JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 25, No. 11: 1512~1518, Nov. 2021

단일세포 RNA-SEQ의 유전자 발현 군집화를 위한 변이 자동인코더 기반의 차원감소와 군집화

지상문*

Variational Autoencoder Based Dimension Reduction and Clustering for Single-Cell RNA-seq Gene Expression

Sang-Mun Chi*

*Professor, Department of Computer Science, Kyungsung University, Busan, 48434 Korea

요 약

단일세포 RNA-Seq 은 개별 세포의 유전자 발현을 제공하므로 세포마다 차등적인 고해상도 정보를 준다. 단일세포 RNA-Seq 자료에 대하여 군집화는 세포의 유형과 고수준의 생물 과정을 이해하기 위하여 수행된다. 매우 고차원이고 대용량인 단일세포 RNA-Seq을 효과적으로 처리하기 위하여, 본 논문은 변이 자동인코더를 사용하여 고차원의 자료 공간을 저차원의 잠재공간으로 변환하여, 보다 정확한 군집화를 수행할 수 있는 특징공간을 만든다. 차원이 축소된 잠재공간에 다양한 군집화 방법을 적용하는 접근을 다양한 전통적인 단일세포 RNA-Seq 군집화 방법과 성능을 비교하였다. 군집화 실험을 통하여, 제안한 방법은 기존 방법들보다 다양한 군집화 성능기준에서 성능이 개선되었다.

ABSTRACT

Since single cell RNA sequencing provides the expression profiles of individual cells, it provides higher cellular differential resolution than traditional bulk RNA sequencing. Using these single cell RNA sequencing data, clustering analysis is generally conducted to find cell types and understand high level biological processes. In order to effectively process the high-dimensional single cell RNA sequencing data fir the clustering analysis, this paper uses a variational autoencoder to transform a high dimensional data space into a lower dimensional latent space, expecting to produce a latent space that can give more accurate clustering results. By clustering the features in the transformed latent space, we compare the performance of various classical clustering methods for single cell RNA sequencing data. Experimental results demonstrate that the proposed framework outperforms many state-of-the-art methods under various clustering performance metrics.

키워드 : 군집화, 차원감소, 단일세포 RNA시퀀싱, 변이 자동인코더

Keywords : Clustering, Dimension reduction, Single-cell RNA-sequencing, Variational autoencoder

Received 22 August 2021, **Revised** 6 September 2021, **Accepted** 15 September 2021

*** Corresponding Author** Sang-Mun Chi(E-mail:smchiks@ks.ac.kr, Tel:+82-51-663-5146)
Professor, Department of Computer Science, Kyungsung University, Pusan, 48434 Korea

Open Access http://doi.org/10.6109/jkiice.2021.25.11.1512 print ISSN: 2234-4772 online ISSN: 2288-4165

ⓒThis is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License(http://creativecommons.org/licenses/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
Copyright Ⓒ The Korea Institute of Information and Communication Engineering.

Ⅰ. 서 론

유전자 발현은 유전체에 존재하는 일부 유전자들을 RNA로 전사하는 것으로, 이러한 분자적 수준의 생명과정에 대한 지식이 의학과 약품설계에 활발히 응용되고 있다[1-3]. 단일세포 RNA-Seq (scRNAseq: single- cell RNA-sequencing)은 생물조직에서 추출한 세포들의 유전자 발현 프로파일이 하나로 합쳐진 값만을 얻을 수 있는 기존 방법과 달리, 개별 세포 단위의 대용량 고해상도 전사체 분석 결과를 준다. 따라서 생장 단계나 생리적 조건에 따라 달라지는 유전체의 기능적인 요소를 알 수 있게 하므로 세포와 조직의 분자 구성요소, 전사체의 동적 변화, 유전자간의 조절관계를 조사할 수 있게 하여 생장과 병을 이해할 수 있게 한다[4-6]. 본 논문은 scRNAseq 자료를 사용하여 유사한 유전자 발현 양상을 가지는 세포들을 군집화한다. 군집화로부터 생물의 발생, 세포주기 진행, 환경 스트레스에 대한 반응, 병원성 감염, 암 등의 정보를 얻을 수 있다. 단일세포 RNA서열 분석은 세포의 낮은 생존성, mRNA의 낮은 회수율, cDNA 생성의 낮은 효율, 많은 세포를 대상으로 분석함에 따라서 유전자 발현을 검출하지 못하는 dropout 문제가 있다. 또한 유전자 발현의 변동이 크고, 수천-수십만개의 세포에 대하여 수만개 유전자를 분석대상으로 한다. 최근에는 희소하고 변동이 크면서, 대용량 고차원의 특징을 효과적으로 처리하기 위하여 심층학습기반의 방법들이 연구되고 있다[7 – 12].

군집화를 scRNAseq 자료에 적용한 최근의 방법으로 SIMLR[13]와 MPSSC[14]는 커널방법에 기반한 스펙트럴 군집화 방법으로 시간과 공간복잡도가 자료의 개수의 대하여 이차이고 행렬의 분해는 삼차이다. 따라서서 수만-수십만개의 세포로 구성된 자료에는 적용이 어렵다는 단점이 있다[15-16]. 또한, 심층학습 기반이 아닌 기존의 군집화는 시공간 복잡도가 매우 클 뿐만 아니라, 고차원의 scRNAseq 자료에 직접적으로 적용하면 정확도가 하락하는 단점이 있다. 최근에 제안된 일차의 시공간 복잡도를 가진 scDeepCluster, AAE-SC, DCA, [15-17]은 오토인코더에 기반한 방법으로 효율적인 특징표현을 학습하여 이를 군집화에 이용한다. scvis는 심층생성모델을 이용하여 저차원 공간의 구조를 찾아 군집에 이용한다[18]. 이러한 심층학습 기반의 방법은 고차원 자료에 대해 정확성 하락과 계산량 증가를 완화하기 위해 심층학습을 사용하여 복잡하고 고수준의 특징과 문맥정보를 자료로부터 추출한 후에, 비선형 사상을 사용하여 군집화에 적합한 형태로 변환하는 표현을 학습한다. 심층학습 기반의 군집화 방법은 선형의 계산량을 가지고 자료의 표현과 군집화를 통합하므로, 군집화에 t-분포나 자승오차와 같은 간단한 기준만을 사용한다. 따라서 자료 분포의 형상과 연결특성에 종속적인 우수한 성능을 보여주는 기존의 군집화 방법을 적용하기 어렵다. 본 논문에서는 고차원 자료를 선형의 시간복잡도로 자료를 변환한 특징공간에서 다양한 군집화를 수행한다. 본 논문에서는 VAE (Variational AutoEncoder) [19-20]를 사용하여 차원을 감소시킨다. 군집화 방법은 자료의 특성에 따라 성능의 우위가 달라지므로, 차원이 감소된 scRNAseq 자료에 다양한 군집화 방법을 적용하여 적합한 군집화 방법을 찾는다. 간단한 이차원 자료의 경우에도 자료 분포의 형상에 따라 각기 다른 군집화 방법이 최적이므로, 수만 차원의 scRNAseq 자료의 경우에는 자료 분포의 특성이 더 다양하므로 적합한 군집화를 선택해야할 필요성이 더 크다. 또한, 고차원의 scRNAseq 자료에 기존의 군집화 방법을 직접적으로 적용하면 정확도가 떨어지나, 잠재공간으로 변환된 자료에 대하여는 효과적으로 적용될 수 있다.

본 논문의 구성은 다음과 같다. 2장에서 scRNAseq의 고차원 자료를 VAE로 차원을 축소하는 방법을 살펴보고, 3장에서는 심층학습 기반이 아닌 군집화 방법들과 군집화 성능을 측정하는 방법을 알아본다. 4장에서는 scRNAseq 자료에 대한 군집화 실험을 수행하고, 5장에서 결론을 맺는다.

Ⅱ. 단일세포 RNA-Seq 자료의 VAE기반의 차원축소

이장에서는 scRNAseq 자료를 VAE를 사용하여 저차원으로 변환하는 방법을 설명한다.

scRNAseq 자료는 수천-수십만개의 세포에 대한 수만개의 유전자의 발현량을 측정한 값이다. 본 논문에서는 각 세포마다의 유전자 발현량이 하나의 자료를 구성하고, 유전자의 수가 자료의 차원이 된다. 측정대상의 유전자 개수가 많으므로 고차원 자료인데, 이를 VAE를 사용하여 저차원의 잠재변수로 변환한다. VAE는 식 (1)

의 생성과정에 따라 학습자료 X의 확률 $P(X)$를 최대화하는 잠재변수 z와 파라미터 θ를 찾는다.

$$P(X) = \int P(X|z;\theta)P(z)dz \quad (1)$$

여기서, $P(X|z;\theta)$는 파라미터 θ를 가지고 $P(X|z)$를 근사하는 모델이다. 또한, 잠재변수 z는 X를 생성하기에 적합한 값이어야 하는데, 이를 함수 $Q(z|X;\phi)$를 이용하여 결정한다. VAE에서는 확률 모델인 $P(X|z;\theta)$와 $Q(z|X;\phi)$를 신경망을 이용하여 구성하고, 각각 디코더와 인코더 역할을 수행한다. 식 (1)을 Kullback-Leibler 발산 $KL(q,p) = E_q[\log q - \log p]$을 사용하여 다음과 같이 재구성한다[19-20].

$$\log P(X) - KL(Q(z|X;\phi), P(z|X;\theta))$$
$$= E_{z\sim Q}[\log P(X|z;\theta)] - KL(Q(z|X;\phi), P(z)) \quad (2)$$
$$= E_{z\sim Q}[\log P(X,z;\theta) - \log Q(z|X;\phi)] \quad (3)$$

항상 $KL() \geq 0$이므로 식 (2)와 식 (3)은 $\log P(X)$의 하한이 되고, ELBO (Evidence Lower BOund)라 불린다. 하한 ELBO를 최대화하는 파라미터 θ,ϕ를 찾아서 $P(X)$도 최대화되기를 기대할 수 있는 신경망을 구성한다.

본 논문에서는 VAE를 구현할 때 널리 사용되고 있는 가정 중에서 $Q(z|X;\phi)$를 대각성분 이외는 0인 공분산 행렬을 가지는 정규분포라는 가정을 사용하였다. 입력 X에 대해 공분산 $\sigma(X) = \mathrm{diag}(\sigma_1,\sigma_2,...,\sigma_k)$와 평균 $\mu(X)$를 출력으로 갖는 인코더 신경망을 구성 한 후에 잠재변수를 $z = \mu(X) + \sigma(X) \odot \epsilon$ 으로 생성하였는데, $\epsilon \sim N(0,I)$, $\odot$는 원소별 곱셈, k는 잠재공간의 차원, I는 단위행렬이다. 본 논문은 ELBO를 최대화하기 위하여 식 (2)의 형태를 사용하였다. 식 (2)의 $E_{z\sim Q}$를 구할 때 ϵ을 하나만 무작위 추출하여 생성한 하나의 잠재변수만 생성하는 간략화된 방법을 사용하였다. 식 (2)의 첫째 항을 근사하기 위하여 $Q(z|X;\phi)$로 무작위 추출한 z를 디코더 신경망으로 변화하여 $\hat{X}$를 구한다. 또한, $P(X|z;\theta)$가 $N(\hat{X},I)$의 분포를 갖는다고 가정하면 $\log P(X|z;\theta)$은 식 (4)에 비례하므로 이 식을 사용하여 식 (2)를 구성하였다.

$$-0.5 * (X-\hat{X})^2. \quad (4)$$

식 (2)의 둘째 항의 $Q(z|X;\phi)$는 $N(\mu(X),\sigma(X)I)$, $P(z)$는 표준 정규분포라고 가정하면 KL 발산은 식 (5)와 같다[19-20].

$$-\frac{1}{2}\sum_{i=1}^{k}(1+\log(\sigma_i^2)-\mu_i^2-\sigma_i^2). \quad (5)$$

본 논문에서는 인코더와 디코더를 이루는 신경망의 파라미터는 식 (4)와 식 (5)로 구성된 식 (2)를 최대화하도록, 식 (2)의 음의 값을 확률적 경사 하강법으로 최적화 하였다.

Ⅲ. 잠재공간에서의 군집화

이 장에서는 고차원 scRNAseq 자료를 VAE의 인코더로 변환하여 만들어진 잠재공간의 특징벡터를 군집화하기 위하여 본 논문에서 사용한 방법을 알아보고, 군집화의 성능 측정 방법을 설명한다.

군집화는 자료 $\{x_1, x_2, \cdots, x_n\}$들을 유사한 특성을 갖는 자료들끼리 모아서 몇 개의 집단으로 분할하는 것이다. 본 논문은 scikit-learn 버전 0.24.1 [21]의 군집화를 사용하였다. Affinity Propagation은 모든 자료의 쌍에 대해서 아래 식이 수렴할 때까지 반복한다[22].

$$r(i,k) = s(i,k) - \max_{k' \neq k}[a(i,k') + s(i,k')], \quad (6)$$
$$a(i,k) = \min[0, r(k,k) + \sum_{i' \neq i,k} r(i',k)],$$

여기서, $i \neq k$ 일때는 $s(i,k) = -\parallel x_i - x_k \parallel^2$, $s(k,k)$는 사용자가 preference로 지정하는 값으로서 클수록 최종적인 군집의 개수가 증가한다. 또 다른 사용자 지정값 λ는 이전 반복 값을 반영 정도를 결정한다.

$$r_{t+1}(i,k) = \lambda r_t(i,k) + (1-\lambda) r_{t+1}(i,k), \quad (7)$$
$$a_{t+1}(i,k) = \lambda a_t(i,k) + (1-\lambda) a_{t+1}(i,k).$$

식 (6)이 수렴이 된 후에, $r(i,k) + a(i,k)$인 행렬을 만들고, 각 행에서 가장 큰 k가 이 자료를 대표하는 중심이 되고, 같은 중심을 갖는 자료는 같은 군집에 소속된다. Agglomerative 군집화는 자료들을 차례로 병합하여 원하는 수의 군집을 만든다[21]. 본 논문에서는 자료를 병합하기 위해 Ward 기준을 사용하였는데, 이는 모든 군집의 분산의 합을 최소화하도록 한다. 또한, 연결 제

한을 사용하여 인접한 군집끼리만 병합되도록 하였다. BIRCH는 부군집 (subcluster)들의 정보를 이용하여 나무구조를 만들어 가면서 군집화를 수행한다. 군집화의 특징을 조절하는 파라미터로 threshold는 기존의 부군집에 추가되는 자료와의 거리를 제한하고, branching factor는 하나의 노드에 속하는 부군집의 수를 제한한다 [21]. GMM (Gaussian Mixture Model)은 여러 개의 가우시안 확률밀도함수의 혼합으로 자료를 모델링한다 [21]. 각각의 가우시안 분포함수의 공분산이 각기 다른 원소를 가지는 full, 같은 값을 공유하는 tied가 있다. 또한, 각각의 가우시안 분포의 공분산이 대각원소만 가지는 diag, 하나의 분산만을 갖는 spherical이 있다. 복잡한 파라미터가 없고 사용이 간단하여 널리 이용되는 K-means는 군집의 중심을 찾는다. 최적의 중심들은 각 군집의 분산의 합을 최소화시키도록 선택되어 진다. Spectral 군집화는 자료들 간의 유사도 행렬로부터 고d 윳값이 큰 순서로 고유벡터를 추출하여 자료를 저차원으로 변환하는데 이용한다[23]. 이렇게 저차원으로 변환된 자료에 대하여 K-means 방법으로 군집화 한다.

군집화 성능척도는 논문[15-16]에서 사용한 CA (Clustering Accuracy), NMI (Normalized Mutual Information)와 ARI (Adjusted Rand Index)을 사용하였다. CA는 예측된 군집의 번호를 실제 군집 번호로 일대일 함수로 최적으로 변환했을 때의 일치도이다.

$$\mathrm{CA} = \max_{m \in M} \frac{\sum_{i=1}^{n} \delta(l_i, m(p_i))}{n}, \quad (8)$$

여기서, i는 자료 번호, n은 자료의 개수, M은 군집화 방법이 예측한 군집번호를 실제 군집번호로 사상하는 함수들의 집합, δ는 두 입력 값이 같을 때 1이고 나머지는 0인 함수이다. NMI는 두 클러스터의 상호정보량을 두 클러스터의 엔트로피의 최댓값으로 나누어 정규화한 값이다.

$$\mathrm{NMI} = \frac{I(T,P)}{\max(H(T), H(P))}, \quad (9)$$

여기서, $I(T,P) = \sum_i \sum_j \frac{|t_i \cap p_j|}{n} log \frac{n|t_i \cap p_j|}{|t_i|\ |\mathrm{p_j}|}$ 이고 $H(P) = -\sum_j \frac{|p_j|}{n} log \frac{|p_j|}{n}$, t_i는 실제 군집 번호 i에 속하는 자료들의 집합, p_j는 군집번호 j로 예측된 자료들의 집합, $|\cdot|$는 집합의 크기이다. ARI는 군집화 방법을 수행한 결과에서 각각의 자료 쌍에 대하여 같은 군집에 속하면 1, 아니면 0으로 나타내고, 이 값을 실제 군집화 참값을 사용하여 마찬가지로 구성한 값과 일치하는 비율인 랜드지수를 구하고 이를 재조정한다. 재조정을 위하여 무작위로 생성한 군집화에서 최대 랜드지수에서 평균 랜드지수를 뺀 값으로 랜드지수를 나눈다.

$$\mathrm{ARI} = \frac{\binom{n}{2}(a+d) - [(a+b)(a+c)+(c+d)(b+d)]}{\binom{n}{2}^2 - [(a+b)(a+c)+(c+d)(b+d)]}, \quad (10)$$

단, U와 V를 두 개의 군집화 결과라고 할 때, a는 두 개의 군집화 결과에서 각 자료 쌍에 대하여 같은 군집에 속할 때의 개수, b는 다른 군집에 속하는 자료 쌍의 개수, c는 U에서는 같은 군집에 속하고 V에서는 다른 군집에 속하는 자료 쌍의 개수, d은 U에서는 다른 군집에 속하고 V에서는 같은 군집에 속하는 자료 쌍의 개수이다.

Ⅳ. 단일세포 RNAseq 군집화 실험

이 장에서는 scRNAseq 자료에 대한 군집화 실험을 수행하여 여러 가지 방법의 성능을 비교한다.

실험에 사용한 자료는 PBMC 4k 세포들이며 10X 지노믹스 플랫폼 (10X PMBC)을 사용하여 얻어졌다[24]. 세포 개수는 4,271개, 유전자 개수는 16,449개이며, 8개의 군집으로 구성되어져 있다. 스펙트럴 기반의 군집화 방법인 MPSSC, SIMLR의 경우에는 공간복잡도가 이차이므로 많은 자료를 대상으로 실험할 수 없으므로, 논문[15-16]에서 선택한 세포 개수를 2,100개로 이루어진 자료로 비교실험 하였다. 자료를 구성하는 각 세포들의 값의 합이 모두 같도록 크기를 정규화하고, 로그변환을 수행한 후에, 평균이 0, 분산이 1이 되도록 정규화 하였다.

제안한 방법은 Tensorflow 2.4.1 gpu 버전을 사용하여 VAE를 구현하였고, NVIDIA RTX 3080Ti GPU로 구성된 리눅스 시스템으로 실험하였다. VAE 인코더의 은닉 계층의 크기는 (2048, 32), 디코더의 은닉 계층의 크기는 (32, 128)이고, VAE 입력과 출력의 크기는 유전자 개수 16,449개이다. 신경망은 학습률 0.004, 배치크기 16, 반복회수 30번으로 하였고, 최적화 방법으로

Adam[25]을 사용하였다. 본 논문에서는 2장의 VAE기반의 차원축소를 사용하여 고차원 scRNAseq 자료를 잠재공간(=32차원)으로 변환하였다. VAE로 생성한 10개의 잠재공간의 자료에 대해 Affinity propagation, Agglomerative, BIRCH, GMM, K-means, Spectral 군집화를 적용한 결과의 평균을 표시하였다. 비교 대상의 방법은 PCA+k-means, SIMLR, MPSSC, DEC, scvis, scDeepCluster와 AAE-SC로서 논문[15-16]의 결과를 사용하였다.

Table. 1 Performance Comparison of several scRNAseq clustering methods.

Method	CA	NMI	ARI
PCA+k-means	56.93	63.58	48.62
SIMLR	62.13	72.29	51.93
MPSSC	76.29	73.59	65.87
DEC	61.62	60.53	52.05
scvis	85.30	75.35	75.05
scDeepCluster	82.58	77.52	72.91
AAE-SC	**87.26**	81.31	81.32
VAE+Affinity propagation	85.54	80.99	82.12
VAE+Agglomerative	84.21	78.90	79.04
VAE+BIRCH	84.49	79.54	79.57
VAE+GMM	81.54	78.42	77.15
VAE+K-means	86.06	**81.46**	**82.76**
VAE+Spectral	85.38	80.97	81.30

표 1에서 Affinity propagation의 성능은 preference = ‐100, -150, -200, -250, -300, -350, -400에 대하여 λ = 0.5, 0.6, 0.7, 0.8, 0.9을 적용하였을 경우에 최고의 성능을 보이는 preference = -350, λ = 0.9일 때의 값이다. λ = 0.9에서는 preference의 넓은 범위에서 좋은 성능을 보였다. Agglomerative는 자료를 병합하기 위한 기준인 linkage 파라미터로 ward, complete, average, single을 실험하였고, 가장 성능이 좋은 ward를 표에 나타내었다. BIRCH의 threshold = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, branching factor = 10, 20, 30, 40, 50, 60, 70에 대하여 실험한 결과 threshold = 0.2, branching factor = 50에서의 가장 높은 성능을 표에 나타내었다. threshold 0.1, 0.2, 0.3에서 넓은 범위의 branching factor에서 높은 성능을 보였다. GMM은 공분산의 종류를 full, tied, diag, spherical에 대하여 실험하였는데, 성능이 비슷하였다. 표 1에는 가장 성능이 좋은 tied의 결과를 나타내었다. K-means는 scikit-learn의 기본적인 파라미터를 그대로 사용하였다. Spectral 군집은 유사도 행렬을 구성할 때에 affinity 파라미터를 nearest_neighbors를 사용할 때는 고유벡터의 개수 n_components = 4, 8, 12, 16, 20, 최근접 그래프를 구성할 때 사용하는 근접한 자료의 개수 n_neighbors=20, 40, 60, 80을 사용하였다. 또한 affinity 파라미터를 rfb를 사용하여 radial basis 함수를 사용하고, n_components = 4, 8, 12, 16, 20개를 사용하였다. 표 1에는 가장 높은 성능을 보인 nearest_neighbors를 사용하고 n_components = 8 n_neighbors = 40 일 때를 나타내었다. VAE 인코더의 은닉 계층의 크기를 (1024, 32)와 (512, 128, 32)로 변화를 주고, 여러 군집화방법을 적용하였을 경우에는 표 1과 유사하지만 약간 낮은 성능을 보였다.

일반적으로 군집화 개수를 미리 지정하는 방법인 Agglomerative, BIRCH, GMM, K-means, Spectral의 성능이 높았다. 사전실험을 통하여 군집의 수를 미리 지정하지 않고, 파라미터로 조정하여 군집을 수행하는 방법인 DBSCAN[26], OPTICS[27]는 높은 성능을 얻지 못하였다. Affinity propagation은 파라미터를 적절히 지정하였을 경우에 표 1의 높은 성능을 얻었다. 제안한 방법들은 CA 기준으로 AAE-SC를 제외한 기존의 심층학습 기반의 군집화보다 세 개의 성능기준에서 성능이 높았다. 특히, VAE+K-means 방법은 안정적으로 높은 성능을 보였다. 이는 Kullback-Leibler 발산을 사용하여 잠재공간의 분포를 표준 정규분포에서 크게 벗어나지 않게 하였기 때문이고, 군집화에 잠재변수의 특성에서 평균만을 사용하였기 때문이라 판단된다.

Ⅴ. 결 론

본 논문에서는 단일세포 RNA 서열을 사용하여 군집화를 통하여 세포유형을 구분하였다. 매우 고차원이고 대용량인 자료를 처리하기 위하여 제안한 방법은 변이 자동인코더로 차원이 축소된 잠재공간에서 다양한 군집화를 적용할 수 있는 장점을 활용하여 성능을 향상시켰다.

본 논문에서는 변이 자동인코더가 생성하는 잠재공간의 특징 중에서 평균만을 사용하였으나, 향후에는 보

다 자료 분포의 형상과 연결특성에 종속적인 군집화를 수행하기 위하여 공분산행렬의 활용이 필요하다 판단된다.

ACKNOWLEDGEMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(2021R1I1A304651111)

References

[1] S. K. Kim, J. Lund, M. Kiraly, K. Duke, M. Jiang, J. M. Stuart, A. Eizinger, B. N. Wylie, and G. S. Davidson, “A gene expression map for Caenorhabditis elegans,” *Science*, vol. 293, no. 5537, pp. 2087-2092, Sep. 2001.

[2] M. N. Arbeitman, E. E. Furlong, F. Imam, E. Johnson, B. H. Null, B. S. Baker, M. A. Krasnow, Ma. P. Scott, R. W. Davis, and K. P. White, “Gene expression during the life cycle of Drosophila melanogaster,” *Science*, vol. 297, no. 5590, pp. 2270-2275, Sep. 2002.

[3] F. Emmert-Streib, M. Dehmer, and B. Haibe-Kains, “Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks,” *Front. Cell Dev. Biol.*, vol. 2, no. 38, Aug. 2014.

[4] A. E. Saliba, A. J. Westermann, S. A. Gorski, and J. Vogel, “Single-cell RNAseq: advances and future challenges,” *Nucleic Acids Res.*, vol. 42, no. 14, pp. 8845-8860, Aug. 2014.

[5] C. W. Shields, C. D. Reyes, and G. P. Lopez, “Microfluidic cell sorting: a review of the advances in the separation of cells from debulking to rare cell isolation,” *Lab Chip*, vol. 15, no. 5, pp. 1230-1249, Mar. 2015.

[6] J. Tanevski, T. Nguyen, B. Truong, N. Karaiskos, M. Er. Ahsen, X. Zhang, C. Shu, K. Xu, X. Liang, Y. Hu, H. V. V. Pham, L. Xiaomei, T. D. Le, A. L. Tarca, G. Bhatti, R. Romero, N. Karathanasis, P. L.oher, Y. Chen, Z. Ouyang, D. Mao, Y. Zhang, M. Zand, J. Ruan, C. Hafemeister, P. Qiu, D. Tran, T. Nguyen, A. Gabor, T. Yu, E. Glaab, R. Krause, P. Banda, DREAM SCTC Consortium, G. Stolovitzky, N. Rajewsky, J. Saez-Rodriguez, and P. Meyer, “Predicting cellular position in the Drosophila embryo from single-cell transcriptomics data,” *bioRxiv*, 2019. doi: doi.org/10.1101/796029.

[7] T. Ching, D. S. Himmelstein, B. K. Beaulieu-Jones, A. A. Kalinin, B. T. Do, G. P. Way, E. Ferrero, P. M. Agapow, M. Zietz, M. M. Hoffman, W. Xie, G. L. Rosen, B. J. Lengerich, J. Israeli, J. Lanchantin, S. Woloszynek, A. E. Carpenter, A. Shrikumar, Ji. Xu, E. M. Cofer, C. A. Lavender, S. C. Turaga, A. M. Alexandari, Z. Lu, D. J. Harris, D. DeCaprio, Y. Qi, A. Kundaje, Y. Peng, L. K. Wiley, M. H. S. Segler, S. M. Boca, S. J. Swamidass, A. Huang, A. Gitter, and C. S. Greene, “Opportunities and obstacles for deep learning in biology and medicine,” *J. R. Soc. Interface*, vol. 15, no. 141, Apr. 2018.

[8] J. Ding, A. Condon, and S. P. Shah, “Interpretable dimensionality reduction of single cell transcriptome data with deep generative models,” *Nat Commun.*, vol. 9, no. 2002, May. 2018.

[9] G. Eraslan, L. M. Simon, M. Mircea, N. S. Mueller, and F. J. Theis, “Single-cell RNA-seq denoising using a deep count autoencoder,” *Nat Commun.*, vol. 10, no. 390, Jan. 2019.

[10] T. Wang, T. S. Johnson, W. Shao, Z. Lu, B. R. Helm, J. Zhang, and K. Huang, “BERMUDA: a novel deep transfer learning method for single-cell RNA sequencing batch correction reveals hidden high-resolution cellular subtypes,” *Genome Biol.*, vol. 20, no. 165, Aug. 2019 .

[11] M. Amodio, D. 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,” *Nat. Methods*, vol. 16, pp. 1139-1145, Oct. 2019.

[12] L. Xiong, K. Xu, K. Tian, Y. Shao, L. Tang, G. Gao, M. Zhang, T. Jiang, and Q. C. Zhang, “SCALE method for single-cell ATAC-seq analysis via latent feature extraction,” *Nat Commun.*, vol. 10, no. 4576, Oct. 2019.

[13] B. Wang, J. Zhu, E. Pierson, D. Ramazzotti and S. Batzoglou, “Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning,” *Nat. Methods*, vol. 14, pp. 414-416, Mar. 2017.

[14] S. Park and H. Zhao, “Spectral clustering based on learning similarity matrix,” *Bioinformatics*, vol. 34, no. 12, pp. 2069-2076, Feb. 2018.

[15] T. Tian, J. Wan, Q. Song, and Z. Wei, “Clustering single-cell RNA-seq data with a model-based deep learning approach,” *Nature Mach. Intell.*, vol. 1, pp. 191-198, Apr. 2019.

[16] Y. Wu, Y. Guo, Y. Xiao, and S. Lao, “AAE-SC: A scRNA-Seq Clustering Framework Based on Adversarial

Autoencoder," *IEEE Access*, vol. 8, pp. 178962-178975, Sep. 2020.

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

[18] J. Ding, A. Condon, and S. P. Shah, "Interpretable dimensionality reduction of single cell transcriptome data with deep generative models," *Nat. Commun.*, vol. 9, no. 2002, May. 2018.

[19] C. Doersch, "Tutorial on Variational Autoencoders," arXiv:1606.05908v3, 2021.

[20] D. P. Kingma and M. Welling, "An Introduction to Variational Autoencoders," *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, pp. 307-392, Nov. 2019.

[21] F. Pedregosa, G. Varoquaux, A. GRamfort, V. Miche, and B. Thirion, "Scikit-learn: Machine Learning in Python," *JMLR*, vol. 12, pp. 2825-2830, 2011.

[22] B. J. Frey and D. Dueck, "Clustering by Passing Messages Between Data Points," *Science*, vol. 315, no. 5814, pp. 972-976, Feb. 2007.

[23] U. Luxburg, "A Tutorial on Spectral Clustering," *Statistics and Computing*, vol. 17, pp. 395-416, 2007.

[24] 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, J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C.r McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson, and J. H. Bielas, "Massively parallel digital transcriptional profiling of single cells," *Nat. Commun.*, vol. 8, no. 14049, Jan. 2017.

[25] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *ICLR (Poster)*, 2015.

[26] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "DBSCAN revisited, revisited: why and how you should (still) use DBSCAN," *ACM Transactions on Database Systems*, vol. 42, no. 3, pp. 1-22, 2017.

[27] E. Schubert and M. Gertz, "Improving the Cluster Structure Extracted from OPTICS Plots," *Proc. of the Conference LWDA*, pp. 318-329. 2018.

지상문(Sang-Mun Chi)

1991년 서울대학교 수학교육학과 졸업(이학사)
1993년 한국과학기술원 수학과 졸업(이학사)
1998년 한국과학기술원 전산학과 졸업(공학박사)
1993년 ~ 2000년 삼성전자 무선사업부 선임연구원
2001년 ~ 현재 경성대학교 소프트웨어학과 교수
※관심분야 : 딥러닝, 생물정보학, 계산금융