이 학 석 사 학 위 논 문

대칭양정치행렬자료의
Predictive Clustering

2020년 8월

서울대학교 자연과학대학원

통계학과

김 재 민

이 학 석 사 학 위 논 문

대칭양정치행렬자료의
Predictive Clustering

2020년 8월

서울대학교 자연과학대학원

통계학과

김 재 민

초록

군집분석은 표본공간의 관측치들을 이용하여, 특정한 기준으로 그룹을 만들어주는 것이다. 비지도 학습의 일종으로 Gaussian Mixture Model, K-means, Level-set 방법 등 다양한 방법들이 존재한다. 이 때, 사전지식이 없고, 실제 그룹들이 변수들을 기준으로 잘 분리되어 있다면, 연구의 초기단계에 활용하기 좋은 학습방법이라고 볼 수 있다. 이를 위해 군집분석을 통해 분리한 그룹들이 실제 그룹과 높은 비율로 일치한다면, 이는 해당 방법을 통한 군집분석에 더 높은 신뢰도를 매길 수 있는 과정이 될 것이다. 본 논문은 실제 그룹이 나와있는 데이터에 K-sphere방법을 적용하는 예측적 군집분석[1]을 적용할 것이다. 이 논문에서 주로 다룰 K-sphere방법은 이미 잘 알려진 K-means방법과 Conformal Predictor[2]를 종합하여 나온 군집분석 방법으로 기존의 K-means방법을 많이 보완한 측면이 있다. 이를 바탕으로 공분산행렬 전처리를 거친 이미지 데이터의 군집분석을 적용하고, 실제 그룹과 일치하는 지의 여부를 교차 검증하고자 한다.

주요어: 군집분석, Conformal Prediction, K-means, 공분산 행렬

학번: 2018-24077

목차

제 1 장 서론

Conformal Predictor[2]는 유사하다고 생각되는 그룹 내에서 Exchangability를 만족하는 성질의 Conformity score를 정의 할 수 있다면, Probability Coverage를 만족하는 영역을 형성할 수 있다는 특징이 있다. 이를 군집분석에 적용할 경우, 유사하다고 분류된 한 군집 내에서 특정 관측치가 어느정도의 유사도를 갖는지를 정의할 수 있다. 이에 따라, 각 군집에서도 정도를 나누어 noise가 있는 경우, 이를 군집에서 제외할 수 있으며, 다르게 분류된 군집내의 관측지들이 어느 정도 유사도를 갖는다고 판단하여, 이를 병합할 수 있다. 병합을 통해서는 sphere 형태가 아닌 경우에도 올바른 군집형태를 형성할 수 있으며, 이는 시뮬레이션에서 다룰 예정이다. 해당 내용은 shin의 논문에서도 잘 나타나 있다.[1]

K-means 방법은 군집분석에서 흔히 쓰이는 방법이나, 1)몇몇 문제에서 원하지 않은 결과를 내기도 하고, 2)적절한 K를 선택하는 기준이 명확하지 않다는 문제를 내포하고 있다. 또한 군집은 하나의 제한된 영역이어야 될 것인데, 3)넓은 표본공간 전채를 K개의 Voronoi Tesslation으로 분할하여, 굳이 군집을 부여하지 않아도 될 관측치에도 군집을 부여한다는 점 또한 하나의 문제점으로 볼 수 있다.[1]

기존에 양정치행렬 데이터는 고차원 데이터를 공분산 행렬로 전처리 하는 과정에서 발생하며, 이는 의학 이미지, 패턴분석에서 차원축소의 개념으로 활용된다. 기존의 (가로 픽셀) * (세로 픽셀) * (채널수)로 고차원의 형태인 하나의 사진자료에서는 색상(Red, Green, Blue) 또는 색의 변화(Gradient)등의 다양한 변수들을 추출할 수 있고, 이 변수들의 공분산 구조를 뽑아낼 수 있다. 많은 경우

에 이 공분산 구조가 분류문제에서 유의미한 결과를 도출한다는 것이 나타났다.[3]

양정치 행렬의 형태로 차원축소한 데이터에 K-means방법과 Conformal Prediction를 적용하여 군집분석을 잘 적용할 수 있는지 확인할 것이다. 이 둘을 종합하여 적용하는 것은 K-sphere[1]로 잘 분리된 데이터에 분포가정없이 Probability coverage를 만족하는 범위를 형성해주는 것을 볼 수 있다. 이미지 데이터의 분류에 관해서는 ETH-80 데이터[4; 5]를 활용하였으며, 비지도 학습인 군집분석의 결과가 실제의 label과 잘 맞는지를 확인하여, 유의미하게 적용이 가능한지 확인해보겠다. 분류학습과는 다르게 군집분석은 비지도학습이다. 따라서 데이터에 대한 사전지식이 없는 경우, 초기 연구단계에서도 적용이 가능하다. Conformal Prediction을 이전에 적용되지 않던 양정치행렬 전처리를 거친 데이터의 군집분석에 적용하는 것이 기존의 연구들과 본 연구의 차이점이라고 볼 수 있다.

제 2 장 Conformal Prediction과 K-Sphere

Predictive Clustering은 분석의 전단계에서 진행하는 Clustering이 실제 데이터의 하나의 구분단위로 묶여질 수 있게 끔 만드는 것이 그 목적이다.

2.1 K-means 군집분석

K-means[6]는 가장 보편적이고 쉽게 사용할 수 있는 군집분석 방법이다. Distance 함수(d(x,y))와 그에 따른 중심점(μ)을 정의할 수 있으면, 그에 따라 상황에 맞는 K-means를 적용할 수 있다.

Algorithm 1: K-means 알고리즘

Result: $\mu_i, \mathcal{C}_i \ i = 1, ..., k$

initialize $\mu_{i0} \ i = 1, ..., k$;

while $\mathcal{C}_{i(n-1)} = \mathcal{C}_{in}$ *for all* i **do**

  $\mathcal{C}_{i(n-1)} = \{x_j : i = \text{argmin}_l \, d(x_j, \mu_{l(n-1)})\}$;

  중심점 μ_{in}을 각 $\mathcal{C}_{i(n-1)}$에서 추출한다.;

end

일반적으로 잘 알려진 K-means알고리즘은 1차원 벡터의 적용하며, 이때 $d(x, y)$와 군집 $\mathcal{C}$의 중심점 $\mu_{\mathcal{C}}$는 다음과 같이 적용한다.

$$d(x, y) = ||x - y||_2, \tag{2.1}$$

$$\mu_{\mathcal{C}} = \sum_{x \in \mathcal{C}} x / n_{\mathcal{C}}, \tag{2.2}$$

본 논문에서는 1차원 행렬이 아닌 2차원 행렬, 특히 양정치행렬에 대해 다루고 있다. 행렬의 거리함수 d(X,Y)와 여러 행렬데이터의 중심점을 구하는 방식은 다양하다. 그 중 하나의 형태인 Euclidean distance와 Extrinsic mean을 이용할 예정이며, 그 정의는 아래와 같다. 이 때, S는 표본 $\mathbf{X}_i$이 분포하는 표본공간으로 양정치행렬이 있는 manifold를 의미한다. 따라서, $\mathbf{X}_i(i = 1, \cdots, n)$ 는 S에 분포 Q로 존재한다고 가정하자. 이때 중심점(Extrinsic mean)과 거리함수(Euclidean distance)는 다음과 같이 정의할 수 있다.

Definition 1. $\mathbf{X}, \mathbf{Y} \in Sym^+(p)$ 이며, $\mathbb{Q}$분포를 따르는 $\mathbf{X}, \mathbf{Y}$에 대해 Euclidean distance는 다음과 같이 정의된다.

$$d_E(\mathbf{X}, \mathbf{Y}) = tr(\mathbf{X} - \mathbf{Y})^2 = \|\mathbf{X} - \mathbf{Y}\|^2, \tag{2.3}$$

여기서 $\|\cdot\|$는 Frobenius norm이다.

위의 Euclidean distance는 다면체(manifold :$\mathcal{S}$)를 유클리드 공간으로 보내고, 해당 유클리드 공간상의 거리를 측정한 것이다. 따라서, 이는 추후 연구의 방향성을 넓히는 데에 이용할 수 있을 것으로 보인다.

Definition 2. (Extrinsic mean) 위에서 이야기 했듯이, 다면체를 유클리드 공간으로 보낸 상태에서, 거리함수를 이용한 분포상의 분산을 계산할 수 있으며, 그 분산을 최소화 하는 모든 값을 Extrinsic mean이라 한다.

$$\mu_{\mathbf{E}}(\mathbf{X}) = \underset{\mathbf{P} \in S}{\operatorname{argmin}}\, \sigma^2_{\mathbf{E}}(\mathbf{P}), \tag{2.4}$$

이 때, $\sigma^2_{\mathbf{E}}(\mathbf{P})$는 $\sigma^2_{\mathbf{E}} : S \to \mathbb{R}$로 정의된 분산함수로 다음과 같이 정의된다.

$$\sigma^2_{\mathbf{E}}(\mathbf{P}) = \mathbf{E}(\|\mathbf{X} - \mathbf{P}\|^2) = \int_{\mathcal{S}} \|\mathbf{X} - \mathbf{P}\|^2 \, \mathbf{Q}(d\mathbf{X}) \tag{2.5}$$

Definition 3. (Sample Extrinsic mean)위에서 정의한 extrinsic mean의 표본상에서의 추정값은 우선 분포 $\mathbf{Q}$의 표본상의 추정치인 $\hat{\mathbf{Q}}_n$에서의 Extrinsic mean이다.

$$\hat{\mathbf{Q}}_n = \frac{1}{n}\sum_{i=1}^{n}\delta(\mathbf{X}_i) \tag{2.6}$$

이 때, 분산함수 $\hat{\sigma}^2_{\mathbf{E}}(\mathbf{P})$는 $\hat{\sigma}^2_{\mathbf{E}} : S \to \mathbb{R}$는 표본상에서 다음과 같이 정의된다.

$$\hat{\sigma}^2_{\mathbf{E}}(\mathbf{P}) = \frac{1}{n}\sum_{i=1,\cdots,n} ||\mathbf{X} - \mathbf{P}||^2 \tag{2.7}$$

이에 대한 Sample Extrinsic mean $\hat{\mu_{\mathbf{E}}}$는 다음과 같이 정의할 수 있다.

$$\hat{\mu}_{\mathbf{E}}(\mathbf{X}) = \underset{\mathbf{P}\in S}{\operatorname{argmin}}\, \hat{\sigma}^2_{\mathbf{E}}(\mathbf{P}) \tag{2.8}$$

Theorem 1. *Q를 $Sym^+(p)$상의 분포라 하며, $\sigma^2(\mathbf{P}) < \infty$ 가 몇몇 $\mathbf{P} \in Sym^+(p)$에서 성립할 때, 다음이 성립한다.[7]*

1. *분포 $\mathbf{Q}$를 따르는 확률변수 $\mathbf{X} \in Sym^+(p)$에 대하여, extrinsic mean은 다음의 평균으로 정의된다.*

$$\mu_{\mathbf{E}}(\mathbf{X}) = \mathbf{E}_{\mathbf{Q}}(\mathbf{X}) = \int \mathbf{X}\mathbf{Q}(d\mathbf{X}) \tag{2.9}$$

2. *표본 $X_1, \cdots, X_n \overset{iid}{\sim} \mathbf{Q}$에 대하여, Extrinsic Sample mean은 다음과 같이 산술평균으로 정의된다.*

$$\hat{\mu}_{\mathbf{E}}(\mathbf{X}) = \bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i \tag{2.10}$$

위의 Theorm의 증명을 포함한 다양한 행렬의 거리와 중심점에 관한 내용은 [7]에서 찾아볼 수 있다. Definition1과 Theorem 2.2에 의해서 양정치 행렬에서도

손쉽게 거리와 중심점을 잡을 수 있으며, K-mean의 적용 역시 가능하다.

K-means는 가장 직관적이고 간단히 적용할 수 있는 방법이지만, 1) K의 선택기준을 잡는 것에 대한 명확한 기준이 없고, 2) 보로노이 분할은 우리가 일반적으로 찾고자 하는 군집의 형태와 잘 맞지 않는 모양을 가지고 있으며, 3) 군집이 구체의 형태와 유사하고 잘 분리되었을 때는 효율이 좋지만, 그렇지 않을 때는 결과값의 가치가 크게 떨어진다는 단점이 있다.

2.2 Conformal Predictor

Conformal Prediction[2]는 분포가정을 사용하지 않고도 데이터에 대한 분석을 할 수 있게 해주는 일반적인 방법 중 하나이다. 기본적으로 다음의 알고리즘을 따른다.

Algorithm 2: Conformal 알고리즘

Result: $\mathbf{C}_n(\alpha)$

1. 기존의 데이터 셋에 $\mathbf{Y}_{n+1} = y$를 합친 데이터 셋 $\mathcal{A}(y) = \{Y_1, \cdots, Y_n, y\}$를 정의한다.

2. unconformity score $R_i(y) = \phi(\mathbf{Y}_i, \mathcal{A})$를 모든 $i = 1, \cdots, n, n+1$에 대해서 계산한다.
   이 때, ϕ는 분포상에서 교환성(exchangability)이 성립한다.

3. $\pi(y)$를 다음과 같이 정의한다.

$$\pi(y) = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbf{I}(\mathbf{R}_i(y) \geq \mathbf{R}_{n+1}(y)).$$

4. 위의 과정을 모든 y에 대하여 적용하고 다음의 $\mathbf{C}_n(\alpha)$을 구한다.

$$\mathbf{C}_n(\alpha) = \{y : \pi(y) \geq \alpha\} = \left\{ y : \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbf{I}\Big(\mathbf{R}_i(y) \geq \mathbf{R}_{n+1}(y)\Big) \geq \alpha \right\}$$

위의 기본적인 Conformal 알고리즘의 $\pi(y)$는 $H_0 : \mathbf{Y}_{n+1} = y$의 p-value로 볼 수 있다. 또한 위의 결과로 나타나는 $\mathbf{C}_n$은 α에 대응 되는 신뢰구간으로 볼 수 있다.

이 때 중요한 가정은 $H_0 : \mathbf{Y}_{n+1} = y$에서 Conformity 또는 Unconformity

score인 R_i의 교환성이 성립한다는 것이다. 결국 $\pi(y)$는 $\{k/(n+1) : k = 1, \cdots, n+1\}$에서 균등분포를 따르며, 모든 분포 $\mathbf{P}$와 그 분포를 따르는 새로운 관측값 Y에 대해, 다음 Probability coverage를 만족한다.[2]

$$\mathbf{P}^{n+1}(Y \in \mathbf{C}_n) \geq 1 - \alpha$$

하지만, 알고리즘 2는 매관측값마다 계산을 진행하기에 속도가 현저히 떨어진다는 단점이 있으며, 이를 보완하는 Split Conformal Prediction 알고리즘이 있다. 새 관측값 하나에 대해 알고리즘을 적용하는 것이 아닌 새로운 여러 관측값들에 대해 알고리즘을 적용하는 것으로, 본 논문에서도 Split Conformal Prediction 알고리즘을 사용할 예정이다.

Algorithm 3: Split Conformal 알고리즘

Result: $\mathbf{C}_n(\alpha)$

1. 전체 데이터셋 $\mathcal{Y}$를 $\mathcal{Y}_1, \mathcal{Y}_2$로 나눈다.
2. $\mathcal{Y}_1$의 함수로부터 $\mathbf{Q}$를 계산한다.
3. $\mathbf{R}_i = \phi(\mathbf{Q}, Y_i)$ for $Y_i \in \mathcal{Y}_2$로 Unconformity Score을 구한다.
4. $\mathbf{R}_i$의 1−분위수 t_α를 구한다.
5. $\mathbf{C}_n(\alpha) = \{y : \phi(\mathbf{Q}, y) \leq t_\alpha\}$

이 경우에도 도출된 신뢰집합(Confidence set : $\mathbf{C}_n(\alpha)$)는 분포와 무관하게 다음의 Probability Coverage가 만족된다.[2]

$$\inf_P \mathbf{P}^{n+1}(Y \in \mathbf{C}_n) \geq 1 - \alpha$$

2.3 K-Sphere

K-sphere[1]는 K-means에 Conformal Prediction을 적용한 방법이다. K-means는 중심점까지의 거리가 가까운 순으로 군집을 탐지하기에, 군집의 형태가 표본공간을 보로노이 테셀레이션(Voronoi Tesslation)의 형태로 나누게 된다. 포본공간이 $\mathbb{R}^2$이고, 군집이 두개인 경우, K-means로 형성된 군집은 2차원 공간을 한 직선으로 둘로 쪼갠 형태가 된다. 이는 군집의 명확한 형태를 반영했다고 보기는 힘들며, 이후의 분석에 이용하기도 힘들다. 일반적인 Conformal알고리즘에 비해 속도를 빠르게 하기위해 Split Conformal 알고리즘을 사용하였으며, K-sphere은 다음과 같다.

Algorithm 4: K-Sphere 알고리즘[1]

Result: $\hat{\mathcal{C}}_i \ i = 1, ..., \hat{l}$

1. 데이터 $\mathcal{Y}$를 $\mathcal{Y}_1, \mathcal{Y}_2$로 나눈다.
2. K-means를 $\mathcal{Y}_1$에 적용하여, $c_1, \cdots, c_k$의 중심점을 구한다.
3. $\mathcal{Y}_2$의 각 데이터에 가장 가까운 중심점까지의 거리를 R_i를 계산한다. 이 때, $c_{j(i)}$는 $\mathcal{Y}_i$와 가장 가까운 중심점이다.
4. R_i의 1-α 분위수 t_α를 구한다.
5. y_i가 c_j에 가장 가깝고, R_i가 t_α보다 작을 경우, y_i를 $\mathcal{C}_j$에 배정한다.
6. 1부터 5를 여러 k에서 반복적으로 측정하고, 가장 $\mu(\mathcal{C}_j)$가 작은 k를 $\hat{k}$라 한다.
7. 교집합이 공집합이 아닌 $\mathcal{C}_j$들을 병합하여 최종적인 $\hat{l} \leq \hat{k}$개의 $\hat{\mathcal{C}}_i$를 뽑는다.

다만 이런 상황에서 위에서 정의한 $R_i = \min_j ||Y_i - c_{(j)i}||$ 는 Conformal Prediction의 기본 조건인 교환성(exchangability)를 만족하지 않는다. 따라서, 아래와 같은 R_i를 적용할 경우, 더 Conformal Prediction의 기본가정을 만족하는 C를 얻을 수 있다.

$$R_i = \min_{j \in \{1, \cdots, k\}} \left[\frac{||Y_i - c_{(j)i}||^2}{\hat{\sigma}_j^2} + 2d \log \hat{\sigma}_j - 2 \log \hat{\pi}_j \right] \tag{2.11}$$

이 때, $\hat{\pi}_j = n_j/n$으로 n_j는 $\mathcal{Y}$의 K-means에 의해 나타난 j번째 보로놀리 테셀레이션인 $\mathbf{V}_j$에 속한 관측치의 개수를 의미한다. 그리고 $\hat{\sigma}_j$는 다음과 같이 정의된다.

$$\hat{\sigma}_j^2 = \frac{1}{n_j} \sum_{Y_i \in \mathbf{V}_j} ||Y_i - \bar{Y}_j||.$$

양정치 행렬의 Euclidean norm과 extrinsic sample mean은 벡터의 유클리드 거리와 평균값과 동일하므로, 위와 동일한 distance하에서 간단히 K-sphere을 적용할 수 있었다. 다만 중심점(c_j) 역시 행렬의 형태를 띈 $\mathbf{C}_j$
로 표기하고, $|| \cdot ||$이 Frobenius norm으로 변경된다.

다만, 위의 알고리즘을 그대로 적용하여 $\mathcal{C}_{\hat{\mathbf{k}}}$를 얻을 경우, 각 군집들은 공통영역을 표본공간 내에서 갖기도 한다. 따라서, 일단 공유하는 공통영역이 있을 때, 이를 병합하여, $\hat{\mathcal{C}}_{\hat{l}}$로 놓도록 한다.($\hat{l}$: $\hat{k}$개의 $\mathcal{C}_i$를 병합한 후의 개수)

시행횟수에 따라 결과에 변동도 발생한다. 실제로 군집의 병합은 조심스럽게 이루어져야 한다. 약하게 연결된 두개의 군집을 하나의 군집으로 취급하는 경우가 발생할 수 있고, 원래 연결되어야 했을 두개의 군집을 병합하지 못하는 경우도 발생할 수 있다. 이를 탐지하기 위해서는 이는 반복적으로 $\mathcal{Y}_1, \mathcal{Y}_2$ 복원추출하여,

K-sphere를 반복시행(Bagging)한 결과들을 기반으로 안정적인 군집의 형태를 추출하는 것이 타당하다.

2.4 K의 선택

위에서 추출한 $\hat{\mathcal{C}}$는 병합하기 전에는 정확한 구체(sphere)의 형태이다. K는 이 구체의 개수를 몇개로 할 것인가에 대한 문제이다. 각 구체는 중심점(c_j)과 반지름(r_j)로 정의할 수 있으며, $\mathbf{B}(C_j, r_j) = \{Y : ||Y - C_j|| \leq r_j\}$로 표현 할 수 있다. 위의 내용을 응용하여 Conformal Predictor를 적용하는 표본공간내의 집합 $\mathcal{M}_k$는 다음과 같이 정의할 수 있으며 Conformal Predictor의 기본성질과 종합하면 다음과 같은 결과를 얻는다.

Lemma 2. **K** $= k$*로 K-Sphere의 결과를 얻은 상황에서 중심* $\mathbf{C}_j$*와 반지름* r_j *에 대하여,*

$$\mathcal{M}_k = \cup_{j=1}^{k} \mathbf{B}(C_j, r_j). \tag{2.12}$$

이 때, r_j*는 아래와 같이 정의되며*

$$r_j = \hat{\sigma}_j \sqrt{\max(t_\alpha + 2\log\hat{\pi}_j - 2d\log\hat{\sigma}_j, 0)}. \tag{2.13}$$

Conformal Predictor의 성질로 $\inf_{\mathbf{P}} \mathbf{P}^{n+1}(\mathbf{Y} \in \mathcal{M}_k) \leq 1 - \alpha$*가 만족한다.*

K가 커질수록 다양한 모양의 군집을 찾아낼 수 있다. 이는 이후의 실험을 통해 보여줄 수 있을 것이다. 하지만 너무 커질 경우, 계산량이 많아지는 문제가 발생한다. 최적의 k는 전체 표본공간 상에서 가장 작은 부피를 차지하게 하는 k로 선택하도록 한다.

이 때, $\mathcal{M}_k$의 부피는 Importance Sampling[8]을 이용하며, 사용된 분포 $f(\mathbf{x})$는 아래와 같다.

$f(\mathbf{x}) = \frac{1}{k}\sum_{i=1}^{k} f_i(\mathbf{x}) \qquad$ where $f_i(\mathbf{x})$: uniform distribution in i-th sphere

이는 아래와 같은 사실을 이용하며, 토대가 $\mathcal{M}_k$를 포함하는 확률변수 $\mathbf{X}$와 확률밀도함수 $f : \mathbf{X} \to \mathbb{R}$ 에 대하여,

$$Vol(\mathcal{M}_k) = \int_{\mathcal{M}_k} 1dx = \int_{\mathcal{M}_k} \frac{1}{f(x)} f(x)dx = \mathbf{E}\left(\frac{1}{f(x)}\right) \tag{2.14}$$

이며, 부피는 아래와 같은 알고리즘으로 $\hat{\mathbf{E}}(1/f(x))$를 추정하는 것과 동일해진다.

Algorithm 5: $\mathcal{M}_k$의 부피추정

Result: $\hat{\mathbf{E}}\left(\frac{1}{f(x)}\right)$

1. f(x)에서 표본$\{x_1, \cdots ,_n\}$들을 임의추출한다.
2. 각 표본에 해당하는 $1/f(x_i)$값을 구한다.
3. $\hat{\mathbf{E}}\left(\frac{1}{f(x)}\right) = \frac{1}{n}\sum_{i=1}^{k} \frac{1}{f(x_i)}$

제 3 장 데이터 설명

3.1 실험 데이터

3.2 시뮬레이션 1 : Mickey mouse Problem

Mickey mouse Problem은 일반적으로 K-means로 풀리지 않는 형태로 잘 알려져 있다. 중심점이 두 귀의 접합부에 위치하여, 가장 큰 군집을 포함하게 군집이 형성된다. 이는 아래의 그림에서 확인할 수 있다.

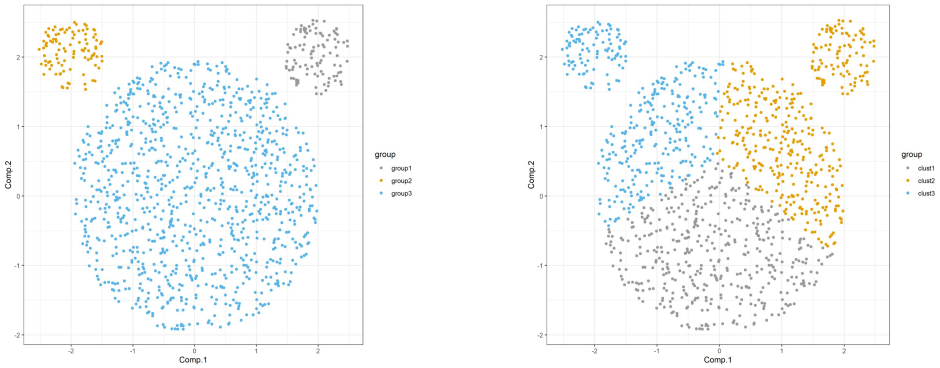

그림 3.1 (좌) Mickey Mouse 데이터 (우) K-means 군집분석 시

Mickey Mouse problem을 푸는 방법은 GMM 등 실제로 여러가지가 존재한다. K-sphere를 통해 잘 분리된 경우에 각 군집을 잘 구분해 내는지를 확인해보도록 하겠다. 해당 데이터는 2차원 공간의 원안에서 각각 균등하게 100, 1000, 100개의 관측치를 가지는 세개의 원으로 미키마우스의 모양의 데이터를 추출했다.

3.3 시뮬레이션 2 : Wishart-분포 데이터

양정치 행렬 표본을 쉽게 뽑아낼 수 있는 분포는 대표적으로 Wishart 분포가 있다. 3차원 공간에 표현이 가능한 점을 이용하기 위해 2×2의 양정치행렬을 이용하였다. Figure2를 보면 알 수 있듯이 구체에 가깝지 않은 형태를 띠고 있기에, 구체와 다른 형태일 때, K-sphere가 잘 작동하는지, 양정치 행렬의 Sample Extrinsic mean과 Euclidean distance를 적용했을 때도 잘 적용이 되는지를 확인해볼 수 있을 것이다. Wishart분포는 $\mathbf{W}_p(\mathbf{\Sigma}, n)$ 로 나뉘며, $\mathbf{\Sigma}$와 자유도(n)를 모수로 갖는다.

그림 3.2 (좌) Wishart 시뮬레이션 데이터 (우) K-means 적용시

각각의 Cluster는 다른 분포를 따르며, 잘 분리되게끔 만들었다. 각 그룹별로 100, 200, 100개의

- 그룹1 : $\mathbf{\Sigma}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, n_1 = 2$
- 그룹2 : $\mathbf{\Sigma}_2 = \begin{pmatrix} 5 & 4.5 \\ 4.5 & 5 \end{pmatrix}, n_2 = 10$
- 그룹3 : $\mathbf{\Sigma}_3 = \begin{pmatrix} 1 & -0.9 \\ -0.9 & 1 \end{pmatrix}, n_3 = 50$

그룹별 산포는 Figure 2(좌)에서 확인할 수 있다.

3.4 시뮬레이션 3 : ETH-80 데이터

ETH-80 데이터는 [4; 5] 에서 사용된 데이터로 10종의 사물의 총 3280개의 사진을 담고 있다. 각 사물별로 410개의 사진이 있으며, 사진의 크기는 동일하다. 한 사물에도 다른 색상을 가진 사진이 있고 다른 각도에서 촬영한 사진이 있어, 완전히 한 군집이 뭉쳐있는 형태로 보기는 힘들 것으로 예상되는 표본이었다. 해당 사진의 각각의 픽셀을 바탕으로 공분산행렬의 형태로 전처리를 진행한 이후의 데이터 셋에 대해 K-sphere 군집분석을 진행하도록 하겠다.

그림 3.3 (좌, 중) ETH-80 데이터, (우)라플라시안 필터 적용시 형태변화

그림 3.4 ETH-80 토마토와 배사진

제 4 장 이미지 전처리(CovD)

고차원에서는 K-sphere의 연산량이 크게 증가하며, 따라서 효율성이 떨어진다. 이미지 데이터는 기존의 (가로 픽셀) * (세로 픽셀) * (채널수)로 고차원의 형태이다. 이를 낮은 차원의 데이터로 새롭게 가공하는 방법 중 하나가 몇몇 변수들의 사진내에서의 공분산 구조를 추출하는 것이다. 하나의 사진자료에서 색상(Red, Green, Blue) 또는 색의 변화(Gradient)등의 다양한 변수들을 추출하여, 이 변수들의 공분산 구조(Covariance Descriptor)를 뽑아낼 수 있다. 이런 공분산 구조가 이미지 분류문제에서 유의미한 결과를 도출한다는 것이 나타났다.([3])

이미지의 각 픽셀은 (Red, Greed, Blue) 3개의 변수를 포함하고 있다. 여기에 더해 변화율은 윤곽선(edge)과 관련된 변수로 사물의 형태와 연관성이 있다. 따라서 이를 반영하고자 변화율(Gradient)와 관련된 변수를 한가지 추가하였다. 즉, 각 픽셀별로 (Red, Green, Blue, Gradient) 4가지 변수가 있었으며, 이를 바탕으로 4공분산 행렬로 각각의 사진데이터를 차원축소했다.
사진데이터를 단순히 저차원의 공분산 행렬로 만들었을 때, 처리된 데이터가 잘 분리가 되지 않는 현상이 일어났기 때문에 10종의 사물 중 2종((cow, tomato), (pear, tomato))을 선별하여 이에 대해 잘 작동하는지 분석을 진행하였다.

4.1 라플라시안 필터

Gradient관련 변수로는 라플라시안 필터를 통하여 각 픽셀별로 Greyscale의 Gradient를 추가하였다. 라플라스 필터는 다음과 같이 4방향의 변화율을 동일하게 반영하는 것으로 아래의 행렬을 사진의 각 픽셀에 적용하게 된다. 이는 아래의 식을 사진의 각 부분에 적용하는 것과 동일한 효과를 가진다.

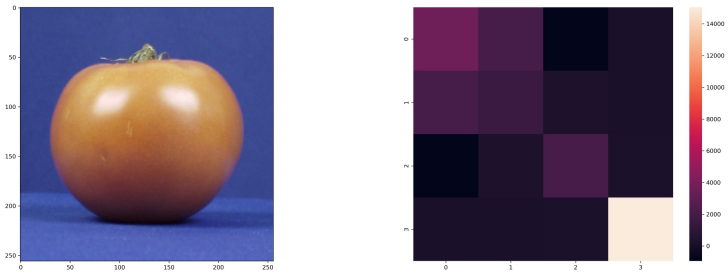

그림 4.1 (좌) Wishart 시뮬레이션 데이터 (우) K-means 적용시

$$\text{Laplacian Filter} : \begin{pmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{pmatrix} \qquad \Delta f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}$$

사물의 회전으로 형태에 변화를 준 ETH-80 데이터에 적합할 것이라 생각했다. 다음과 같은 커널을 모든 그림의 픽셀에 적용하면 그림3.3과 같이 그림이 변화하여, 사물의 형태를 반영할 수 있게 된다.

제 5 장 결과

5.1 시뮬레이션 1 : Mickey mouse Problem

원하는 형태로 Cluster가 잘 분리된 것을 확인 할 수 있었다. 이는 일반적인 K-means 군집분석(그림 3.1)에 비해 개선 된 것으로 볼 수 있으며, 군집이 나뉜 것은 다음과 같이 그림상으로 확인할 수 있다. Figure4는 K = 11, α = 0.05의 상황이다.

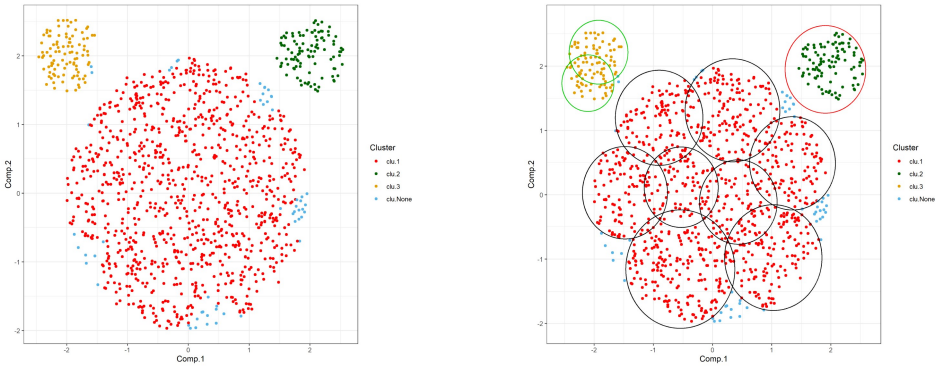

그림 5.1 Mickey Mouse 데이터의 K-sphere 적용

	K = 11, α = 0.05			Bagging n = 100		
	그룹1	그룹2	그룹3	그룹1	그룹2	그룹3
군집1	1.000	0.000	0.000	1.000	0.000	0.000
군집2	0.000	0.980	0.000	0.000	1.000	0.000
군집3	0.000	0.000	0.938	0.000	0.000	0.997
Noise	0.000	0.020	0.062	0.000	0.000	0.003

다만 해당 데이터와 같이 noise가 없는 경우, 오히려 군집에 포함되어야 할

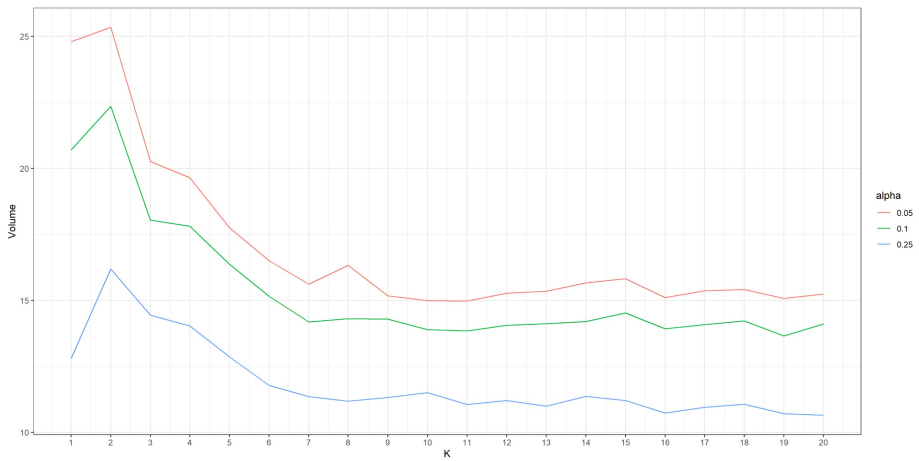

그림 5.2 K vs Volume 그래프

것이 Noise로 분류 되는 경우가 발생한다. 이 경우엔 α를 조정하거나, 배깅(Bagging)을 이용하여, 정확도를 향상시킬 수 있다. 그 결과는 위의 표를 참고하면 알 수 있다. 다만 이 경우, 위에서 설정한 α값보다 작은 Probability Coverage를 갖게 되므로, 이항분포의 분위수를 활용하는 것이 타당한 것으로 보인다. 위의 경우는 Noise가 없는 데이터의 형태였기에 보다 좋은 결과를 나타냈다.

5.2 시뮬레이션 2 : Wishert-분포 데이터

이 경우, 차원의 증가로 인한 것인지 K에 따른 분석결과의 차이가 많이 발생하였으며, 2차원에 비해 stable하지 않았다. 그러나 $\alpha = 0.05$에서 부피가 최저인 지점은 K = 35로 나타났다. K-means 방법과 달리 각 그룹을 혼동하여 군집화하는 경우가 발생하지 않았으며, Noise로 분류한 비율도 $\alpha = 0.05$와 유사하게 나타났다.

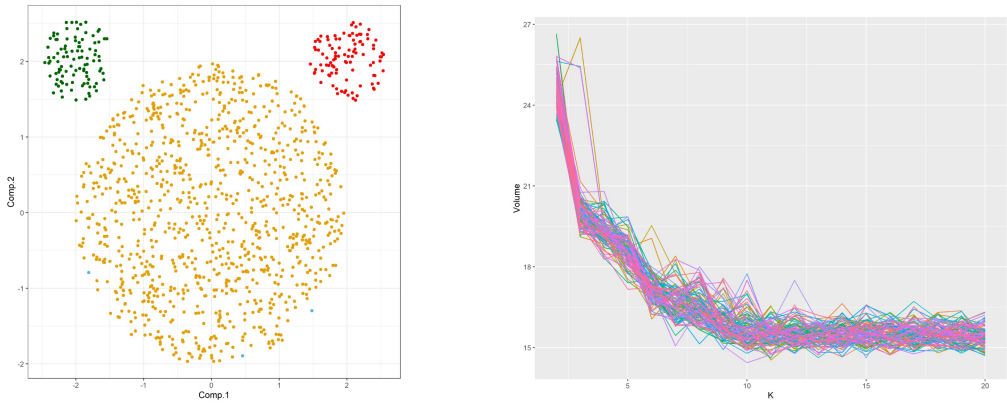

그림 5.3 Mickey Mouse 데이터의 K-sphere($\alpha = 0.05$)와 Bagging(n = 100) 적용

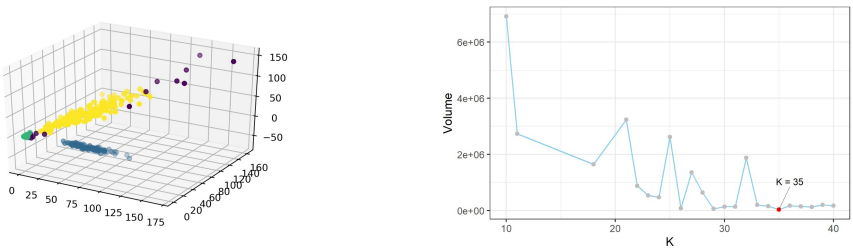

그림 5.4 Wishart 데이터의 K-sphere 적용

K = 35, $\alpha = 0.05$			
	그룹1	그룹2	그룹3
군집1	0.980	0.000	0.000
군집2	0.000	0.935	0.000
군집3	0.000	0.000	1.000
Noise	0.020	0.065	0.000

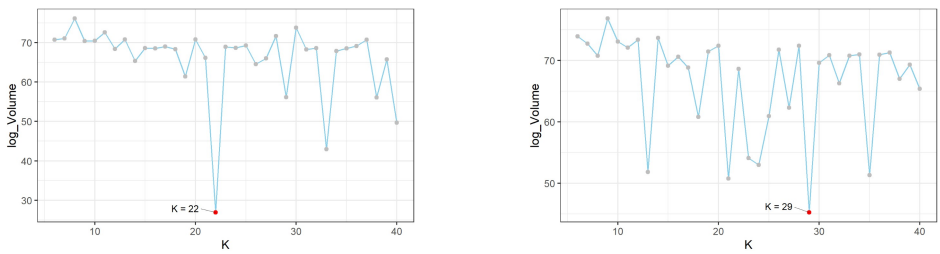

그림 5.5 ETH-80 소(좌)와 배(우)의 K vs log-Volume

5.3 ETH-80 데이터

5.3.1 토마토 vs 소

그림 5.5에서 확인 할 수 있듯이, K = 22에서 가장 낮은 부피를 기록하였으며, 두 사물을 혼동하지 않고, 각각 0.822, 0.963로 각 그룹의 사물을 잘 탐지하였다. 토마토의 경우, 그림 3.4의 두번째 그림과 같이 초록색이 없는 토마토를 노이즈로 분류하는 경우가 발생하였다. 단순히 4×4 행렬로 CovD를 진행했음에도, 두 그룹이 잘 분류가 되어서 좋은 결과로 나타난 것으로 보인다.

5.3.2 토마토 vs 배

위의 예시보다 조금 더 유사한 사물을 군집분석하는 과정으로 토마토와 배 사진의 군집분석을 진행하였다. 두 종류의 데이터 모두 시행마다 결과는 조금씩 다르게 나타났다. $\alpha = 0.05$로 놓았을 경우 전체 데이터를 하나로 묶는 현상이 자주 일어나서, $\alpha = 0.1$로 조정하여 분석을 진행하였다.

Tomato vs Cow(K = 22)			Tomato vs Pear(K = 29)		
$\alpha = 0.1$	cow	tomato	$\alpha = 0.1$	pear	Tomato
군집1	0.963	0.000	군집1	0.888	0.000
군집2	0.000	0.822	군집2	0.000	0.924
Noise	0.037	0.178	Noise	0.112	0.176

결과는 위와 같으며, $\alpha = 0.05$로 놓았을 때는 최적의 K를 선택했음에도 모든 군집이 하나로 묶이는 현상이 빈번하게 일어났다. 여기서 더 나아가 α를 더 키우면 좀 더 세분화된 군집을 얻을 수도 있을 것으로 보인다. 다만 ETH-80 데이터는 소, 토마토, 배와 같은 하나의 사물을 구분점으로 보고 정리가 된 데이터이기 때문에 최대한 α를 낮게 설정하였고, 그 결과 적정선에서 군집이 형성되었으며, 두 그룹을 혼동하는 경우는 발생하지 않았다. 이는 데이터를 CovD로 전처리한 것에서 잘 분리가 되었기 때문으로 보인다.

제 6 장 맺음말

K-sphere는 행렬데이터의 분류에서도 잘 작동하는 것을 확인 할 수 있었다. 이와 마찬가지로 다양한 형태의 데이터에 응용이 가능할 것으로 보이며, Generalized K-sphere와 K-elipsoid방법[1] 또한 다만 α를 작게 할 경우 군집을 너무 크게 설정하여. 실제로 나뉘어야 할 군집이 병합이 되는 경우가 발생하기도 했다. 마찬가지로 너무 크게 할 경우, 하나의 분류로 묶여야 할 군집이 다르게 분류되는 현상도 발생했다. 따라서 또한 이 때, α의 선택에 연구의 목적에 따른 조절이 가능하며, 선택의 기준점을 마련하는 것도 필요함을 알 수 있다.

K-sphere는 고차원과 2개 이상의 군집에서도 잘 작동하나, 여기에는 기본적으로 데이터가 군집이 잘 나뉘는 것을 표현하고 있어야 된다는 전제가 붙는다. 그렇기 때문에 위의 논문의 방식인 CovD가 아닌 다른 방식으로 데이터를 표기하거나, Euclidean Distance이 다른 Distance를 이용한 방법을 적용할 수도 있다. 이를 통해 카테고리별로 잘 분리된 데이터 형태를 만들 수 있다면, 좀더 완성도 높은 예측적 군집분석도 가능할 것으로 보인다.

참고문헌

[1] J. Shin, A. Rinaldo, and L. Wasserman, "Predictive Clustering," *ArXiv eprint*, vol. 1903.08125v2, 2019.

[2] G. Shafer, A. Gammerman, and V. Vovk, *Algorithmic Learning in a Random World.* Manhatten, New York: Springer, 2005.

[3] K.-X. Chen and X.-J. Wu, "Component spd matrices: A low-dimensional discriminative data descriptor for image set classification," *Computational Visual Media*, vol. 4, no. 3, pp. 245–252, 2018.

[4] K.-X. Chen, J.-Y. Ren, X.-J. Wu, and J. Kittler, "Covariance descriptors on a gaussian manifold and their application to image set classification," *Pattern Recognition*, p. 107463, 2020.

[5] K.-X. Chen, X.-J. Wu, J.-Y. Ren, R. Wang, and J. Kittler, "More about covariance descriptors for image set coding: Log-euclidean framework based kernel matrix representation," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

[6] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning.* Manhatten, New York: Springer, 2009.

[7] A. Schwartzman, "Random ellipsoids and false discovery rates: statistics for diffusion tensor imaging data," 01 2006.

[8] G. H. Givens and J. A. Hoeting, *Computational Statistics.* Hoboken, New Jersey: John Wiley Sons, Inc., 2012.