Single-cell RNA-seq 解析における Star clustering と GMVAE を用いたデータのクラスタ数の探索

EP19011 伊藤啓太　　指導教授：山下隆義

1. はじめに

病気や障害の原因は，異常な細胞の遺伝子情報をタンパク質として発現していることである．病気や障害の原因究明には，タンパク質を発現している異常な細胞を特定する必要がある．細胞の解析手法には，深層生成モデルの Gaussian-mixture Variational AutoEncoder (GMVAE) [1] を用いた Single-cell RNA-seq 解析手法がある．しかし，遺伝子発現量データは細胞の種類が不明である．そのため，GMVAE に真のクラスタ数を入力することが不可能であり，適切な潜在空間の獲得が困難である．そこで，本研究では，Star clustering によりクラスタ数を予測し，GMVAE に入力する．これにより，GMVAE を用いて遺伝子発現量データの最適な潜在空間を獲得できる．

2. 先行研究

Single-cell RNA-seq 解析とは，1 つの細胞に含まれる各遺伝子の発現量を解析する手法である．Grønbech ら [2] は，生の遺伝子発現量データを GMVAE で処理することで，各細胞の潜在的な表現を獲得可能であることを示した．GMVAE には，クラス分類ネットワークが含まれ，データの真のクラスタ数を入力する必要がある．しかし，データにおける真のクラスタ数は不明なため，最適な潜在空間の獲得が困難となる場合がある．

3. 提案手法

本研究では，真のクラスタ数が不明なデータに対して最適なクラスタ数を探索し，GMVAE に用いる手法を提案する．提案手法によるクラスタ数の探索手順を以下に示す．

1. Star clustering を用いてデータのクラスタ数を予測する
2. 予測したクラスタ数を図 1 に示す GMVAE に入力する
3. GMVAE を学習し，データのクラスタ数を予測する
4. GMVAE に入力したクラスタ数と GMVAE で予測したクラスタ数が一致する場合，GMVAE に入力したクラスタ数を最適なクラスタ数とする
5. GMVAE に入力したクラスタ数と GMVAE で予測したクラスタ数が不一致な場合，GMVAE で予測したクラスタ数を GMVAE に入力する．そして，3 から 5 の手順を最適なクラスタ数を予測するまで繰り返す

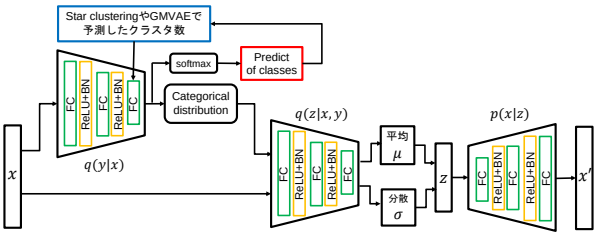


図 1：GMVAE のモデル図

提案手法で探索した最適なクラスタ数による GMVAE を用いることで，データの最適な潜在空間が獲得可能となる．

4. 評価実験

本実験では，提案手法を用いて真のクラスタ数が不明なデータの最適なクラスタ数を探索可能かを評価する．

4.1. 実験概要

本実験は，末梢血単核細胞データ（クラスタ数 9）を対象とする．GMVAE の尤度関数は，負の 2 項分布を使用する．そして，200 エポックの Warm up を実施し，500 エポック学習させる．評価には，Accuracy，Normalization Mutual Information score(NMI) と Entropy を使用する．NMI はクラスタリングの性能を評価し，値は 1 に近いほどクラスタリング性能が高いことを示す．Entropy は予測クラスタに属するデータの真のラベルの偏り度合いを評価し，値が 0 に近いほどクラスタリングができていることを示す．

4.2. 実験結果

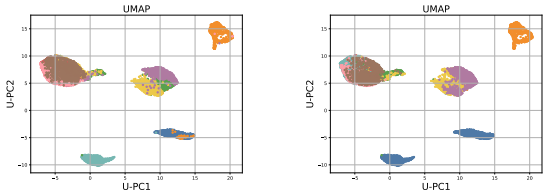
従来手法との比較結果を表 1 に示す．従来手法は真のクラスタ数である 9 を入力とした時の評価結果である．クラスタ数 10 は，提案手法により求めた最適なクラスタ数である 10 を入力とした時の評価結果である．表 1 から，クラスタ数 10 の Accuracy と NMI がクラスタ数 9 よりも高く，クラスタ数 10 の Entropy がクラスタ数 9 よりも低いことが分かる．これらから，提案手法で求めたクラスタ数である 10 が GMVAE で予測する最適なクラスタ数であると考える．

表 1：評価結果

クラスタ数	9（従来手法）	10（提案手法）
Accuracy	0.537	0.661
NMI	0.607	0.683
Entropy	1.150	1.076

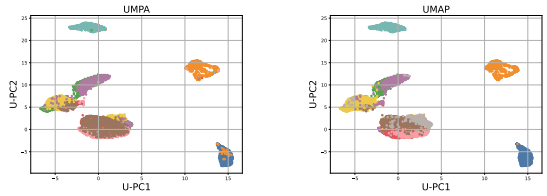
4.3. ラベル分布の比較

従来手法および提案手法の真のラベル分布と予測ラベル分布を図 2，図 3 に示す．提案手法は，青色と水色のクラスタを個々のクラスタと予測したが，従来手法は 1 つの青色のクラスタと予測した．また，図 3 から提案手法は，黄色のクラスタを黄色と灰色の 2 つのクラスタに分離した．これにより，クラスタ数を 1 つ多く予測したと考える．これは，黄色の細胞種が 2 つの細胞種の特徴を持っているためである．これらから，提案手法はデータに即した予測結果であるといえる．



(a) 真のラベル　　(b) 予測ラベル

図 2：従来手法



(a) 真のラベル　　(b) 予測ラベル

図 3：提案手法

5. おわりに

本研究では，GMVAE を用いたクラスタ数の探索手法により，GMVAE のクラスタリング性能が向上し，データに即した予測が可能であることと，予測クラスタ数を真のクラスタ数に近づけることが可能であることを示した．今後は，疾患のあるデータと疾患のないデータを用いて潜在空間を比較することで，疾患原因の細胞種の特定をしていく．

参考文献

[1] N. Dilokthanakul, *et al.*, “Deep unsupervised clustering with gaussian mixture variational autoencoders”, arXiv, 2016.

[2] C. Grønbech, *et al.*, “scVAE: variational auto-encoders for single-cell gene expression data”, Bioinformatics, 2020.