. A good cluster is dense inside and sparse outside. The samples of the same cluster should be dense enough, and the samples of different clusters should be sparse enough. The Silhouette coefficient is to calculate the average distance a between a specific sample in the sample space and other samples in the cluster, and the average distance b between the sample and all samples in the nearest cluster. The Silhouette coefficients of all samples in the whole sample space are taken as the arithmetic mean, which is the performance criterion of clustering. The formula to calculate the silhouette coefficient can be expressed as below,

$$S(i) = \frac{b(i) - a(i)}{max[b(i), a(i)]} \tag{7}$$

The range of values for the silhouette coefficient is [-1, 1], and the closer the sample distance of the same category, the farther the distance of the different categories, the higher the value of silhouette coefficient.

TABLE 1. Overview of 5 scRNA-seq datasets.

Dataset	Cells	Genes	Clusters
GSE60361	3005	19972	9
GSE65525	2717	24175	4
GSE72056	4645	23686	7
GSE76312	2287	23384	7
GSE103322	5902	23685	9

B. DATA SOURCES

5 datasets are collected from ArrayExpress and GEO databases to measure the performance of clustering. Table 1 shows the brief information about datasets we used. GSE60361 studies the cells type of somatosensory cortex and hippocampus by appplying a recently developed, highly accurate and sensitive single-cell scRNA-seq method (STRT/C1) [23]. GSE60361 contains 9 types of cell, including Interneurons cells, S1 Pyramidal cells, CA1 Pyramidal cells, Mural cells, Endothelial cells, Microglia cells, Ependymal cells, Astrocytes cells and Oligodendropcytes cells. GSE65525 analyzes mouse embryonic stem cells, revealing in detail the population structure and the heterogeneous onset of differentiation after LIF withdrawal [24]. GSE65525 contains 4 types of cell, including Horizontal cells, Retinal ganglion cells, Amacrine cells and Bipolar cells. GSE72056 studies the diversity of expression states within melanoma tumors, and then obtained freshly resected samples, dissagregated the samples, sorted into single cells and profiled them by single-cell RNA-seq [25]. GSE72056 is composed of malignant cells and non-malignant cells, clusters of non-malignant cells are annotated as T cells, B cells, macrophages cells, endothelial cells, caner-associated fibroblasts(CAFs) and NK cells. GSE76312 has more than 2,000 single cells from patients with chronic myeloid leukemia, and then gene expression profiling was performed by single cell sequencing [26].

All of the data has undergone a basical filter process which in detail screen out the gene expressed in less than 5 cells.

C. THE EFFECTIVENESS OF DENOISE LAYER

In this section, we use toy examples to measure the effectiveness of the denoise layer in our algorithm.

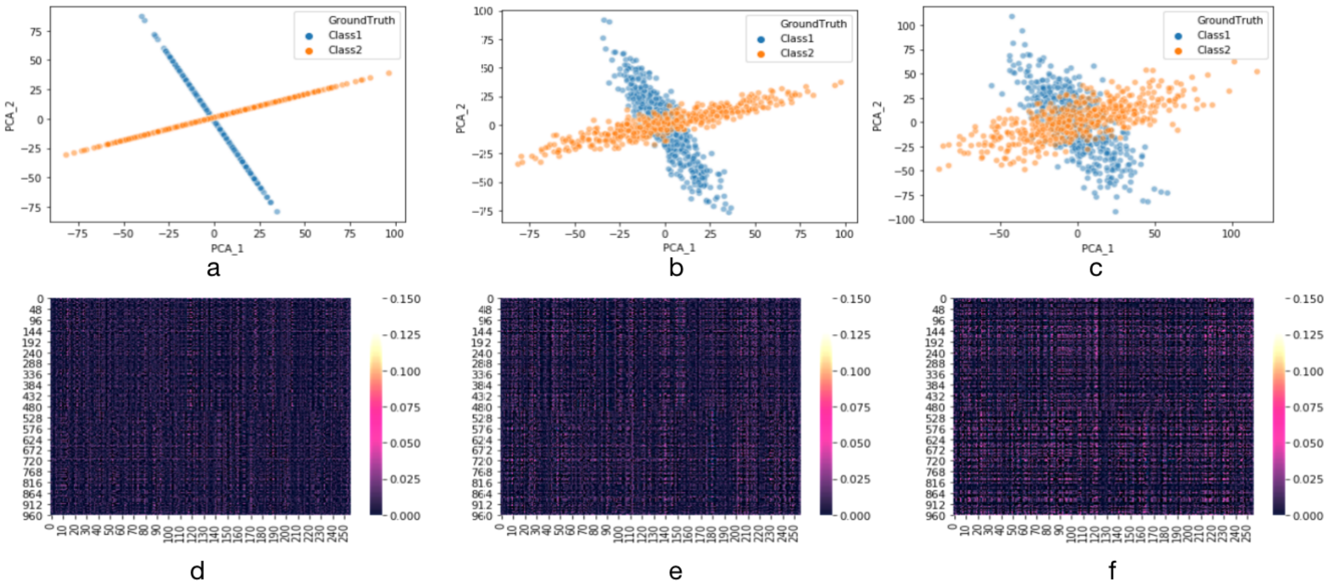

FIGURE 5. Effectiveness of denoise layer a.b.c. show the simulation data with no noise, σ=5 μ=0 and σ=10 μ=0 respectively.

The process of generating toy examples is described as following. Firstly, a set of data (N, d) which obeys a normal distribution is generated, where N is the number of samples for toy example and d is the number of hidden subspaces. Generates another set of data which obeys the normal distribution (d, D), where D is the number of sample properties, multiplying the two sets of data to the data with dimensions (N, D), as the first class. Use the same method to make the second class. In this experiment, $N = 2500$, $D = 3000$, $d = 1$.

To measure the effectiveness of the denoise layer in our algorithm, we generate two clusters of pure data samples artificially as in Figure.5a and add Gaussian white noise of different levels to samples as shown in Figure.5bc, then use DSC and DSCD to cluster them. We extract the final noise terms in LOSS function, which is marked as B in previous depiction, under the three different noise condition. Take the opposite number of the negative numbers in the matrix to ensure the noise start from 0. Demonstrate them in the format of heatmap in which the lighter the matrix is, the noisier the data is. When we add more noise to the original sample, the noise term will burden the noise and become lighter, thus purify the data and get higher accuracy. The heatmaps shown in Figure.5 confirm this. The heatmaps is an approach to demonstrate our novel self-expressive denoise layer is able to capture the noise containing in the data. We continuously add more gaussian noise into our toy example data and the heatmap shows the noise matrix capturing more noise, which becoming lighter.

To demonstrate the statement, we apply DSC and DSCD to the three datasets respectively and compare the clustering results with the index ARI. The results are shown in Figure6. We achieve better ARI under all the noise conditions. In specific, when there is no noise, result of the two methods are close to each other. As the noise level rises up, DSCD shows great superiority over DSC whom has no means to deal with noise.

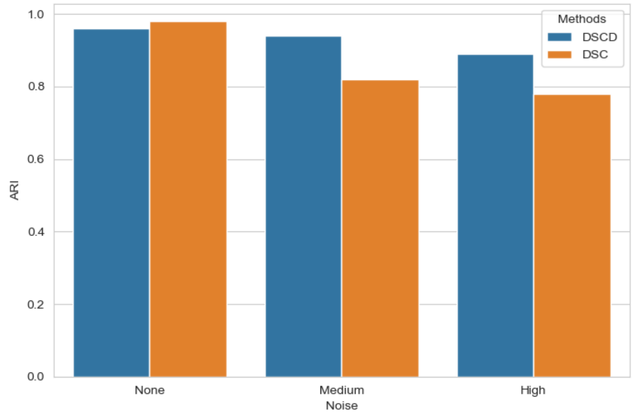

FIGURE 6. Comparision on ARI between DSC and DSCD.

D. EXPERIMENT ON REAL DATASET

After the experiment on toy examples, we apply the proposed algorithm to the actual scRNA-seq data. We first make a brief introduction to these methods. Justina Zurauskiene *et al.* proposed a new method named pcaReduce [9] combining principal component analysis and hierarchical clustering, and established a framework to describe the consistency of cell state. Matthew Amodio's study [21] is a multi task framework called SAUCIE, which integrates four important functions: clustering, batch processing and correction, visualization and inference. Because of it's innovative use of the parameters

TABLE 2. Comparison on Adjusted Rand index(ARI) between DSCD and other state of art single cell clustering methods.

dataset	pcaReduce	NMF	SAUCIE	CIDR	DCA	DSCD
GSE60361	0.400	0.183	0.176	0.369	0.337	**0.422**
GSE65525	0.622	0.655	0.477	0.683	0.805	**0.821**
GSE72056	0.291	0.315	0.286	0.104	0.229	**0.321**
GSE76312	0.025	0.024	0.056	**0.075**	0.029	0.053
GSE103322	0.321	0.384	0.195	0.309	**0.530**	0.388

TABLE 3. Comparison on Adjusted Rand index(ARI) between DSCD and DSC.

dataset	DSC-l_1	DSC-l_2	DSCD-l_1	DSCD-l_2
GSE60361	0.339	0.331	0.420	**0.422**
GSE65525	0.666	0.666	**0.821**	0.819
GSE72056	0.193	0.192	**0.321**	**0.321**
GSE76312	0.026	0.019	**0.053**	0.051
GSE103322	0.160	0.159	0.384	**0.388**

TABLE 4. Comparison on Silhouette Coefficient(SC) between DSCD and other state of art single cell clustering methods.

dataset	pcaReduce	NMF	SAUCIE	CIDR	DCA	DSCD
GSE60361	-0.066	-0.039	-0.235	-0.027	-0.009	**−0.015**
GSE65525	0.115	**0.185**	-0.085	-0.051	-0.027	0.022
GSE72056	0.011	**0.026**	0.001	-0.024	0.007	0.007
GSE76312	-0.07	**0.029**	-0.038	-0.018	-0.1183	-0.017
GSE103322	0.01	0.0086	-0.001	0.024	0.021	**0.07**

TABLE 5. Comparison on Silhouette Coefficient(SC) between DSCD and DSC.

dataset	DSC-l_1	DSC-l_2	DSCD-l_1	DSCD-l_2
GSE60361	-0.017	**0.003**	-0.014	-0.015
GSE65525	0.006	0.006	**0.022**	**0.022**
GSE72056	0.002	0.002	0.001	**0.007**
GSE76312	-0.079	-0.105	**−0.017**	-0.019
GSE103322	-0.021	-0.021	**0.012**	0.011

of the middle layer and self built binary activation function, it achieved good results. Gökcen Eraslan built a method called DCA which is based on Autoencoder and merges ZINB into the loss function [27]. After obtaining the imputed data, k-means is used for clustering. CIDR [28] takes dropout into account and achieves a fast results. NIMFA [29] is an open-source Python library that provides a unified interface to nonnegative matrix factorization algorithms. NMF [30] is an algorithm for non-negative matrix factorization, which learns holistic, not parts-based, representations. To sum up, we select five advanced methods including two deep-learning based methods and three traditional algorithms to prove the superiority of our newly proposed model over the above. Our comparison method used the default parameters and network structure of the open source code, all of which can be found in the github repository in the referenced paper.

1) ADJUSTED RAND INDEX

The following Table 2 and Table 3 show the comparison on Adjusted Rand index(ARI) between DSCD and other algorithms in real dataset experiment and the competition between DSC and DSCD. The experiment results of GSE65525, GSE72056 and GSE60361 are better than the other five state-of-art existing Single-cell clustering algorithms, and the performances of the other two datsets are also better than the most of the existing algorithms where DSCD occupies second and third rank on GSE76312 and GSE103322. The last line shows the total results on the five data sets in which DSCD gets the first place. It can be shown that our DCSD clustering algorithm is effective and performs excellent.

2) SILHOUETTE COEFFICIENT

The following Table 4 shows the comparison on Silhouette Coefficient(SC) between DSCD and other state-of-art algorithms which are designed specifically for single-cell clustering. Table 5 demonstrates the result competition of DSCD and DSC. The experiment results of GSE60361, GSE76312 and GSE103322 are better than other existing Single-cell clustering algorithms while rank highly on the other three results.At the same time, DSCD shows great advantage over DSC on Silhouette Coefficient.

V. CONCLUSION

A major challenge in developmental biology is to understand the genetic and cellular processes driving organ formation and differentiation of the diverse cell types that comprise the embryo [31]. Single-cell RNA sequencing technology is widely used in the quantitative study of single-cell RNA expression,which can help us to improve the understanding of human diseases. Many research team uses the technology of clustering, which is of great significance to scRNA-seq data analysis, to find cells subtype from a large number of heterogeneous cells. However, the current single-cell clustering methods have great limitations, such as lack of integration, not suitable for a large number of high-dimensional samples and so on.

Deep learning has achieved great success in fields such as computer vision, natural language processing, speech recognition and so on. However, deep learning has not been fully utilized in unsupervised tasks. With the continuous research in recent years, deep learning has become the hope of processing modern high-dimensional biological data sets [21], while auto-encoder can learn the features of data itself and reveal the structure of data, without defining similarity or distance measurement in the original data space like other dimension reduction methods [21]. Besides, sparse subspace clustering can obtain self-expressive matrix by finding sparse self-expressiveness, so as to reveal the characteristics of data points in multiple sub-spaces [10]. In this paper, we introduce self-expressive denoise layer and improve the deep neural network structure for unsupervised subspace clustering of single-cell clustering. Based on the deep auto-encoder and self-expressive denoise layer, the input data is mapped to the low dimensional space to get the self-expressive matrix, so as to mine the complex subspace structure in the data.

Experiment results show that the clustering effect of this algorithm is excellent, the evaluation index 'Adjusted Rand index' and 'Silhouette Coefficient' are better than other existing single cell clustering methods. The evaluation index ARI of this algorithm achieved at least 1.8 times of other existing algorithm.

Based on the deep neural network of unsupervised subspace clustering, our future research will be devoted to further improve the self-expressive denoise layer. Firstly, we need to increase its ability to cope with dropout, and further improve the accuracy of clustering according to the characteristics of data. The second step is to increase the ability of multitasking which enables the network to solve other types of problems in the analysis of single-cell sequencing, such as batch effect and visualization.

REFERENCES

[1] R. Petegrosso, Z. Li, and R. Kuang, "Machine learning and statistical methods for clustering single-cell RNA-sequencing data," *Briefings Bioinf.*, Jun. 2019.

[2] J. Tan, G. Doing, K. A. Lewis, C. E. Price, K. M. Chen, K. C. Cady, B. Perchuk, M. T. Laub, D. A. Hogan, and C. S. Greene, "Unsupervised extraction of stable expression signatures from public compendia with an ensemble of neural networks," *Cell Syst.*, vol. 5, no. 1, pp. 63–71, Jul. 2017.

[3] F. Buettner, K. N. Natarajan, F. P. Casale, V. Proserpio, A. Scialdone, F. J. Theis, S. A. Teichmann, J. C. Marioni, and O. Stegle, "Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells," *Nature Biotechnol.*, vol. 33, no. 2, pp. 155–160, Feb. 2015.

[4] C. A. Vallejos, D. Risso, A. Scialdone, S. Dudoit, and J. C. Marioni, "Normalizing single-cell RNA sequencing data: Challenges and opportunities," *Nature Methods*, vol. 14, no. 6, pp. 565–571, Jun. 2017.

[5] P. V. Kharchenko, L. Silberstein, and D. T. Scadden, "Bayesian approach to single-cell differential expression analysis," *Nature Methods*, vol. 11, no. 7, pp. 740–742, Jul. 2014.

[6] S. C. Hicks, F. W. Townes, M. Teng, and R. A. Irizarry, "Missing data and technical variability in single-cell RNA-sequencing experiments," *Biostatistics*, vol. 19, no. 4, pp. 562–578, Oct. 2018.

[7] U. Shaham, K. P. Stanton, J. Zhao, H. Li, K. Raddassi, R. Montgomery, and Y. Kluger, "Removal of batch effects using distribution-matching residual networks," *Bioinformatics*, vol. 33, no. 16, pp. 2539–2546, Aug. 2017.

[8] E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev, and S. A. McCarroll, "Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets," *Cell*, vol. 161, no. 5, pp. 1202–1214, May 2015.

[9] J. Žurauskienė and C. Yau, "PcaReduce: Hierarchical clustering of single cell transcriptional profiles," *BMC Bioinf.*, vol. 17, no. 1, p. 140, Dec. 2016.

[10] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, Nov. 2013.

[11] X. F. Weiwei, S. W. Wang, and X. Li, "Overview of sparse subspace clustering," *J. Automat.*, no. 8, pp. 3–14.

[12] G. X. Y. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, and J. Zhu, "Massively parallel digital transcriptional profiling of single cells," *Nature Commun.*, vol. 8, p. 14049, Jan. 2017.

[13] D. Sinha, A. Kumar, H. Kumar, S. Bandyopadhyay, and D. Sengupta, "DropClust: Efficient clustering of ultra-large scRNA-seq data," *Nucleic Acids Res.*, vol. 46, no. 6, p. e36, Apr. 2018.

[14] M. Guo, H. Wang, S. S. Potter, J. A. Whitsett, and Y. Xu, "SINCERA: A pipeline for single-cell RNA-seq profiling analysis," *PLOS Comput. Biol.*, vol. 11, no. 11, Nov. 2015, Art. no. e1004575.

[15] T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. Mauck, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija, "Comprehensive integration of single-cell data," *Cell*, vol. 177, pp. 1888–1902, Jun. 2019.

[16] G. Eraslan, L. M. Simon, M. Mircea, N. S. Mueller, and F. J. Theis, "Single-cell RNA-seq denoising using a deep count autoencoder," *Nature Commun.*, vol. 10, no. 1, p. 390, Dec. 2019.

[17] V. Svensson, R. Vento-Tormo, and S. A. Teichmann, "Exponential scaling of single-cell RNA-seq in the past decade," *Nature Protocols*, vol. 13, no. 4, pp. 599–604, Apr. 2018.

[18] O. Du Rr and B. Sick, "Single-cell phenotype classification using deep convolutional neural networks," *J. Biomol. Screening*, vol. 21, no. 9, pp. 998–1003, 2016.

[19] S. R. Rao, R. Tron, R. Vidal, and Y. Ma, "Motion segmentation via robust subspace separation in the presence of outlying, incomplete, or corrupted trajectories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Anchorage, AK, USA, Jun. 2008, pp. 1–8.

[20] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. Mach. Learn. (ICML)*, Helsinki, Finland, Jun. 2008, pp. 1096–1103.

[21] M. Amodio, D. van Dijk, K. Srinivasan, W. S. Chen, H. Mohsen, K. R. Moon, A. Campbell, Y. Zhao, X. Wang, M. Venkataswamy, A. Desai, V. Ravi, P. Kumar, R. Montgomery, G. Wolf, and S. Krishnaswamy, "Exploring single-cell data with deep multitasking neural networks," *Nature Methods*, vol. 16, no. 11, pp. 1139–1145, Nov. 2019.

[22] P. Ji, T. Zhang, H. Li, M. Salzmann, and I. Reid, *Deep Subspace Clustering Networks*.

[23] A. Zeisel, A. B. Muñoz-Manchado, S. Codeluppi, P. Lönnerberg, G. La Manno, A. Juréus, S. Marques, H. Munguba, L. He, C. Betsholtz, C. Rolny, G. Castelo-Branco, J. Hjerling-Leffler, and S. Linnarsson, "Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq," *Science*, vol. 347, no. 6226, pp. 1138–1142, Mar. 2015.

[24] A. M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, and M. W. Kirschner, "Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells," *Cell*, vol. 161, no. 5, pp. 1187–1201, May 2015.

[25] I. Tirosh, B. Izar, S. M. Prakadan, M. H. Wadsworth, D. Treacy, J. J. Trombetta, A. Rotem, C. Rodman, C. Lian, and G. Murphy, "Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq," *Sci.*, vol. 352, no. 6282, pp. 189–196.

[26] A. Giustacchini, S. Thongjuea, N. Barkas, P. S. Woll, B. J. Povinelli, C. A. G. Booth, P. Sopp, R. Norfo, A. Rodriguez-Meira, N. Ashley, L. Jamieson, P. Vyas, K. Anderson, Å. Segerstolpe, H. Qian, U. Olsson-Strömberg, S. Mustjoki, R. Sandberg, S. E. W. Jacobsen, and A. J. Mead, "Single-cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia," *Nature Med.*, vol. 23, no. 6, pp. 692–702, Jun. 2017.

[27] G. Eraslan, L. M. Simon, M. Mircea, N. S. Müller, and F. J. Theis, "Single-cell RNA-seq denoising using a deep count autoencoder," *Nature Commun.*, vol. 10, no. 1, p. 390, Dec. 2019.

[28] P. Lin, M. Troup, and J. W. K. Ho, "CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data," *Genome Biol.*, vol. 18, no. 1, p. 12, Dec. 2017.

[29] M. Zitnik and B. Zupan, "NIMFA: A python library for nonnegative matrix factorization," *J. Mach. Learn. Res.*, vol. 13, no. 30, pp. 849–853, 2012.

[30] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, p. 788, 1999.

[31] M. Guo, H. Wang, S. S. Potter, J. A. Whitsett, and Y. Xu, "SINCERA: A pipeline for single-cell RNA-seq profiling analysis," *Plos Comput. Biol.*, vol. 11, no. 11, 2015, Art. no. e1004575.

ZHIYE WANG was born in Hebei, China. He is currently pursuing the bachelor's degree with the Computer Science Department, Tongji University. His research interests include machine learning, neural networks, and bioinformatics.

YIWEN LU is currently pursuing the B.Sc. degree with Tongji University, Shanghai, China. Her research interests include deep learning, image processing, and data mining.

CHANG YU received the B.Sc. degree from Tongji University, Shanghai, China. His research interests include statistical analysis of data, machine learning, and parametric and non-parametric statistics.

TAO ZHOU received the B.Sc. degree from Tongji University, Shanghai, China. He is currently working with the Industries and Commercial Bank of China. His research interests include machine learning and data mining.

RUIYI LI is currently pursuing the Ph.D. degree in computer science with the Department of Computer Science and Technology, Tongji University, Shanghai, China. Her research interest includes bioinformatics.

SIYUN HOU is currently pursuing the Ph.D. degree in computer science with the Department of Computer Science and Technology, Tongji University, Shanghai, China. His research interest includes deep learning.

• • •