creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

이학석사학위논문

지승영

t-SNE using different distance
(다른 거리를 이용한 t-SNE)

BY

지승영

2022 년 2 월

지승영

t-SNE using different distance
(다른 거리를 이용한 t-SNE)

지도교수 임 요 한

이 논문을 이학석사 학위논문으로 제출함

2021 년 10 월

서울대학교 대학원

통계학과

지 승 영

지승영의 이학석사 학위논문을 인준함

2021 년 12 월

위 원 장	장원철	(인)
부위원장	임요한	(인)
위 원	원중호	(인)

초록

본 연구에서는 차원 축소 방법 중 하나인 t-SNE(t-분포 확률적 임베딩)를 고차원 자료에 적용하여 본다. 이때 고차원 자료의 거리 측정에서 유클리드 거리나 맨하탄 거리가 아닌 Interpoint type distance를 이용하는 방법에 대하여 고민한다. 우선 t-SNE, 거리의 집중(Concentration of Distance), Interpoint type distance에 대하여 살펴본다. 그 후 유클리드 거리와 맨하탄 거리, Interpoint type distance를 이용하여, 실제 고차원 유전자 데이터에 t-SNE를 적용하여 본다. 그리고 유전자 데이터에 대한 사전 스크리닝을 진행한 결과와 함께 비교하여, 그 결과를 의논하여 본다.

주요어: 고차원 데이터, 차원 축소, t-SNE(t-분포 확률적 임베딩), 거리의 집중 (Concentration of Distance), Interpoint type distance, MADD, ADDM

학번: 2020-29533

목차

초록 **i**

1 서론 **1**

2 t-SNE **3**

2.1 t-SNE의 차원 축소 과정 3

2.2 crowding problem 5

2.3 t-SNE의 특징 6

2.4 t-SNE의 약점 6

3 거리들 **8**

3.1 기존 거리 8

3.1.1 유클리드 거리 8

3.1.2 민코프스키 거리(Minkowski Distance) 8

3.1.3 p-Norm 9

3.2 거리의 집중(Concentration of Distances) 9

3.3 Interpoint type distance 11

3.3.1 Mean Absolute Difference of Distances (MADD) 12

3.3.2 Absolute Difference of Distances from the Mean (ADDM) 13

4 고차원 데이터에 활용 **14**

4.1 데이터 설명 14

4.2 t-SNE 결과 15

4.3 평가 척도 20
4.3.1 실루엣 점수(silhouette score) 20
4.3.2 다변량 분산분석(MANOVA) 21
4.4 평가 결과 23
4.4.1 실루엣 점수 23
4.4.2 다변량 분산분석 결과 30
4.4.3 실루엣 점수와 다변량 분산분석 결과 요약 34

5 결론 **35**

참고문헌 **36**

Abstract **38**

그림 목차

2.1 표준정규분포와 자유도가 1인 t분포. 6

4.1 t-SNE 결과 1 : 5%의 변수 이용 16
4.2 t-SNE 결과 2 : 10%의 변수 이용 17
4.3 t-SNE 결과 3 : 30%의 변수 이용 18
4.4 t-SNE 결과 4 : 50%의 변수 이용 19
4.5 t-SNE 결과 5 : 100%의 변수 이용 20
4.6 실루엣 분석 결과 1 : 5%의 변수 이용 26
4.7 실루엣 분석 결과 2 : 10%의 변수 이용 27
4.8 실루엣 분석 결과 3 : 30%의 변수 이용 28
4.9 실루엣 분석 결과 4 : 50%의 변수 이용 29
4.10 실루엣 분석 결과 5 : 100%의 변수 이용 30

표 목차

4.1 5%, 10%, 30%의 설명변수 이용 시 실루엣 점수표 24
4.2 50%, 100%의 설명변수 이용 시 실루엣 점수표 25
4.3 5%, 10%, 30%의 설명변수 이용 시 다변량 분산분석 결과표 32
4.4 50%, 100%의 설명변수 이용 시 다변량 분산분석 결과표 33
4.5 실루엣 점수와 다변량 분산분석 결과 요약표 34

제 1 장 서론

디지털 전환이 이루어 지고, 산업이 발달하면서 많은 양의 데이터가 축적되고 있다. 설명 변수의 개수를 데이터의 차원이라고 할 때, 데이터의 양도 많고, 데이터의 차원도 크다면, 데이터의 의미를 직관적으로 파악하기 힘들다. 또한 고차원 데이터의 경우, 필요 없는 설명 변수를 가지고 있는 경우도 많고, 계산을 하기에도 많은 시간이 소요된다. 고차원 데이터의 이러한 문제들을 해결하기 위한 하나의 방법은, 데이터의 차원을 축소하는 것이다.

차원 축소(Dimension reduction)란 고차원 공간의 데이터를 저차원 공간의 데이터로 변환 시키는 것이다. 이 때, 고차원 공간의 데이터로부터 정보의 손실을 줄이고, 의미 있는 성질들을 남기는 것을 목표로 하여, 저차원 공간의 데이터로 고차원 공간의 데이터를 변환 시킨다.

대표적으로는 선형 데이터 축소 기법인 주성분분석(Principal Component Analysis)가 있다. 여기서 선형의 의미는, 축소된 차원의 데이터가 기존 고차원 데이터의 선형 함수로 표현된다는 것이다. 이와 같은 선형 차원 축소 방법은, 고차원 데이터의 구조를 제대로 보존하지 못하는 경우가 많다. 해당 연구에서는 비선형 방식의 차원 축소 방법인 t-SNE(t-분포 확률적 임베딩)에 대하여 살펴볼 것이다.

비선형 방식 차원 축소는 저차원으로 축소된 데이터가 고차원 데이터의 비선형 함수로 표현되는 것을 말한다. t-SNE를 비롯하여, Sammon mapping, Isomap, Locally Linear Embedding 등 비선형적 방식의 차원 축소 방법들은 고차원 공간 데이터들의 거리 관계를 저차원 공간에서 보존되도록 차원을 축소한다. 따라서, 비선형 방식의 차원 축소 방식에서, 고차원 공간의 거리가 서로를 구분 지을 수 있는 의미 있는 거리가 되는 것이 중요하다.

하지만 차원의 저주 문제 중 한 가지로, 고차원 데이터의 경우 차원이 커질수록 임의의 두 점 사이의 p-Norm 거리가 비슷해지는 거리의 집중(Concentration of Distance)현상이 나타난다. 따라서 고차원 공간에서는 기존의 거리의 의미가 옅어지게 된다.

현재까지 고차원 공간의 거리에 대한 많은 연구가 있었으나, 이들의 비선형 방식의 차원 축소에의 적용은 큰 주목을 받지 못하였다. 본 연구에서는 최근 제안된 고차원 공간의 거리들을 활용한 t-SNE에 대하여 고민하여 본다. 또한 고차원 설명 변수들에 대한 사전 스크리닝 절차의 영향에 대하여 고민하여 본다.

제 2 장 t-SNE

t-SNE(t-분포 확률적 임베딩)는 고차원 데이터를 저차원으로 차원 축소하는 비선형 차원 축소 방법이다([3]). 이는 고차원 데이터들 사이의 쌍별 거리를 확률화하고, 이를 이용해 저차원 공간으로 임베딩 된 데이터들의 쌍별 거리 관계가 고차원에서와 같이 유지되도록 차원을 축소한다.

t-SNE는 고차원 데이터를 2차원 또는 3차원으로 줄여, 데이터를 시각화하는 것에 많이 이용된다. 다음에 언급될 내용들은 t-SNE의 차원 축소 과정과 차원 축소 시의 문제점인 crowding problem, t-SNE의 특징과 약점이다. 해당 내용들은 [3]을 참고하였다.

2.1 t-SNE의 차원 축소 과정

t-SNE는 고차원 공간의 유사성(similarity)를 계산하기 위한 분포로 정규분포(Gaussian distribution)를 사용하고, 저차원 공간의 유사성(similarity)을 계산하기 위한 분포로 t-분포를 사용하는 확률적 임베딩 방법이다.

고차원 공간의 데이터 점(datapoint)을 x_i 라 하고, 이에 대응되는 저차원으로 차원 축소 된 점(map point)을 y_i 라고 한다. t-SNE의 첫 단계는 다음과 같이 고차원 공간의 데이터들 간의 유클리드 거리를 유사성(similarity)을 나타내는 조건부 확률 $p_{j|i}$ 로 변환하는 것이다. (σ_i의 값을 정하는 방식은 펄플렉서티(perplexity)를 이용한다. 해당 내용은 생략한다.)

$$p_{j|i} = \frac{\exp\left(-\left\|x_i - x_j\right\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\left\|x_i - x_k\right\|^2 / 2\sigma_i^2\right)}$$

이는 고차원 공간의 데이터 x_i가 주어졌을 때, 다른 고차원 공간의 데이터 x_j

와의 거리를 정규분포를 이용해 확률화 한 것이다. 그리고 이를 $p_{ij} = p_{ji}$ 와 같이 대칭적(symetric)이고, $\sum_j p_{ij} > \frac{1}{2n}$($n$은 데이터의 개수)와 같은 성질을 가지는 쌍별 유사성(pairwise similarity)으로 이용하기 위하여

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2} \tag{2.1}$$

를 이용한다.

그리고 두 번째 과정으로는 다음을 생각해 볼 수 있다. 저차원 공간으로 축소된 점(map point) y_i를 이용하여, 고차원 공간의 쌍별 유사성(pairwise similarity) p_{ij}에 대응되는 저차원 공간의 쌍별 유사성(pairwise similarity) q_{ij}를 아래와 같이 만든다. 이 때는 자유도가 1인 t-분포를 이용한다.

$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \|y_k - y_l\|^2\right)^{-1}} \tag{2.2}$$

그 후 위의 (2.1), (2.2)의 p_{ij}, q_{ij}를 이용한 다음 (2.3)과 같은 비용 함수 C를 최소화한다. 이를 통해 고차원 공간에서의 데이터 점들 사이의 관계가 저차원으로 매핑(mapping) 된 점들 사이의 관계에서도 유지되도록 차원을 축소한다.

$$C = KL(P\|Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{2.3}$$

여기서 $p_{ii} = q_{ii} = 0$라고 설정한다. 그리고 C의 그레디언트는 다음과 같이 주어진다.

$$\frac{\delta C}{\delta y_i} = 4 \sum_j \left(p_{ij} - q_{ij}\right)\left(y_i - y_j\right)\left(1 + \|y_i - y_j\|^2\right)^{-1}$$

해당 내용들과는 논외로 위의 (2.1)에서 설명된 p_{ij}를 쓰는 이유는 다음과 같다. n이 데이터의 개수일 때, p_{ij}는 $\sum_j p_{ij} > \frac{1}{2n}$과 같은 성질을 만족시킨다. 그리고 이는 모든 고차원 공간의 점 x_i가 비용 함수에 기여할 수 있도록 해준다.

만약 정규분포를 이용한 다음과 같은

$$p_{ij} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma^2\right)}{\sum_{k \neq l} \exp\left(-\|x_k - x_l\|^2 / 2\sigma^2\right)},$$

를 이용한다면, 고차원 공간의 점 x_i가 이상치(outlier) 일 경우에, p_{ij}의 값이 매우 작아지게 된다. 그리고 이는 비용 함수에 영향을 미치지 못하게 된다.

정리하자면 t-SNE는 정규분포를 이용하여 거리를 확률화한 (2.1)의 p_{ij}와 자유도가 1인 t분포를 이용하여 거리를 확률화한 (2.2)의 q_{ij}와 비용함수인 (2.3)을 통해서, 고차원에서의 데이터 구조를 유지시키며 저차원으로 차원을 축소시키는 기법이라고 할 수 있다. 그리고 t-분포를 저차원 공간의 쌍별 유사성(pairwise similarity)을 계산하기 위한 분포로 사용했기 때문에, 이를 t-SNE(t-분포 확률적 임베딩)라고 부른다. 후에 언급될 crowding problem을 완화시키기 위하여, 두터운 꼬리(heavy tail) 분포인 t-분포를 이용하였다.

2.2 crowding problem

Crowding problem 이란 고차원의 데이터 점들(datapoints)을 2차원 공간으로 매핑(mapping) 시킬 때, 고차원에서 적당한 거리가 떨어진 점들을 2차원 공간이 수용하기에는, 2차원 공간의 면적이 충분하지 않다는 문제이다. 따라서 2차원 공간으로 축소된 점(mapping point)들의 경우, 고차원 공간에서보다 멀게 놓이게 된다.

t-SNE는 저차원 공간의 거리를 확률화하는 분포를 두터운 꼬리 분포인 자유도가 1인 t-분포를 이용하여, 이 문제를 완화한다. 그림 2.1에서 확인할 수 있듯이 t-분포는 정규분포와 비교하여 같은 확률 값의 차이(y축 값의 차이)에 대하여, 더 넓은 거리의 차이(x축 값의 차이)를 만들어 저차원 공간으로 점을 매핑(mapping)하게 된다. 따라서 고차원 공간의 데이터 점(datapoint) x_i 로부터 적당한 거리에 떨어진 대부분의 점들은, 2차원 맵(map)에서 더 멀리 떨어지게 될 것이다.

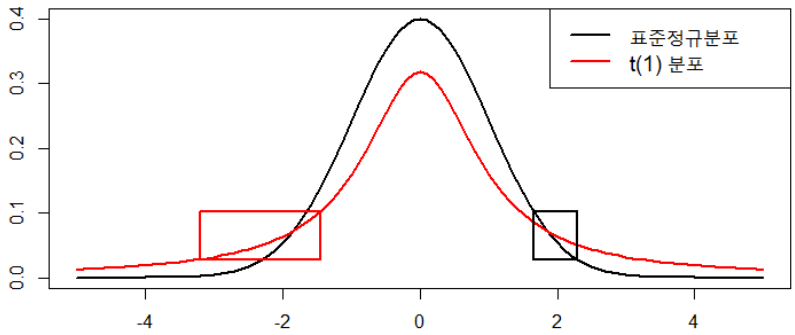

그림 2.1 표준정규분포와 자유도가 1인 t분포.

2.3 t-SNE의 특징

t-SNE는 첫째로, 비슷하지 않은 고차원 데이터 점들(dissimilar datapoints)을 큰 쌍별 거리들(large pairwise distances)로 모델링한다. 둘째로, 비슷한 고차원 데이터 점들(similar datapoints)을 작은 쌍별 거리들(small pairwise distances)로 모델링한다. 셋째로, 저차원 맵(map)에 긴 범위의 힘을 작용시켜서, 최적화의 초기에 분리된 비슷한 점들(similar points)이 서로 끌어당겨질 수 있도록 한다.

2.4 t-SNE의 약점

[3]에 언급된 t-SNE의 약점으로는 다음과 같은 것들이 있다.

(1) 4차원 이상으로의 축소

고차원 데이터를 d-차원 데이터로 축소한다고 하자. 4차원 이상($d \geq 4$) 으로의 차원 축소 시에 어떻게 작동할지 확실하지 않다. 고차원 공간에서 두터운 꼬리 분포(heavy tail distribution)인 t-분포를 이용할 시, 두터운 꼬리(heavy tail) 부분이 상대적으로 확률 질량(probability mass)의 많은 부분을 차지한다. 그래서 이 때문에, d-차원 표현(d-dimensional representation)에서 데이터의 국소적 구조(local structure)를 잘 보존하지 못할 수 있다.

(2) 내재적 차원의 저주

t-SNE는 주로 데이터의 국소적 성질들(local properties of the data)에 기반하여, 데이터의 차원을 축소한다. 따라서 고차원인 경우, 유클리드 거리를 이용할 수 있는 가정인 국소적 선형 가정(local linearity assumption)이 깨질 수 있게 된다. 이에 따라 t-SNE의 성능이 떨어질 수 있다. 이 문제를 해결할 수 있는 한 가지 방법은 고변동의 다양체(highly varing manifold)를 여러 개의 비선형 층으로 효율적으로 표현할 수 있는 오토 엔코더(autoencoder)와 같은 모델의 데이터 표현을 이용해 t-SNE를 수행하는 것이다.

제 3 장 거리들

집합 X에서의 거리는 다음 성질들을 만족시키는 함수 $f : X \times X \to \mathbb{R}$로 정의된다. 임의의 $x, y, z \in X$에 대하여,

(1) $\forall x, y \in X,\ f(x, y) \geq 0 \quad (\ f(x, y) = 0 \ \Leftrightarrow\ x = y\ )$

(2) $\forall x, y \in X,\ f(x, y) = f(y, x)$ (*대칭성* , *symmetry*)

(3) $\forall x, y, z \in X,\ f(x, y) + f(y, z) \geq f(x, z)$ (*삼각 부등식* , *triangle inequality*)

일반적으로 거리함수는 d로 나타내지만, 본 연구에서는 d를 데이터의 차원을 나타내는 것에 쓰기로 한다.

3.1 기존 거리

3.1.1 유클리드 거리

유클리드 거리는 일반적으로 사용하는 거리이다. d-차원 유클리드 거리 $f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$는 다음과 같이 정의된다.

$$f(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_d - y_d)^2}$$

$\forall x, y \in \mathbb{R}^d$, 여기서 $x = (x_1, x_2, \cdots, x_d),\ y = (y_1, y_2, \cdots, y_d)$ 이다.

3.1.2 민코프스키 거리(Minkowski Distance)

d-차원 민코프스키 거리 $f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$는 다음과 같이 정의된다.

$$f(x, y) = \left(\sum_{i=1}^{d} |x_i - y_i|^p \right)^{\frac{1}{p}}.$$

$\forall x, y \in \mathbb{R}^d$, 여기서 $x = (x_1, x_2, \cdots, x_d)$, $y = (y_1, y_2, \cdots, y_d)$ 이다.

$p \geq 1$ 인경우 민코프스키 거리는 거리의 정의를 만족한다. 하지만 $0 < p < 1$ 경우 삼각 부등식을 만족시키지 못해, 거리의 정의를 만족하지는 못한다. 하지만 본 연구에서는 거리로 혼용해서 이용하도록 한다. 민코프스키 거리는 $p = 1$ 인 경우 맨하탄 거리가 되고, $p = 2$ 인 경우 유클리드 거리가 된다.

3.1.3 p-Norm

p-Norm $\|\cdot\|_p$는 다음과 같이 정의된다.

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{d} |x_i|^p \right)^{1/p}$$

여기서 $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ 이다.

p-Norm이 위의 유클리드 거리와 민코프스키 거리를 유도하므로, 앞으로 p-Norm과 이로 유도되는 위의 거리들을 혼용해서 사용하도록 한다. 또한 $0 < p < 1$ 경우 이는 엄밀히 Norm이 아니지만, 편의상 Norm으로 언급하도록 한다. 그리고 이로 유도되는 거리 또한 정의를 만족하는 거리(Distance)가 아니지만, 거리로 언급하도록 한다.

3.2 거리의 집중(Concentration of Distances)

거리의 집중(Concentration of Distances)이란 고차원에서 두 점 사이의 거리가 대부분 비슷해져서, 거리 함수가 점들을 잘 구분하지 못하는 것을 말한다. [6]에서는 거리의 집중 현상이 생기는 조건에 대해서 언급하며, 조건을 만족하는 여러 가지 예시를 들고 있다. [6]에서의 거리의 집중과 관련된 정리를 [7]에서 다음과 같이 나타내고 있다.

정리 1: Beyer et al

$\mathrm{X}^{(j)} : 1 \leq j \leq n$ *를 d-차원 i.i.d. 랜덤 벡터들이라고 하자. 만약*

$$\lim_{d\to\infty} \mathrm{Var}\left(\frac{\|\mathrm{X}^{(j)}\|}{\mathrm{E}\left(\|\mathrm{X}^{(j)}\|\right)}\right) = 0$$

이라면, 임의의 $\epsilon > 0$ 에 대하여

$$\lim_{d\to\infty} \mathbf{P}\left[\frac{\max_j \|\mathrm{X}^{(j)}\| - \min_j \|\mathrm{X}^{(j)}\|}{\min_j \|\mathrm{X}^{(j)}\|} \leq \epsilon\right] = 1$$

이다.

여기서, $\|\cdot\|$ *는 임의의 양함수이다.*

즉, 위 정리의 가정 하에서는 차원이 커지면 임의의 두 점 사이의 거리가 모두 비슷해진다. 그렇기 때문에, 거리의 판별적인 의미가 사라지게 된다. 또한 [8]에서는 위 정리의 역 또한 샘플의 크기(sample size)가 충분히 크다면 성립한다는 것을 보였다. 또한 [7]에서는 정리 2와 Proposition 1을 보였다.

정리 2: D.Francois et al

$\mathrm{X} = (X_1, \cdots, X_d)$ *가 각 성분이* $X_i \sim \mathcal{F}$ *로 i.i.d. 성분들을 가지는 랜덤 벡터라고 하자. 그러면 다음이 성립한다.*

$$\lim_{d\to\infty} \frac{\sqrt{\mathrm{Var}\left(\|X\|_p\right)}}{\mathrm{E}\left(\|X\|_p\right)} = 0$$

여기서, $p > 0$ *이다.*

Propositon 1: D.Francois et al

만약 데이터들이 동일한 분포가 아니더라도(not identically distributed), 데이터

들이 정규화(normalization)되었다면, 정리 2의 결론은 여전히 성립한다.

위 정리들에 따르면, d차원 랜덤 벡터의 각 성분이 독립이고 같은 분포를 따를 때, p-Norm을 이용하게 되면 거리의 집중(Concentration of Distance) 현상이 발생한다. 또한 Proposition 1에 따르면 정리 2의 동일 분포 가정은 데이터가 정규화된다는 것으로 대체될 수 있다.

또한 [7]에서는 고차원 데이터의 경우, 변수들 사이에 큰 상관관계나 의존성(dependency)이 있는 경우의 거리의 집중(Concentration of Distance) 정도가, 모든 변수가 독립(independent)인 경우보다 작다는 것을 보였다. 이 결론으로부터 거리의 집중 현상은 데이터가 임베딩(embedding) 된 차원보다는 데이터의 내재적 차원(intrinsic dimension)에 더 의존한다는 것을 생각해 볼 수 있다.

3.3 Interpoint type distance

Interpoint type distance는 두 점 사이의 거리를 구할 때, 다른 점들을 함께 이용하여 두 점 사이의 거리를 구하는 방식이다. [5]에서는 Interpoint type distance를 이용한 분류기(classifier)가 고차원 자료에서 거리의 집중(Concentration of Distance)이 발생하는 경우, 유클리드 거리를 이용한 분류기 보다 잘 작동한다는 것을 보였다. 또한 적절한 가정 하에서, 제안된 분류기(classifier)가 HDLSS(High Dimension Low Sample Size)의 점근적 체제(asymptotic regime)에서 작동하는 경우(샘플 사이즈가 고정되어 있고, 차원이 무한대로 갈 때), 제안된 분류기의 오분류 확률의 최적성을 보였다.

본 연구에서는 Interpoint type distance를 t-SNE의 고차원 자료의 거리를 측정하는 것에 사용해보고, 그 성능을 살펴보고자 한다.

3.3.1 Mean Absolute Difference of Distances (MADD)

[5]에서는 고차원에서의 거리의 집중(Concentration of Distance) 문제를 해결하기 위해, 다음과 같은 거리를 제시하였다. 이 거리는 다음과 같은 사실에 기반한다. 2개의 클래스(class)를 구분하는 문제를 생각한다. 분포를 평균이 같고, 등방성공분산을 가지는 다변량 정규분포라고 가정한다.

즉, class-1,2 는 평균 벡터가 $\mu_{1,d} = (\mu_1, \mu_1, \cdots, \mu_1)^T$ and $\mu_{2,d} = (\mu_2, \mu_2, \cdots, \mu_2)^T$ 이고, 공분산행렬은$\mathbf{\Sigma}_{1,d} = \sigma_1^2 \mathbf{I}_d$ and $\mathbf{\Sigma}_{2,d} = \sigma_2^2 \mathbf{I}_d$ 인 다변량 정규분포이다.

$\mathbf{X} = (X_1, \ldots, X_d)^T$ and $\mathbf{X}^* = (X_1^*, \ldots, , X_d^*)^T$ 가 class-1 로부터의 독립 관측치들(independent observations) 이고, $d^{-1}\|\mathbf{X} - \mathbf{X}^*\|_p^p = d^{-1}\sum_{i=1}^d |X_i - X_i^*|^p$는 d개의 독립이고 동일한(independent and identical) 분포의 확률 변수의 평균이다. 따라서 p차 적률(moment)이 유한(finite)하다면 강대수법칙(Strong Law of Large Numbers)에 의해,

$$d^{-1}\|\mathbf{X} - \mathbf{X}^*\|_p^p = d^{-1}\sum_{i=1}^{d} |X_i - X_i^*|^p \overset{a.s}{\to} E|X_1 - X_1^*|^p \text{ as } d \to \infty \tag{3.1}$$

이다. 마찬가지로 $\mathbf{Y} = (Y_1, \ldots, Y_d)^T$ and $\mathbf{Y}^* = (Y_1^*, \ldots, , Y_d^*)^T$ 가 class-2 로부터의 독립 관측치들(independent observations) 이라고 하자. 그렇다면

$$d^{-1}\|\mathbf{Y} - \mathbf{Y}^*\|_p^p \overset{a.s}{\to} E|Y_1 - Y_1^*|^p \text{ as } d \to \infty \tag{3.2}$$

이다. 마지막으로 만약 $\mathbf{X}$ 가 class-1의 확률 변수이고, $\mathbf{Y}$ 가 class-2의 확률 변수라고 한다면,

$$d^{-1}\|\mathbf{X} - \mathbf{Y}\|_p^p \overset{a.s}{\to} E|X_1 - Y_1|^p \text{ as } d \to \infty \tag{3.3}$$

이다. 따라서 다음과 같은 거리 ρ_p^*를 생각하자. $\mathcal{X}_1, \mathcal{X}_2$는 각각 class-1, class-2의 표본집단이다. 그리고 $\mathbf{z}$ 는 class-1의 미래의 관측치이고 $\mathbf{w}$는 표본집단의 원소라고

가정하자.

$$\rho_p^*(\mathbf{z},\mathbf{w}) = |\mathcal{X}_1 \cup \mathcal{X}_2 - \mathbf{w}|^{-1} \sum_{t \in \{\mathcal{X}_1 \cup \mathcal{X}_2 - \mathbf{w}\}} |\|\mathbf{z}-\mathbf{t}\|_p - \|\mathbf{w}-\mathbf{t}\|_p|$$

위의 (3.1),(3.2),(3.3)에 의해 p차 적률(moment)이 유한(finite)하다면, 다음이 성립한다.

$$\rho_p^*(\mathbf{z},\mathbf{w}) \overset{a.s}{\to} 0 \Leftrightarrow \mathbf{z} \text{ 와 } \mathbf{w} \text{ 는 같은 클래스(class)이다.}$$

위의 거리 ρ_p^*를 수정한 ρ를 다음과 같이 정의하고, 이를 Mean Absolute Difference of Distances (MADD)라고 한다. 여기서 d 는 데이터의 차원이다.

$$\rho(\mathbf{z},\mathbf{w}) = \frac{1}{|\mathcal{X}_1 \cup \mathcal{X}_2| - 2} \frac{1}{\sqrt{d}} \sum_{t \in \{\mathcal{X}_1 \cup \mathcal{X}_2 - \{\mathbf{w},\mathbf{z}\}\}} |\|\mathbf{z}-\mathbf{t}\|_2 - \|\mathbf{w}-\mathbf{t}\|_2|$$

3.3.2 Absolute Difference of Distances from the Mean (ADDM)

[9]에서는 위의 MADD를 조금 수정한 다음과 같은 거리 δ를 제시한다. $\bar{\mathbf{t}}$는 $\mathbf{z},\mathbf{w}$를 제외한 데이터들의 평균이다. 또한 d 는 데이터의 차원이다.

$$\delta(\mathbf{z},\mathbf{w}) = \frac{1}{\sqrt{d}} |\|\mathbf{z}-\bar{\mathbf{t}}\|_2 - \|\mathbf{w}-\bar{\mathbf{t}}\|_2|$$

여기서, $\bar{\mathbf{t}} = \frac{1}{|\mathcal{X}_1 \cup \mathcal{X}_2| - 2} \sum_{t \in \{\mathcal{X}_1 \cup \mathcal{X}_2 - \{\mathbf{w},\mathbf{z}\}\}} \mathbf{t}$ 이다.

제 4 장 고차원 데이터에 활용

5개의 라벨이 있는 5746차원의 고차원 유전자 데이터 67개에 대하여 t-SNE를 적용해보고, 저 차원에서 각 라벨들의 밀집된 정도를 비교해 볼 것이다. t-SNE의 고차원 공간의 거리를 계산할 때, 기존에 사용하던 유클리드 거리, 맨하탄 거리를 비롯해 Interpoint type의 거리인 MADD와 ADDM을 이용할 것이다.

또한 t-SNE를 적용하기에 앞서 사전 스크리닝을 진행할 것이다. 5개의 라벨이 각 설명변수에서 평균이 같은지에 관해 분산분석(ANOVA)을 적용하여 볼 것이다. 그 후 5746개의 p-value 중 작은 p-value를 가지는 설명변수들을 사용하여, t-SNE를 적용해 볼 것이다. 그리고 실루엣 점수와 다변량 분산분석(MANOVA)을 이용해, t-SNE의 결과를 평가할 것이다.

4.1 데이터 설명

t-SNE를 적용할 데이터는 난소 장액성 선암(protein phosphorylation analysis)의 단백질 인산화 데이터이다. 해당 암은 5개의 종류로 구분되며, 각각 A(Differentiated), B(Immunoreactive), C(Proliferative), D(Mesenchymal), E(Stromal)의 5개 라벨로 표기하였다. 각 데이터는 한 명의 환자를 의미한다. 하나의 데이터는 A, B, C, D, E 중 하나의 암 종류와 5746곳의 위치에 대하여 단백질 인산화 정도를 수치로 나타내는 5746차원의 벡터로 이루어져 있다. 전체 관측치의 개수는 67개이다. 5가지 종류의 암의 관측치의 개수는 $(n_A, n_B, n_C, n_D, n_E) = (21, 10, 7, 21, 8)$이다.

4.2 t-SNE 결과

t-SNE를 이용하여 고차원 유전자 데이터를 2차원으로 차원 축소하였다. R 버전 4.1.1에서 Rtsne 패키지 0.15 버전의 Rtsne 함수를 사용하였다. Rtsne 함수의 입력값으로는 각 거리 종류에 따른 거리 행렬과, perplexity를 이용하였고, 나머지는 디폴트 값을 이용하였다. 거리 행렬의 경우, 데이터를 표준화(Standardization)한 경우와 표준화하지 않은 경우 모두 계산하였다. Perplexity는 5, 10, 15, 20 값을 이용하였다.

5746개의 각 변수에 대하여, 분산분석 결과 p-value의 값이 작은 5%, 10%, 30%, 50%, 100% 인 5가지 경우로 나누어서 나타내었다. 즉 각 변수의 평균이 같다는 귀무가설을 강하게 기각하는 변수 순서로, 변수를 선택하였다.

모든 경우에서 전반적으로 차원 축소 시 군집화가 잘되지 않았다. P-value 값이 작은 5%, 10%, 30%의 변수만을 이용한 경우, p-value 값이 작은 50%, 100% 변수를 이용한 경우보다는, 군집화가 잘 되었다. 또한 전체적으로 유클리드 거리와 맨하탄 거리가 MADD(Mean Absolute Difference of Distance)와 ADDM(Absolute Difference of Distances from the Mean) 보다 군집화를 잘하였다. 표준화를 한 경우와 그렇지 않은 경우의 차이는 미미하였지만 표준화를 한 경우가 군집화가 조금 더 잘 되었다. Perplexity의 경우 전반적으로 15인 경우에 가장 군집화가 잘 되었다.

다음의 그림은 5%, 10%, 30%, 50%, 100% p-value 그룹별로 t-SNE를 적용한 결과이다. 공간의 제약상 표준화된 변수를 사용한 경우와 perplexity는 15를 이용한 경우만을 나타낸다. 데이터의 5개의 라벨들은 A, B, C, D, E로 표기하였다. 왼쪽 위는 유클리드 거리, 오른쪽 위는 맨하탄 거리, 왼쪽 아래는 MADD, 오른쪽 아래는 ADDM을 고차원 공간의 거리로 이용한 t-SNE 결과이다.

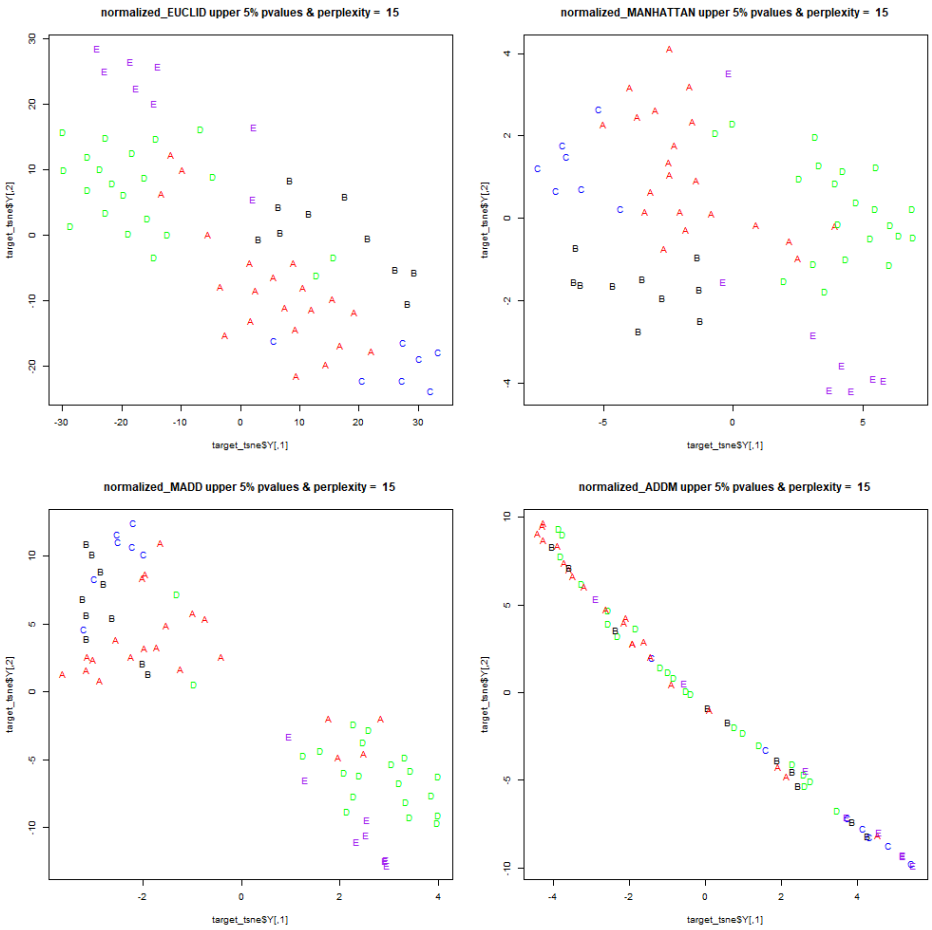

그림 4.1 t-SNE 결과 1 : 5%의 변수 이용

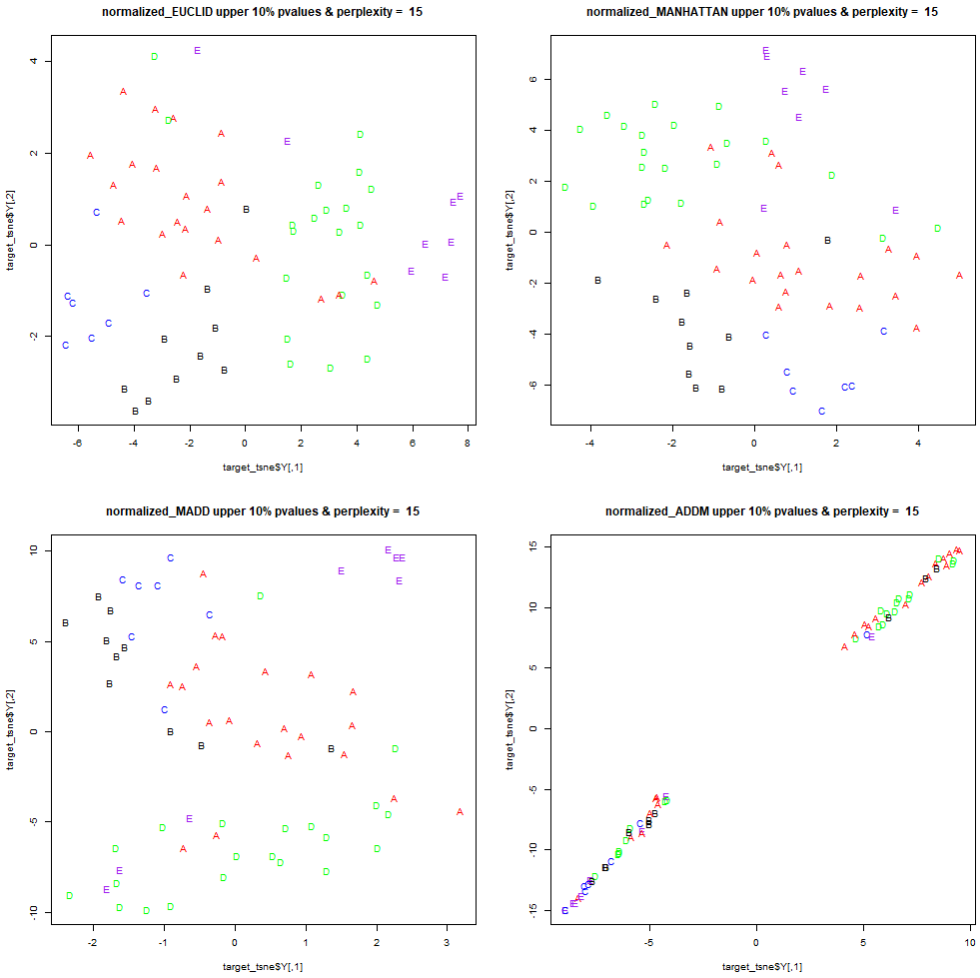

그림 4.2 t-SNE 결과 2 : 10%의 변수 이용

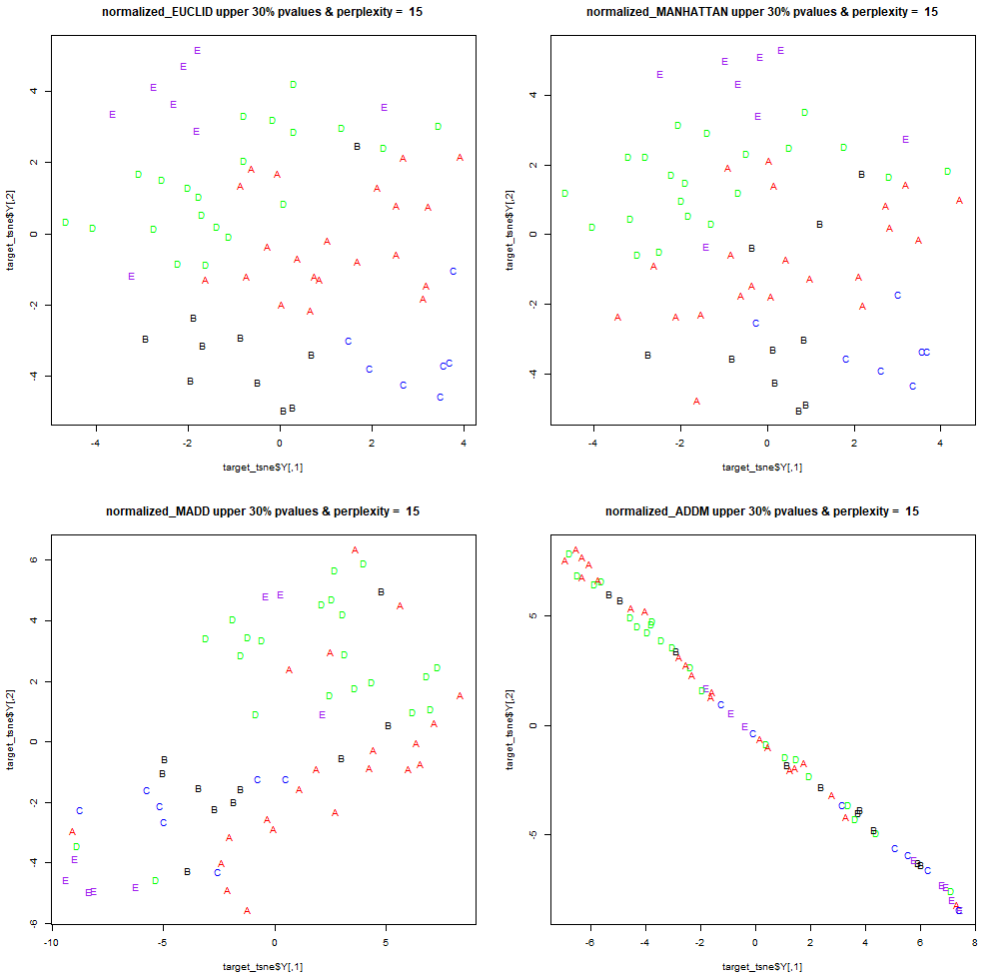

그림 4.3 t-SNE 결과 3 : 30%의 변수 이용

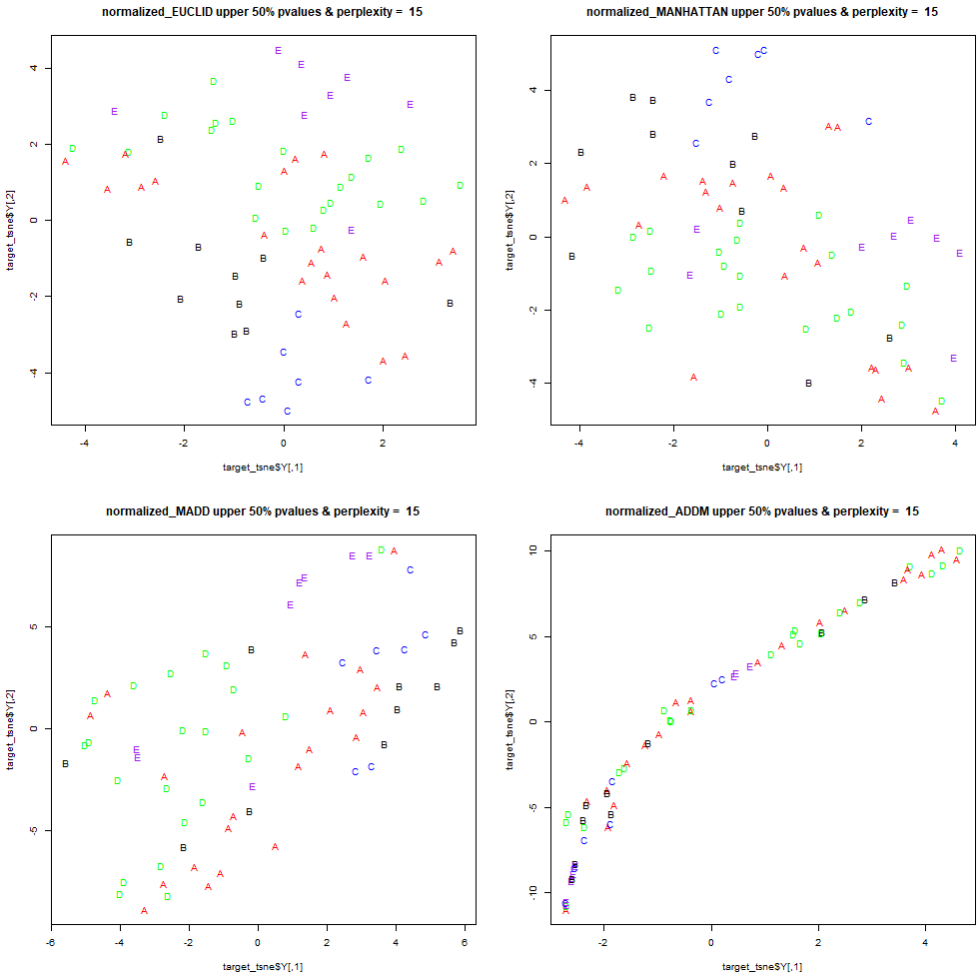

그림 4.4 t-SNE 결과 4 : 50%의 변수 이용

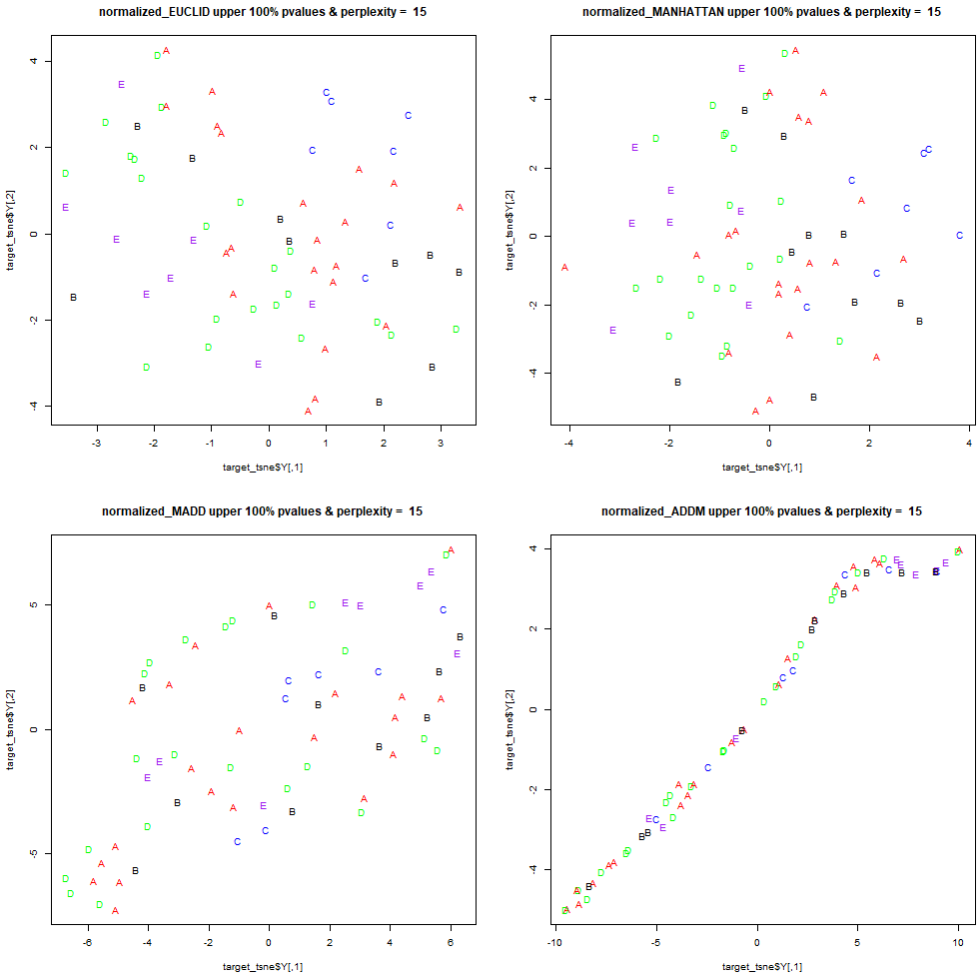

그림 4.5 t-SNE 결과 5 : 100%의 변수 이용

4.3 평가 척도

4.3.1 실루엣 점수(silhouette score)

실루엣 점수는 군집화(clustering) 알고리즘의 결과인 군집들이 얼마나 잘 군집화 되었는지를 수치적으로 표현한 방법이다. 이는 각 군집이 다른 군집에 비하여

얼마나 응집되어 있는지를 표현한다. t-SNE는 군집화 알고리즘은 아니지만, 2차원으로 차원이 축소된 데이터가, 얼마나 각 라벨을 분리시켰는지를 측정하기 위해 실루엣 점수를 이용한다.

군집을 C_I와 같이 나타내고, 각 데이터는 $y_i \in C_I$ 라고 하자. y_i의 실루엣 점수 s_i의 정의는 다음의 a_i, b_i를 이용하여 정의된다.

$$a_i = \frac{1}{|C_I| - 1} \sum_{y_j \in C_I, y_j \neq y_i} \|y_i - y_j\|_2 \tag{4.1}$$

$$b_i = \min_{J \neq I} \frac{1}{|C_J|} \sum_{y_j \in C_J} \|y_i - y_j\|_2 \tag{4.2}$$

여기서 유클리드 거리가 아닌 다른 거리를 이용할 수 있지만, 본 연구에서는 유클리드 거리를 이용하기로 한다. a_i는 군집 C_I와 y_i가 얼마나 다른지를 평균 거리로 표현한다. b_i 는 데이터 y_i가 속한 군집이 아닌 다른 군집에 대하여, 해당 군집의 모든 데이터들과 y_i의 평균 거리를 구한 후, 그중 최소인 평균 거리를 구한 것이다. 이때 선택된 최소 거리 군집은 C_I가 아닌 군집 중 y_i가 가장 속할 법한 군집으로 생각할 수 있다.

실루엣 점수 s_i는 $|C_I| > 1$인 경우 다음과 같이 정의된다.

$$s_i = \frac{b_i - a_i}{max\{a_i, b_i\}} \tag{4.3}$$

그리고 $|C_I| = 1$인 경우 $s_i = 0$으로 정의된다. 해당 정의에 따라 $-1 \leq s_i \leq 1$ 이 성립한다. 여기서 a_i 보다 b_i가 커서, s_i가 1에 가까울수록, 군집화가 잘 되었다고 평가할 수 있다.

4.3.2 다변량 분산분석(MANOVA)

다변량 분산분석(MANOVA)은 분산분석(ANOVA)의 다변량으로의 확장이다. 다변량 분산분석은 여러 그룹의 2개 이상의 변수의 평균이 같은지에 대한 검정을 시행한다. 다변량 분산분석의 가정과 모델, 가설, Table은 다음과 같다.

가정

g개의 그룹이 있을 때 다변량 분산분석을 위한 기본 가정은 다음과 같다.
$X_{l1}, X_{l2}, \cdots, X_{ln_l} \sim N_p(\mu_l, \Sigma)$, $l = 1, 2, \cdots, g$, n_l 은 l번째 그룹의 표본의 크기, N_p는 p 차원 정규분포, g는 그룹의 개수이다. 즉 그룹별로 평균이 다르고, 공분산 행렬이 같은 다변량 정규분포 가정이 필요하다.

모델

다변량 분산분석의 모델은 다음과 같다.
$\boldsymbol{X_{li} = \mu + \tau_l + \epsilon_{li}}$,
여기서 $\boldsymbol{\Sigma_{l=1}^{g} n_l\ \tau_l = 0}$ and $\boldsymbol{\epsilon_{li} \sim N_p(0, \Sigma)}$

가설

다변량 분산분석의 가설은 다음과 같다.
$\boldsymbol{H_0}$: $\boldsymbol{\tau_1 = \tau_2 = \cdots = \tau_g = 0}$ vs. $\boldsymbol{H_1 : not\ H_0}$

다변량 분산분석 Table

요인	제곱합	자유도
집단 간	$\boldsymbol{B = \sum_{l=1}^{g} n_l \left(\overline{\mathbf{X}}_{l\cdot} - \overline{\mathbf{X}}_{\cdot\cdot}\right)\left(\overline{\mathbf{X}}_{l\cdot} - \overline{\mathbf{X}}_{\cdot\cdot}\right)^\top}$	$\boldsymbol{g-1}$
집단 내	$\boldsymbol{W = \sum_{l=1}^{g} \sum_{i=1}^{n_l} \left(\mathbf{X}_{li} - \overline{\mathbf{X}}_{l\cdot}\right)\left(\mathbf{X}_{li} - \overline{\mathbf{X}}_{l\cdot}\right)^\top}$	$\boldsymbol{\sum_{l=1}^{g} n_l - g}$
전체	$\boldsymbol{\mathbf{T} = \sum_{l=1}^{g} \sum_{i=1}^{n_l} \left(\mathbf{X}_{li} - \overline{\mathbf{X}}_{\cdot\cdot}\right)\left(\mathbf{X}_{li} - \overline{\mathbf{X}}_{\cdot\cdot\cdot}\right)^\top}$	$\boldsymbol{\sum_{l=1}^{g} n_l - 1}$

위의 Table의 $\boldsymbol{B}$ 가 $\boldsymbol{W}$에 비하여 큰 경우 귀무가설을 기각하게 된다. 이때 이용되는 통계량들은 다음과 같다.

통계량

우선, $\boldsymbol{W^{-1}B}$의 고유값들을 $\boldsymbol{\lambda_j}$, $\boldsymbol{j = 1, 2, \cdots, p}$ 라고 하자.

- Wilk's Lambda

  $\boldsymbol{\Lambda} = \dfrac{|\boldsymbol{W}|}{|\boldsymbol{B}+\boldsymbol{W}|}$ 로 정의되고 다음과 같은 방법으로도 계산할 수 있다.

  $\boldsymbol{\Lambda} = \prod_{j=1}^{p} \dfrac{1}{1+\lambda_j}$, 여기서 $\boldsymbol{\Lambda}$ 가 작을 때, 귀무가설 $\boldsymbol{H_0}$를 기각한다. 해당 실제 데이터의 경우 $p = 2, g \geq 2$이므로 $\boldsymbol{\Lambda} \sim \boldsymbol{F_{2(g-1),2(\sum n_l - g - 1)}}$ 을 따른다.

- Pillai's Trace

  $\boldsymbol{T}_{\text{Pil}} = \boldsymbol{Tr}(\boldsymbol{B}(\boldsymbol{B}+\boldsymbol{W})^{-1}) = \sum_{j=1}^{p} \dfrac{\lambda_j}{1+\lambda_j}$, 여기서 $\boldsymbol{T}_{\text{Pil}}$이 클 때, 귀무가설 $\boldsymbol{H_0}$를 기각한다.

4.4 평가 결과

4.4.1 실루엣 점수

t-SNE 결과와 마찬가지로, 공간의 제약상 표준화된 변수를 사용한 경우와 perplexity는 15를 이용한 경우만을 나타낸다. 데이터의 5개의 라벨들은 A, B, C, D, E로 표기하였다. 왼쪽 위는 유클리드 거리, 오른쪽 위는 맨하탄 거리, 왼쪽 아래는 MADD, 오른쪽 아래는 ADDM을 고차원 공간의 거리로 이용한 실루엣 점수이다.

실루엣 점수는 -1 과 1 사이의 값으로 주어진다. 1에 가까울수록, 군집화가 잘 되었다고 평가할 수 있다. 전반적으로 음수 값이 나오는 경우도 많아, 군집화가 잘 되었다고 보기 힘들다. t-SNE 결과와 같이, 유클리드 거리와 맨하탄 거리의 점수가 MADD와 ADDM보다 좋다. 또한 5%와 10%의 변수만 이용한 경우 전반적으로 더 높은 점수가 산출되었다.

실루엣 점수 표

		사용한 거리에 따른 실루엣 점수			
사용한 설명변수 개수	암의 종류	유클리드	맨하탄	MADD	ADDM
5%	A	-0.006	0.031	-0.152	0.049
	B	0.160	0.230	-0.128	-0.392
	C	0.449	0.550	0.298	0.010
	D	0.159	0.318	0.103	-0.197
	E	0.279	0.126	0.277	-0.383
	전체	0.152	0.216	0.03	-0.15
10%	A	0.002	0.056	-0.103	-0.154
	B	0.260	0.186	-0.150	-0.421
	C	0.489	0.463	0.241	-0.160
	D	0.204	0.214	0.423	-0.296
	E	0.134	0.269	-0.523	-0.179
	전체	0.170	0.193	0.041	-0.242
30%	A	0.098	-0.095	-0.140	-0.145
	B	0.236	-0.066	-0.287	-0.351
	C	0.496	0.386	0.120	-0.068
	D	0.073	0.131	0.062	-0.162
	E	0.171	0.219	-0.271	-0.228
	전체	0.161	0.068	-0.087	-0.183

표 4.1 **5%, 10%, 30%**의 설명변수 이용 시 실루엣 점수표

		사용한 거리에 따른 실루엣 점수			
사용한 설명변수 개수	암의 종류	유클리드	맨하탄	MADD	ADDM
50%	A	-0.152	-0.221	-0.178	-0.201
	B	0.048	-0.267	-0.330	-0.332
	C	0.542	0.486	0.176	-0.090
	D	0.065	0.064	0.046	-0.142
	E	0.163	0.132	-0.156	-0.222
	전체	0.056	-0.022	-0.091	-0.193
100%	A	-0.145	-0.202	-0.170	-0.232
	B	-0.183	-0.125	-0.217	-0.287
	C	0.314	0.267	0.042	-0.127
	D	0.132	-0.070	-0.163	-0.136
	E	0.081	0.048	-0.121	-0.141
	전체	-0.072	-0.070	-0.147	-0.188

표 4.2 **50%, 100%**의 설명변수 이용 시 실루엣 점수표

실루엣 그림

다음의 그림은 각 관측치의 실루엣 점수를 그래프로 나타낸 것이다. 같은 색은 같은 라벨의 암 종류를 의미한다. 그림의 좌측 상단에는 전체 실루엣 점수의 평균이 표기되었다. 또한 전체 실루엣 점수의 평균은 그림의 가로방향 빨간 점선으로 표시되었다. 각 라벨 A, B, C, D, E의 관측치들의 실루엣 점수 값이 큰 양수이고, 대체적으로 평균보다 크다면, 해당 라벨은 군집화가 잘 되었다고 생각할 수 있다.

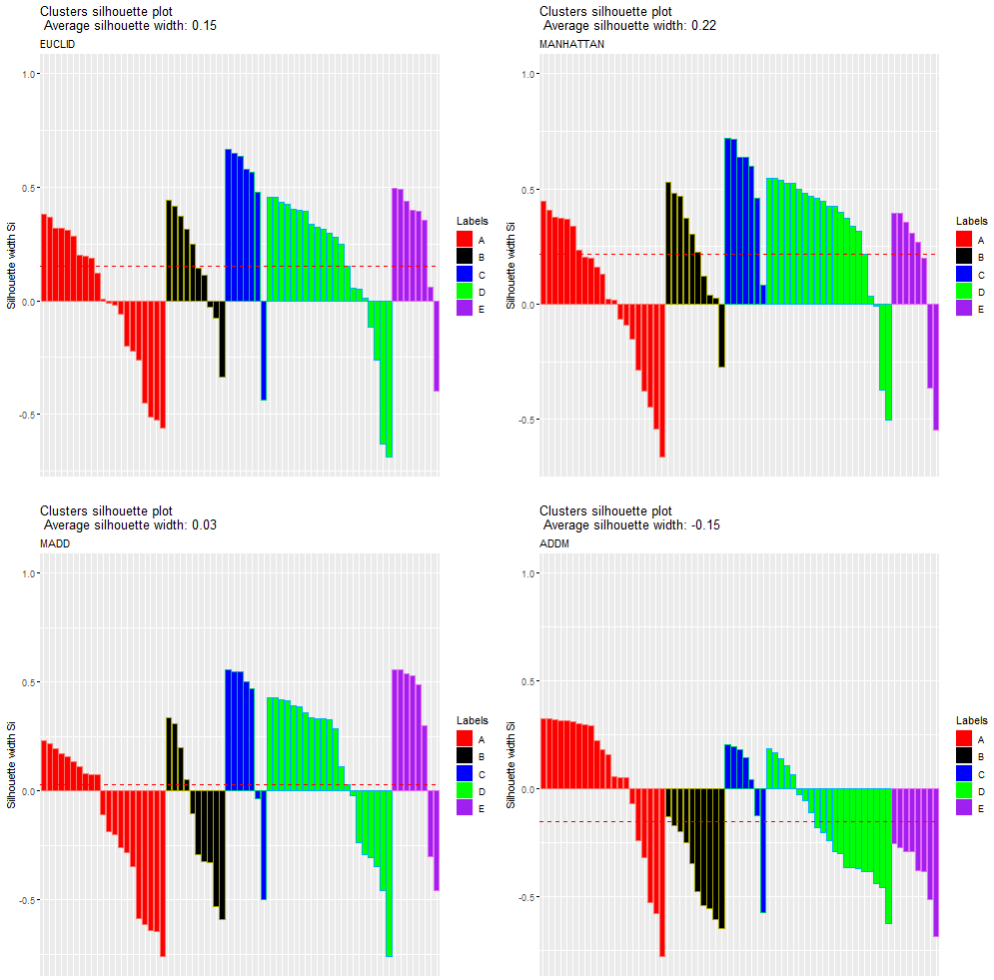

그림 4.6 실루엣 분석 결과 1 : 5%의 변수 이용

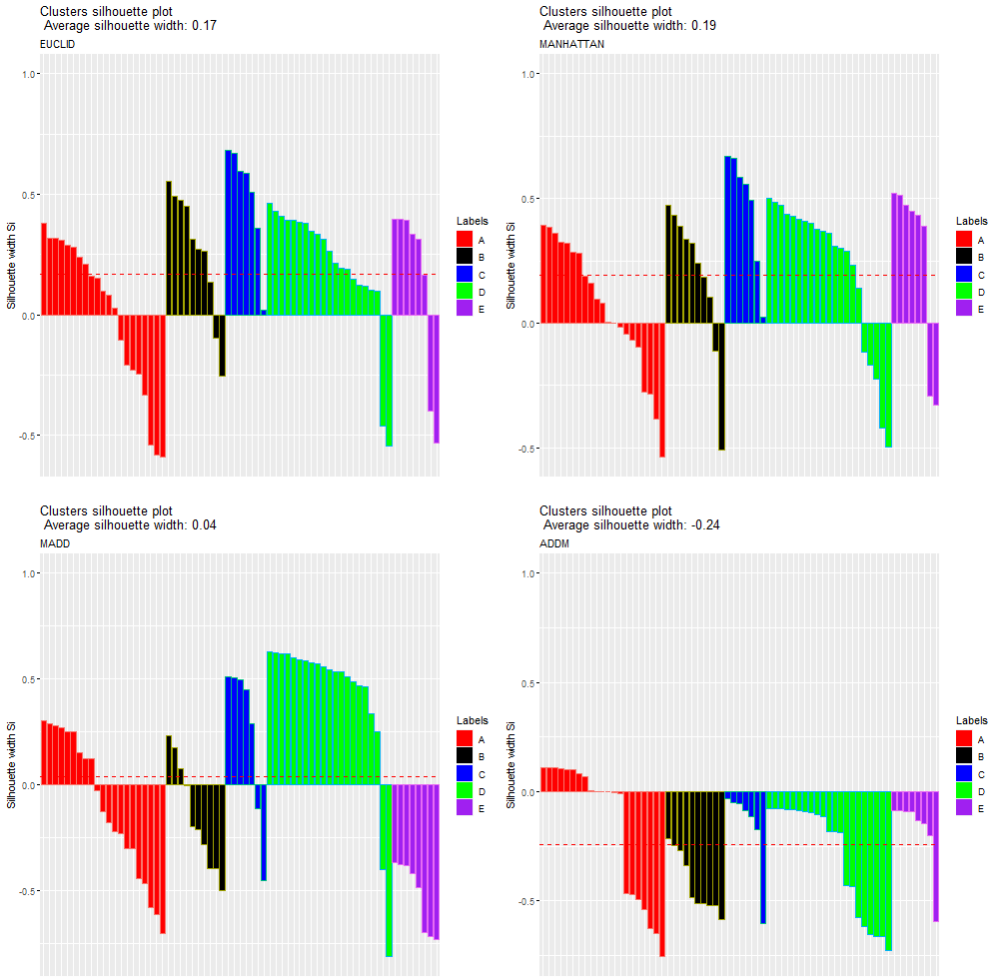

그림 4.7 실루엣 분석 결과 2 : 10%의 변수 이용

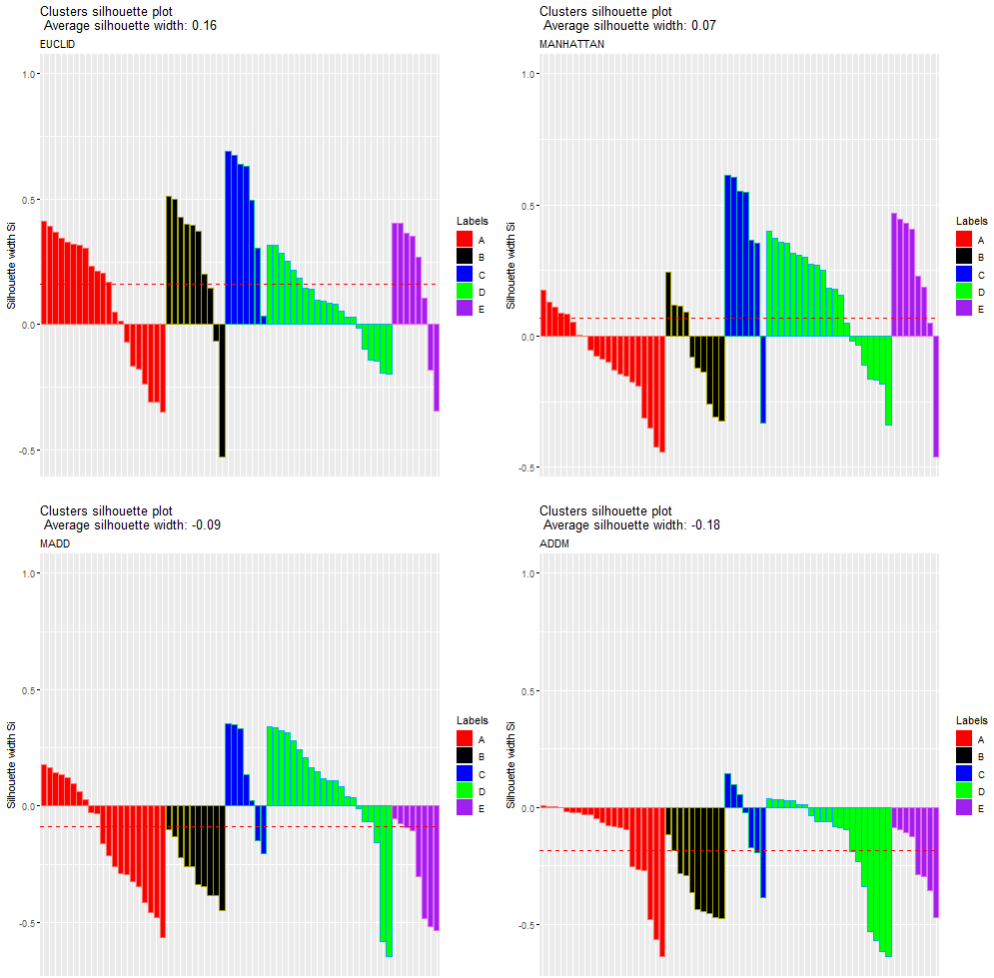

그림 4.8 실루엣 분석 결과 3 : 30%의 변수 이용

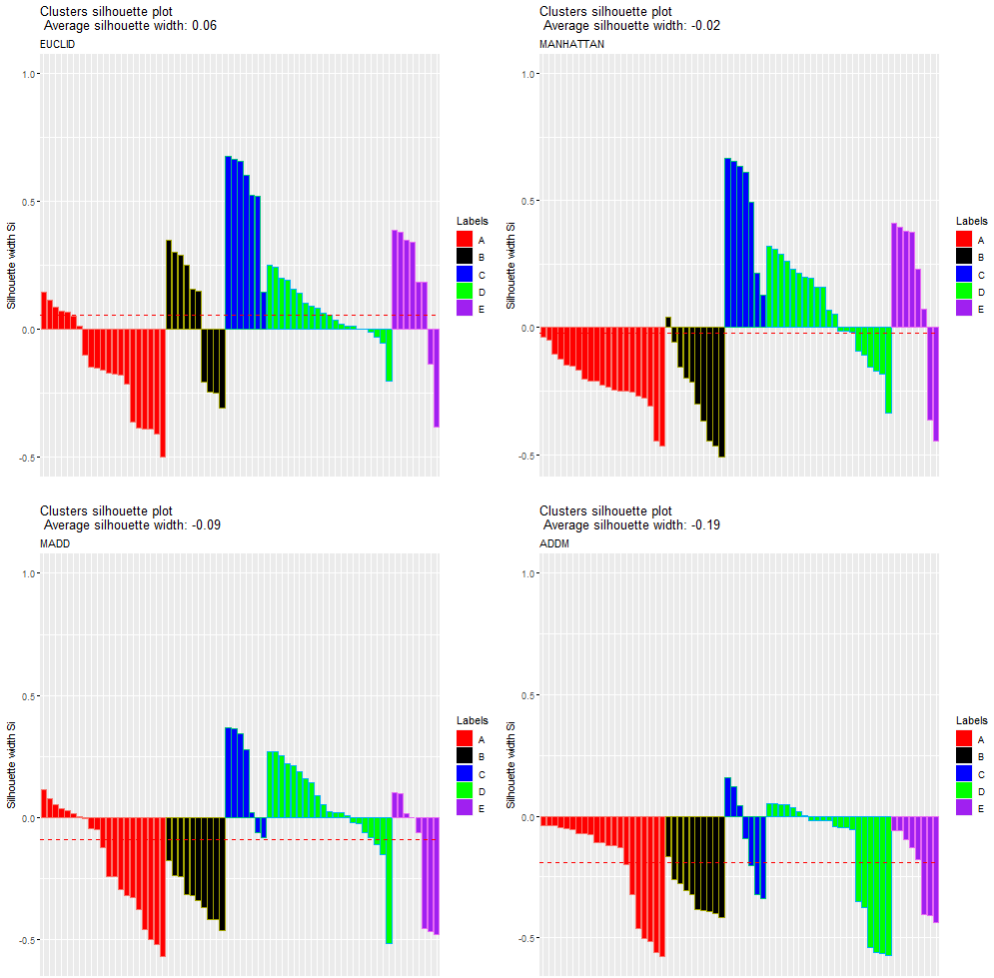

그림 4.9 실루엣 분석 결과 4 : 50%의 변수 이용

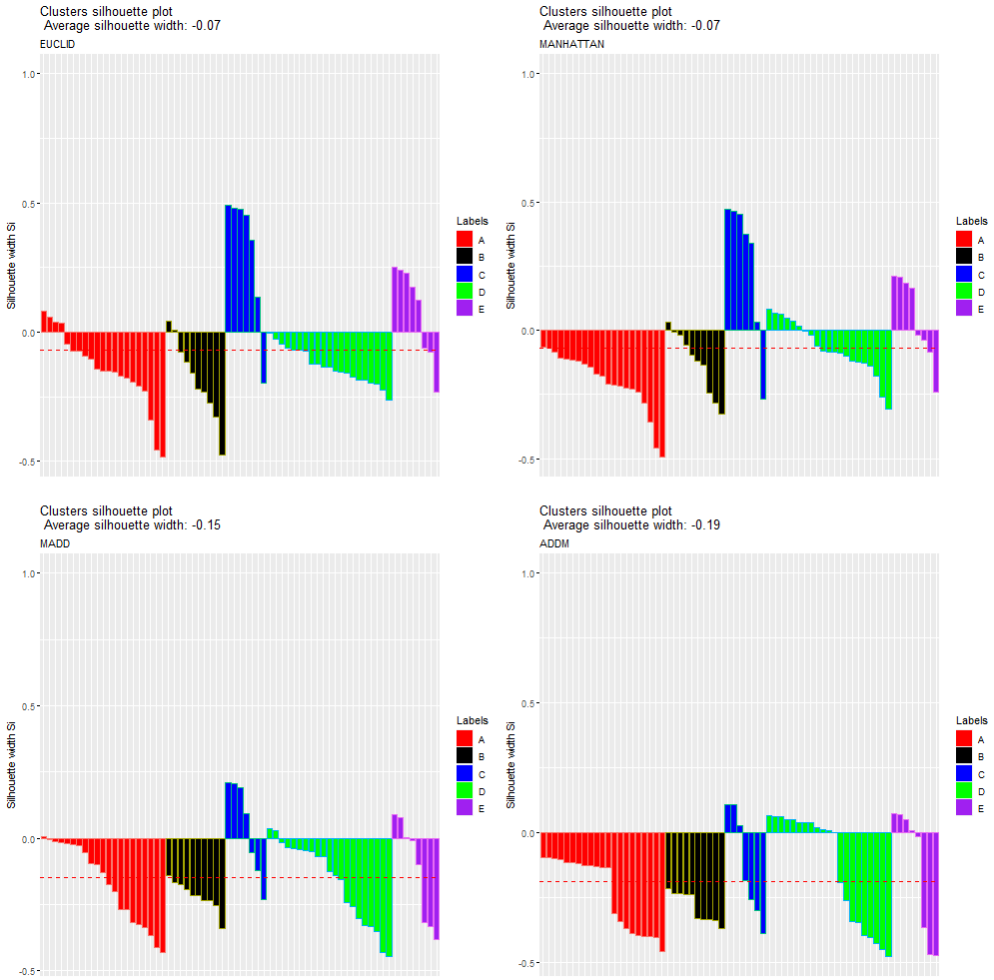

그림 4.10 실루엣 분석 결과 5 : 100%의 변수 이용

4.4.2 다변량 분산분석 결과

사용한 검정

다변량 분산분석의 가정인, 다변량 정규분포와 그룹 간 동일한 공분산 행렬에 대한 내용을 다음의 검정방법을 이용하여 검정하였다. 다변량 정규분포 가정의 경우, 각 그룹별로 유의 수준 0.05에서 Mardia's 검정을 이용하였다. 등분산 가정

의 경우, 유의 수준 0.05에서 Box's M 검정을 하였다. 그 결과 다변량 정규성과 등분산에 대한 검정은 만족, 불만족으로 표기하였다.

다변량 분산분석의 통계량으로 Wilk's Lambda와 Pillai's Trace를 이용하였다. 각각 통계량 값과 p-value를 표기하였다.

검정 결과

		사용한 거리에 따른 검정 결과와 통계량			
설명변수 개수	검정 종류	유클리드	맨하탄	MADD	ADDM
5%	다변량 정규분포 A	만족	만족	만족	불만족
	다변량 정규분포 B	만족	만족	만족	만족
	다변량 정규분포 C	만족	만족	만족	만족
	다변량 정규분포 D	만족	만족	불만족	불만족
	다변량 정규분포 E	만족	불만족	만족	만족
	등분산 가정	만족	불만족	불만족	불만족
	Wilk's Lambda	0.123	0.096	0.165	0.599
	(p-value)	(0.000)	(0.000)	(0.000)	(0.000)
	Pillai's Trace	1.278	1.297	1.036	0.417
	(p-value)	(0.000)	(0.000)	(0.000)	(0.000)
10%	다변량 정규분포 A	만족	만족	만족	불만족
	다변량 정규분포 B	만족	불만족	불만족	만족
	다변량 정규분포 C	만족	만족	만족	불만족
	다변량 정규분포 D	불만족	만족	불만족	만족
	다변량 정규분포 E	만족	불만족	만족	불만족
	등분산 가정	불만족	만족	불만족	만족
	Wilk's Lambda	0.166	0.137	0.329	0.587
	(p-value)	(0.000)	(0.000)	(0.000)	(0.000)
	Pillai's Trace	1.032	1.159	0.782	0.420
	(p-value)	(0.000)	(0.000)	(0.000)	(0.000)
30%	다변량 정규분포 A	만족	만족	만족	만족
	다변량 정규분포 B	불만족	만족	만족	만족
	다변량 정규분포 C	만족	만족	만족	만족
	다변량 정규분포 D	만족	만족	불만족	만족
	다변량 정규분포 E	불만족	불만족	만족	만족
	등분산 가정	만족	만족	만족	불만족
	Wilk's Lambda	0.167	0.232	0.512	0.717
	(p-value)	(0.000)	(0.000)	(0.000)	(0.008)
	Pillai's Trace	1.018	0.847	0.568	0.289
	(p-value)	(0.000)	(0.000)	(0.000)	(0.011)

표 4.3 **5%, 10%, 30%**의 설명변수 이용 시 다변량 분산분석 결과표

		사용한 거리에 따른 검정 결과와 통계량			
설명변수 개수	검정 종류	유클리드	맨하탄	MADD	ADDM
50%	다변량 정규분포 A	만족	만족	만족	불만족
	다변량 정규분포 B	불만족	만족	만족	만족
	다변량 정규분포 C	만족	만족	만족	만족
	다변량 정규분포 D	만족	만족	만족	불만족
	다변량 정규분포 E	불만족	불만족	불만족	만족
	등분산 가정	만족	만족	만족	불만족
	Wilk's Lambda	0.240	0.461	0.494	0.781
	(p-value)	(0.000)	(0.000)	(0.000)	(0.051)
	Pillai's Trace	0.792	0.611	0.578	0.222
	(p-value)	(0.000)	(0.000)	(0.000)	(0.061)
100%	다변량 정규분포 A	만족	만족	만족	불만족
	다변량 정규분포 B	만족	만족	만족	만족
	다변량 정규분포 C	만족	만족	만족	만족
	다변량 정규분포 D	만족	만족	만족	불만족
	다변량 정규분포 E	만족	만족	만족	만족
	등분산 가정	만족	만족	만족	만족
	Wilk's Lambda	0.541	0.474	0.870	0.806
	(p-value)	(0.000)	(0.000)	(0.365)	(0.097)
	Pillai's Trace	0.475	0.542	0.134	0.197
	(p-value)	(0.000)	(0.000)	(0.357)	(0.106)

표 4.4 **50%, 100%**의 설명변수 이용 시 다변량 분산분석 결과표

4.4.3 실루엣 점수와 다변량 분산분석 결과 요약

다음의 표는 실루엣 점수와 다변량 분산분석 결과의 요약이다. 순위가 낮을수록(숫자가 작을수록) 더 큰 점수(p-value가 더 작은 통계량)를 뜻한다. 전반적으로 실루엣 점수, Wilk's Lambda, Pillai's Trace 모두 비슷한 순위를 보인다.

		사용한 거리에 따른 검정 순위			
설명변수 개수	검정 종류	유클리드	맨하탄	MADD	ADDM
5%	실루엣 점수	2	1	3	4
	Wilk's Lambda	2	1	3	4
	Pillai's Trace	2	1	3	4
10%	실루엣 점수	2	1	3	4
	Wilk's Lambda	2	1	3	4
	Pillai's Trace	2	1	3	4
30%	실루엣 점수	1	2	3	4
	Wilk's Lambda	1	2	3	4
	Pillai's Trace	1	2	3	4
50%	실루엣 점수	1	2	3	4
	Wilk's Lambda	1	2	3	4
	Pillai's Trace	1	2	3	4
100%	실루엣 점수	1	2	3	4
	Wilk's Lambda	2	1	4	3
	Pillai's Trace	2	1	3	4

표 4.5 실루엣 점수와 다변량 분산분석 결과 요약표

제 5 장 결론

본 연구에서는 t-SNE와 거리의 집중, Interpoint type distance에 대하여 살펴보았다. 그리고 차원 축소 기법인 t-SNE를 고차원 자료에 활용할 때, 고차원 공간의 다른 거리 사용에 따른 t-SNE 결과를 비교해 보았다. 결과를 평가하는 척도로는 실루엣 점수와 다변량 분산분석(MANOVA)의 Wilk's Lambda와 Pillai's Trace를 사용하였다.

전반적으로 t-SNE의 2차원으로의 차원 축소 결과는 라벨들(A, B, C, D, E)을 잘 분리하지 못하였다. 또한 해당 연구에서는 고차원 공간의 거리의 집중 문제를 완화하기 위해 제시된 Interpoint type distance인 MADD(Mean Absolute Difference of Distance)와 ADDM(Absolute Difference of Distances from the Mean)을 고차원 공간의 거리를 측정하는 것에 사용하였다. 그 결과 기존에 사용하던 유클리드 거리와 맨하탄 거리의 성능이 더 좋게 나타났다.

사전 스크리닝 절차로써 5746개의 변수에 대한 분산분석(ANOVA) 결과, p-value가 작은 변수를 이용하였을 때, t-SNE의 성능이 좀 더 좋게 나타났다. 고차원 자료의 각 변수에 대한 분산분석을 통해서, 의미 없는 변수의 개수를 조금 줄일 수 있다고 생각해 볼 수 있다.

[4]에서는 거리의 집중을 측정하는 여러 가지 인덱스에 대하여 설명하고 있다. 시뮬레이션 상황에서 해당 인덱스들을 이용한다면, Interpoint type distance의 효과와 변수에 대한 사전 스크리닝 절차의 효과를 보다 정밀하게 생각해 볼 수 있을 것으로 기대한다.

참고문헌

[1] Elmore K.L and Richman M.B (2001). Euclidean distance as a similarity metric for principal component analysis, Mathematics, Monthly Weather Review

[2] Frechet M.M. (1906). Sur quelques points du calcl fonctionnel. Rend. Circ. Matem. Palermo 22, 1-72

[3] L.Van der Maaten and G.Hinton (2008). Visualizing Data using t-SNE, Journal of machine learning research 9, 2579-2605

[4] Sushma Kumari and Balasubramaniam Jayaram (2016), Measuring Concentration of Distances - An Effective and Efficient Empirical Index, IEEE Transactions on knowledge and data engineering, Vol 28

[5] Arnab K. Pal, Pronoy K. Mondal, Anil K. Ghosh (2016). High dimensional nearest neighbor classification based on mean absolute differences of interpoint distances, Pattern Recognition Letters 74, 1-8

[6] Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaf(1999), When Is “Nearest Neighbor” Meaningful?, International Conference on Database Theory, 217–235

[7] Damien Francois, Vincent Wertz, Michel Verleysen (2007), The Concentration of Fractional Distances, IEEE Transactions on knowledge anddata engineering, Vol 19

[8] Robert J. Durrant, Ata Kabán(2009), When is 'nearest neighbour' meaningful: A converse theorem and implications, Journal of Complexity 25 385–397

[9] Reza Modarres(2021), Indices of Matrix Dissimilarity

[10] Richard A. Johnson, Dean W.wichern, Applied multivariate statistical analysis(6th), Chapter 6

Abstract

In this study, t-SNE(t-Distributed Stochastic Neighbor Embedding), one of the dimension reduction methods, is applied to high-dimensional data. When applying t-SNE, it is considered to use Interpoint type distance rather than Euclidean distance or Manhattan distance. First, t-SNE, concentration of distance, and Interpoint type distance are examined. After that, t-SNE is applied to high-dimensional gene data using Euclidean distance, Manhattan distance and Interpoint type distance. And compared with the results of pre-screening of gene data, the results are discussed.

Keywords: high-dimensional data, dimension reduction, t-SNE, Concentration of Distance, Interpoint type distance, MADD, ADDM

Student Number: 2020-29533