

## 형상점 자료의 변환을 통한 분류모형의 비교\*

김준홍<sup>1</sup>, 강창완<sup>2</sup>, 최용석<sup>3</sup>

### 요 약

Dryden, Mardia(2016)는 형상분석(shape analysis)을 기하적 공간상에서 형상점들(landmarks)에 의해 나타난 개체들(objects)의 형상(shape)을 측정 및 기술하며 이를 비교하는 분석이라 정의한다. 형상분석에 활용되는 형상점 자료는 배열(array)의 형태를 가지고 있는데, 이는 형상의 분류모형에 적용하기 어려움이 있는 형태이다. 본 연구는 형상점 자료를 분류모형에 적용하기 위한 두 가지 방법을 제시한다. 첫 번째 방법은 GPA(Generalized Procrustes Analysis) 적합을 적용한 후 리만거리, 중심크기 및 형상주성분분석을 통한 형상 주성분점수를 얻어내어 이를 변수로 선택하는 방법이고, 두 번째 방법은 각 표본의 형태행렬(configuration matrix)을 벡터화하는 과정을 진행하여 좌표값들을 모두 변수로써 사용하는 방법이다. 새롭게 얻어진 자료들을 랜덤 포레스트(random forest), 로지스틱 회귀분석(logistic regression analysis), 그리고 서포트벡터머신(support vector machine, SVM) 등 3가지의 분류모형에 투입하여 자료별 오분류율을 비교한다. 이를 위하여 Dryden, Mardia(2016)의 세 군집의 척추뼈 형상자료인 쥐(mice) 자료에 대하여 위의 방법을 적용하였다. 특히, 이 자료는 형상점의 수가 적은 2차원 자료인데 회전 모수가 군집 결정에 영향을 끼쳤으며, 적합과정이 이루어지지 않은 원자료나 중심화 자료의 벡터화가 더 좋은 결과를 보였다.

주요 용어 : 형상분석, GPA적합, 형태행렬, 랜덤 포레스트, 로지스틱회귀, SVM

### 1. 서론

Dryden, Mardia(2016)는 형상분석을 기하적 공간상에서 형상점들에 의해 나타난 개체들의 형상을 측정 및 기술하며 이를 비교하는 분석이라 정의하고 있다. 여기서 형상은 어느 곳에서나 다양하게 존재하는 개체들을 표현할 때 주로 사용되며 이를 군집별로 분류하는 데 활용된다. 대표적인 예로 병원의 엑스레이 자료를 통해 환자들의 병을 진단하거나, 단백질 구조의 형상을 통해 특정 단백질을 식별하는 과정 등이 있다.

기본적으로 형상점 자료는 배열의 형태를 가지고 있는데, 이는 그대로 분류를 위한 모형에 적용하기 어려움이 있는 형태이다. 본 연구는 이러한 형상점 자료를 기존의 방법인 형상 특성변수의 변수화와 새롭게 좌표의 변수화를 제시하려 한다.

\*이 논문은 제1저자 김준홍의 석사학위논문 của 축약본입니다.

<sup>1</sup>46241 부산광역시 금정구 부산대학교 63 부산대학교 통계학과, 석사졸업. E-mail: wnsghd4685@naver.com

<sup>2</sup>47340 부산광역시 부산진구 가야동 산 24 동의대학교 산업경영빅데이터공학과, 교수.

E-mail: cwkwang@deu.ac.kr

<sup>3</sup>(교신저자) 46241 부산광역시 금정구 부산대학교 63 부산대학교 통계학과, 교수. E-mail: yschoi@pusan.ac.kr

[접수 2025년 2월 25일; 수정 2025년 3월 20일; 게재 확정 2025년 3월 23일]

먼저 2절에서는 형상점 자료를 분류모형에 적용하기 위한 배열 자료의 행렬화를 시도하려 한다. 이를 위한 첫 번째 방법은 일반화 프러크루스티즈 분석인 GPA 적합과정을 통해 위치, 척도 및 회전 요인을 고려한 다양한 형상 특성변수들을 변수화시키는 것이고, 두 번째 방법은 개체들의 형태 행렬(configuration matrix)을 벡터화하여 형상점 좌표값들이 변수로 변환되는 방법이다. 이를 위해 고려되는 자료로는 첫번째 방법에서는 중심화 형태행렬을 고려하고 두번째 방법에서는 개체들 형상의 형상점들에 의한 원자료와 중심화 자료, 그리고 위치, 척도 및 회전 요인을 고려한 다른 방법인 북스테인 좌표(Bookstein coordinates)를 사용한다. 그리고 3절에서는 2절에서 제시한 다양한 자료를 얻는 방법을 토대로 Dryden과 Mardia(2016)의 세 군집의 척추뼈 형상자료인 쥐(mice)자료에 대하여 자료들로 변환하고, 군집의 분류를 위해 랜덤 포레스트, 로지스틱 회귀분석, 그리고 SVM 등 3개의 분류모형을 적용하여 오분류율을 비교하려 한다(Chun, 2023; Ko, Kim, 2024; Kim et al., 2023; Lee, Yeo, 2024). 마지막으로 4절에서는 분석결과에 대한 의의 및 한계점을 제시하려 한다.

## 2. 이론적 배경

### 2.1 형상적 자료

기본적으로 형상점 자료는  $k$ 개의 형상점과  $m$ 차원으로 구성된  $k \times m$  형태행렬이 개체 개수  $n$  개만큼 배열로서 합쳐진 자료형태이며, 이중  $m$ 은 형상 데이터 특성상 2차원 혹은 3차원으로 나타내기 때문에 2 혹은 3의 값이다. 여기서  $i$ 번째 형태행렬  $X_i$ 에 대하여 아래와 같이 표현할 수 있다. 여기서  $x_{jl}^{(i)}$ 는  $i$ 번째 형태행렬의  $j$ 번째 형상점을 나타내는 행과  $l$ 번째 차원을 나타내는 열의 원소를 뜻한다.

$$X_i = \begin{pmatrix} x_{11}^{(i)} & x_{12}^{(i)} & \cdots & x_{1m}^{(i)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{k1}^{(i)} & x_{k2}^{(i)} & \cdots & x_{km}^{(i)} \end{pmatrix}, i = 1, \dots, n$$

일반적인 자료는 개체와 변수에 의한 행렬 형태의 다변량 자료로 구성되어 있는 반면, 형상점 자료는 개체별 차원의 형상점에 의한 3차원 배열의 형태로 구성되어 있다. 따라서, 배열 형태의 자료를 분류모형에 사용하기 위해서는 행렬 형태로 변환하여야 한다. 따라서, 배열 형태의 형상점 자료를 새로운 변수 생성을 통해 행렬 형태로 전환하는 과정이 필요하다. 2.2절에서는 GPA 적합을 진행한 형상점 자료들에 대하여 각각 리만거리, 중심크기 및 형상 PC 점수를 계산하여 분류모형의 변수로 선정하는 방법을 제시하고 있다. 그리고 2.3절에서는 원자료의 형태행렬의 중심화와 북스테인 좌표의 벡터화를 통해 좌표값을 변수로 변환하는 방법을 제시하고 있다.

### 2.2 GPA(Generalized Procrustes Analysis)

$m$ 차원 상에서  $k$ 개의 형상점들에 의해 표현되는 중심화 형태행렬  $X_1, \dots, X_n$ 이 있다고 가정한다. 여기서 중심화 형태행렬은 2.2.1절의 중심크기를 구할때 중심화 행렬에 의해 중심점의 변환을 고려한 형태행렬이다. 이들 형태행렬을 크기가  $k \times m$ 인 평균형상  $\mu$ 의 모집단으로부터 추출한 확률표

본이라고 한다면, 다음으로부터 평균형상을 추정하고, 각각의 형태행렬  $X_i$ 에 대한 중심  $\alpha_i$ , 변환  $\Gamma_i$ , 크기  $\beta_i$ 를 고려한 프로크러스티즈좌표  $\hat{X}_i$ 를 얻을 수 있다. 그리고 이렇게 얻어진 프로크러스티즈좌표가 GPA에 의한 적합좌표가 된다. GPA 적합좌표와 추정된 평균형상은 각각 아래 식과 같다.

$$\hat{X}_i = \hat{\beta}_i X_i \hat{\Gamma}_i + 1_k \hat{\alpha}_i^T, i = 1, \dots, n$$

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n \hat{X}_i$$

여기서  $\hat{\alpha}_i$ 는 크기  $m \times 1$ 의 위치벡터 추정치,  $\hat{\beta}_i > 0$ 는 척도모수 추정치,  $\hat{\Gamma}_i$ 는  $m \times m$  크기의 특수 직교회전행렬로, 이는 기존  $n$ 개의 형태행렬들이 각각 위치, 척도 및 회전모수에 대하여 적합이 이루어지는 과정을 표현한다. Dryden, Mardia(2016)은 이렇게 새롭게 생성된 GPA 적합좌표들에 대하여 각각 리만거리, 중심크기 및 형상 PC 점수를 계산하여 이를 형상 특성변수로서 제시하였고, Choi(2021)는 이러한 형상 특성변수들을 SVM과 전통적 관별분석 등의 분류 기법에 사용하였다.

### 2.2.1 중심크기

중심화 행렬  $C = I_k - 1_k 1_k^T / k$ 를 통해 중심점을 (0, 0), 혹은 (0, 0, 0)으로 이동시켜 위치효과를 적합시킨 중심화 형태행렬인  $X_C$ 를 활용한다. 여기서 중심화 행렬  $X_C$ 는 아래와 같이 정의된다.

$$X_C = CX = \left( I_k - \frac{1_k 1_k^T}{k} \right) X$$

이렇게 중심화 형태행렬로 변환된 자료  $X_C$ 의 행렬 놈(norm)을 계산한 결과가 형태행렬의 중심크기(centroid size)이다. 이에 대한 계산식은 다음과 같다. 여기서  $i$ 와  $j$ 는 각각 형상점의 번호와 차원을 뜻한다.

$$S(X) = \|X_C\| = \sqrt{\sum_{i=1}^k \sum_{j=1}^m (x_{ij} - \bar{x}_j)^2}$$

### 2.2.2 리만거리

크기가  $k \times m$ 인 두 형태행렬  $X_1$ 과  $X_2$ 를 고려할 때, 크기가  $(k-1) \times k$  헬머트 부행렬에 의한  $(k-1) \times m$ 의 헬머트화 형태행렬  $X_H = HX$ 와 이에 대한 사전형상  $Z_H = X_H / S(X)$ 의 정의를 통해 두 가지의 사전형상  $Z_1$ 과  $Z_2$ 를 계산할 수 있다. 그리고  $Z_1$ 과  $Z_2$ 는 사전형상 공간  $\mathbb{R}^{(k-1)m}$ 에 속하는 구의 형상  $[X_1]$ 과  $[X_2]$ 에 속하는데, 이를 사전형상구라고 정의한다. 여기서,  $X_1$ 과  $X_2$ 의 사전형상구에서 사전형상  $Z_1$ 과  $Z_2$  간에서 가장 근접하게 큰 원을 리만거리(Riemannian distance)로 아래와 같이 표현한다. 특히, 비정칙값  $\lambda_k, k=1, \dots, m$ 은 비정칙값분해  $Z_2^T Z_1 = U \Lambda V^T$ 를 만족하고  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_m)$ 는  $m \times m$ 인 대각행렬로  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{m-1} \geq |\lambda_m|$ 를 만족하는 비정칙값들을 대각원소로 가지며,  $U$ 와  $V$ 는 직교행렬이다.

$$\rho(X_1, X_2) = \cos^{-1} \left( \sum_{i=1}^m \lambda_i \right)$$

### 2.2.3 형상 PCA(Principal Component Analysis)

Dryden, Mardia(2016)은 프로크러스티즈 잔차(residual)에 대한 PCA를 이용한 형상변동(shape variability)을 설명하였다. 즉, 2.2절의 GPA 적합에 따라 구성된 GPA 적합좌표  $\hat{X}_1, \dots, \hat{X}_n$ 가 구해졌다고 가정하였을 때, 적합좌표  $\hat{X}_i$ 의 형상과 추정된 평균형상  $\hat{\mu}$  간의 차이를 나타내는 프로크러스티즈 잔차

$$R_i = \hat{X}_i - \hat{\mu}, i = 1, \dots, n$$

에 PCA를 적용하여 형상변동을 평가할 수 있다.

또한, GPA 적합 좌표에 대한 크기가  $km \times km$ 인 공분산행렬

$$S = \frac{1}{n-1} \sum_{i=1}^n \text{vec}(\hat{X}_i - \bar{X}) \text{vec}(\hat{X}_i - \bar{X})^T$$

의 고유값-고유벡터를 고려한다. 여기서  $\text{vec}(\cdot)$ 는 행렬을 벡터화하는 함수이고  $\bar{X} = n^{-1} \sum_{i=1}^n \hat{X}_i$ 는 추정된 평균형상과 같다. 그리고, 공분산행렬  $S$ 의 고유값을 크기순으로 나열하면  $\lambda_1 \geq \dots \geq \lambda_p \geq 0$ 이고, 이에 대응하는 고유벡터를 각각  $\gamma_j, j=1, \dots, p$ 라고 나타낼 수 있다. 이는  $j$ 번째 주성분을 제공하고, 이들의 고유벡터는  $\gamma_j^T \gamma_j = 1$ 이며,  $\gamma_i^T \gamma_j = 0, i \neq j$ 를 만족한다. 여기서  $p = \min(n-1, M)$ 이고  $M = 2k-4$ 는 형상공간의 차원을 나타낸다. 따라서 형상분석에서의  $j$ 번째 주성분에 대한  $i$ 번째 표본의 주성분점수는 다음 식을 통해 계산할 수 있다. 본 논문상에서는 주성분분석을 형상 주성분분석(PCA), 주성분점수를 형상 PC 점수라고 표현한다.

$$z_{ij} = \gamma_j^T \text{vec}(\hat{X}_i - \bar{X}), i = 1, \dots, n, j = 1, \dots, p$$

### 2.3 형태행렬의 벡터화

$n$ 개의 형태행렬을 벡터화하는 방법으로, 앞에서 제시한  $i$ 번째 형태행렬  $X_i$ 를  $1 \times km$  크기의 벡터  $Y_i$ 로 벡터화하며 최종적으로  $n$ 개의 표본들을 포함하는 행렬  $Y$ 에 대하여 다음과 같이 표현할 수 있으며, 이는 각각의 좌푯값들이 변수로 변환된 행렬로 재구성된 자료이다.

$$Y_i = (x_{11}^{(i)} \dots x_{k1}^{(i)} | \dots | x_{1m}^{(i)} \dots x_{km}^{(i)}), i = 1, \dots, n$$

$$Y = (Y_1' \dots Y_i' \dots Y_n')'$$

#### 2.3.1 원자료의 중심화(Centralization)

원자료의 경우 크기, 회전 및 위치에 대하여 보정이 이루어지지 않는다. 따라서 중심화를 통하여 위치에 대한 적합을 진행하여 표본들의 위치를 보정하는 전처리 작업을 제시한다.  $i$ 번째 형상 표본의 중심점  $M_i$ 에 대해서 다음과 같이 표현할 수 있다. 여기서  $m$ 은 차원을 나타낸다.

$$M_i = \frac{1}{k} \left( \sum_{j=1}^k x_{j1}, \sum_{l=1}^k x_{l2}, \dots, \sum_{p=1}^k x_{pm} \right)$$

다음으로  $i$ 번째 형태행렬의 중심점  $M_i$ 을 원점  $(0, 0)$ , 혹은  $(0, 0, 0)$ 으로 이동을 시키는 작업을 진행한다. 이러한 작업은  $i$ 번째 형상 표본의 형상점 좌표들을  $M_i$ 의 좌푯값만큼 감하는 방향으로 진행되며, 위의 2.2.1절의 중심크기에서 제시한 중심화 형태행렬  $X_C$ 를 구하는 것과 일치하며, 이를 통해 좌푯값을 변수화하여 행렬화한다.

### 2.3.2 북스테인 좌표(Bookstein Coordinates)

북스테인 좌표는 Bookstein(1986)이 제안한 좌표 체계로서, 두 개의 기저 형상점을 고정시켜 척도 및 이동, 회전모수를 변환하는 좌표 체계이다. 2차원 상의  $k(\geq 3)$ 개의 형상점  $(x_{j1}, x_{j2})$ ,  $j=1, \dots, k$ 가 존재한다고 가정할 때, 두 개의 기저 형상점 사이의 제곱 유클리드거리는

$$D_{12}^2 = (x_{21} - x_{11})^2 + (x_{22} - x_{12})^2 > 0$$

이다. 여기서, 두 개의 기저 형상점은 각각  $(-0.5, 0)$ 과  $(0.5, 0)$ 으로 고정시킨 후 나머지 형상점들에 대하여 아래 식을 통하여 기존 좌표에서 위치, 척도 및 회전에 의해 변환된 북스테인 좌표  $(u_j^B, v_j^B)$ ,  $j=3, \dots, k$ 가 형성된다.

$$u_j^B = [(x_{21} - x_{11})(x_{j1} - x_{11}) + (x_{22} - x_{12})(x_{j2} - x_{12})] / D_{12}^2 - 1/2$$

$$v_j^B = [(x_{21} - x_{11})(x_{j2} - x_{12}) - (x_{22} - x_{12})(x_{j1} - x_{11})] / D_{12}^2$$

## 3. 사례분석

실제 형상점 자료를 이용하여 2절에서 소개한 방법을 통해 두 가지의 변수 선택 방법에 따른 분류 과정을 진행한다. 첫 번째는 2.2절의 GPA 적합을 통해 모든 형상에 대해 크기, 위치, 방향 모수를 변환시킨 후 얻어지는 형상 특성변수들(리만거리, 중심크기, 형상 PC 점수)를 독립변수로 채택하여 분류 과정을 진행하는 과정이다. 그리고 두 번째는 2.3절에서 소개한 방법과 같이 GPA를 통한 적합과정을 배제하고 각 점의 좌푯값을 변수화하여 이들을 독립변수로 채택 후 분류 과정을 진행하는 과정이 된다. 단, 각 점의 좌푯값은 원자료와 중심화 자료, 그리고 북스테인 적합 자료가 사용되며, 북스테인 적합 자료는 2차원 자료에서만 활용할 수 있다. 이러한 분석을 통해, 두 가지 방법 중 어느 방법이 더 효율적인지, 혹은 적합을 진행한 자료와 진행하지 않은 자료 중 어느 자료에서 좋은 결과가 나오는지 확인하는 방향으로 진행한다. 활용 사례로는 Dryden, Mardia(2016)가 제시한 세 군집의 쥐의 척추뼈 형상점 자료를 사용한다.

### 3.1 쥐의 척추뼈 형상점 자료

이 절에서 사용하는 자료는 R의 shapes 패키지에 내장된 쥐(mice)자료이다. Dryden, Mardia(2016)가 제시한 78마리의 척추뼈 형상점 자료이며, 각 개체 당  $x$ 와  $y$ 의 값으로 구성된 2차원 형상점이 6개 존재한다. 또한, 각각 23마리의 큰 쥐 군집( $C_1$ ), 좌우를 반전시킨 30마리의 조절 군집( $C_2$ ), 그리고 23마리의 작은 쥐( $C_3$ ) 총 세 개의 군집으로 구성되어 있다. Figure 1는 세 군집에 포함된 개체들의 형상 그림이며, 중간에 위치한 조절 군집의 방향이 큰 쥐 군집과 작은 쥐 군집의 방향과

다르게 좌우가 반전되어 있음을 확인할 수 있다. 그리고 Figure 2는 군집별 평균 형상 그림과 이에 대한 형상점 위치를 표기한 그림이다. 조절 군집의 좌표 위치 중 1번과 2번의 좌표가 큰 쥐와 작은 쥐 군집과 다르게 반전되어 있음을 확인할 수 있다. 이는 조절 군집은 단순히 좌우만 반전된 것이 아니라, 상하 또한 반전된 자료임을 확인할 수 있다.

Table 1은 원자료의 군집별 좌푯값을 요약한 것으로,  $x_i$ 와  $y_i$ 는 각각  $i$ 번째 좌표의  $x$ 와  $y$  값을 의미한다. 전체적으로 조절 군집의 편차가 크게 형성되어 있음을 확인할 수 있다. 이러한  $x$ 와  $y$  값들은 설명변수로 사용되어 군집에 대한 분류 과정을 진행할 것이다. 분석 기법으로는 랜덤 포레스트와 로지스틱 회귀분석, 그리고 SVM을 사용하였다. 세 가지 분류모형 결과, 랜덤 포레스트에서는 3.95%, 로지스틱 회귀분석에서는 6.41%, SVM에서는 5.26%의 오분류율을 기록하였다. 전체적으로 좋은 분류율을 보이고 있으며, 분석 모형 중에서는 랜덤 포레스트가 가장 좋은 결과를 보였다.

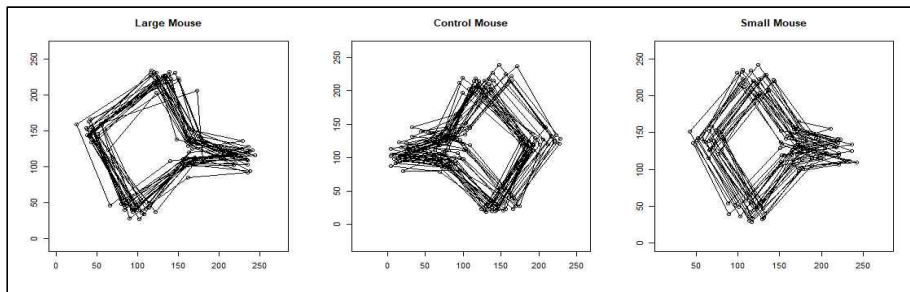


Figure 1. shapes of tree cluster in the raw data

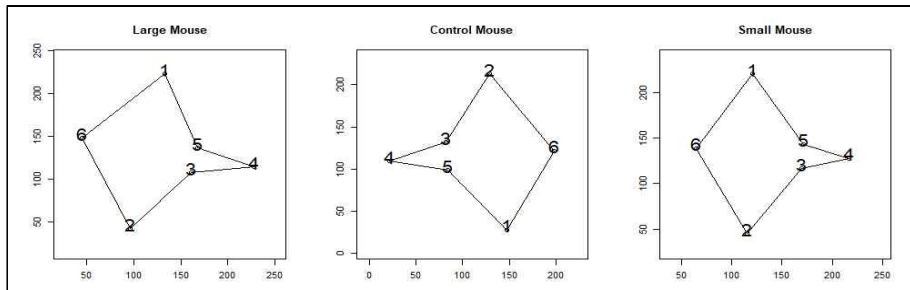


Figure 2. mean shapes of tree cluster in the raw data

Table 1. Summary of coordinates values by cluster (raw data)

| Large Mouse (n=23)   |       |       |       |       |       |       |       |       |       |       |       |       |
|----------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
|                      | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ | $y_6$ |
| Avg                  | 132.7 | 96.5  | 161.1 | 227.2 | 167.4 | 45.3  | 223.3 | 42.2  | 107.5 | 114.1 | 136.9 | 148.5 |
| SD                   | 13.6  | 13.2  | 6.8   | 11.0  | 7.8   | 8.5   | 7.7   | 7.1   | 7.6   | 10.3  | 8.3   | 10.1  |
| Control Mouse (n=30) |       |       |       |       |       |       |       |       |       |       |       |       |
|                      | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ | $y_6$ |
| Avg                  | 147.1 | 129.3 | 82.3  | 22.1  | 84.1  | 198.6 | 28.1  | 212.9 | 132.3 | 109.5 | 99.1  | 121.8 |
| SD                   | 14.2  | 21.1  | 13.9  | 13.7  | 12.2  | 15.4  | 8.3   | 10.0  | 13.0  | 15.5  | 13.0  | 9.5   |
| Small Mouse (n=23)   |       |       |       |       |       |       |       |       |       |       |       |       |
|                      | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ | $y_6$ |
| Avg                  | 121.0 | 115.7 | 169.3 | 216.7 | 171.1 | 65.14 | 220.2 | 44.9  | 116.1 | 127.3 | 143.0 | 138.6 |
| SD                   | 13.4  | 12.8  | 11.2  | 13.4  | 11.4  | 11.4  | 11.1  | 10.0  | 11.5  | 12.9  | 11.3  | 11.1  |

원자료의 경우 위치, 방향 및 크기에 대해서 적합이 이루어지지 않았는데, 이번 중심화 자료는 중심점을 원점 (0, 0)으로 통일시켜 위치에 대한 적합을 진행하였다. Figure 3은 중심화 자료에 대한 세 가지 군집의 형상 그림이며, Figure 2와의 차이를  $x$ 축과  $y$ 축에서 확인할 수 있다. 또한, Figure 4는 중심화 자료에 대한 군집별 평균 형상 그림과 이에 대한 형상점 위치를 표기한 그림으로, Figure 3과 같이 중심점이 (0, 0)으로 통일되었기 때문에  $x$ 와  $y$ 의 값이 수정되었음을 확인할 수 있다.

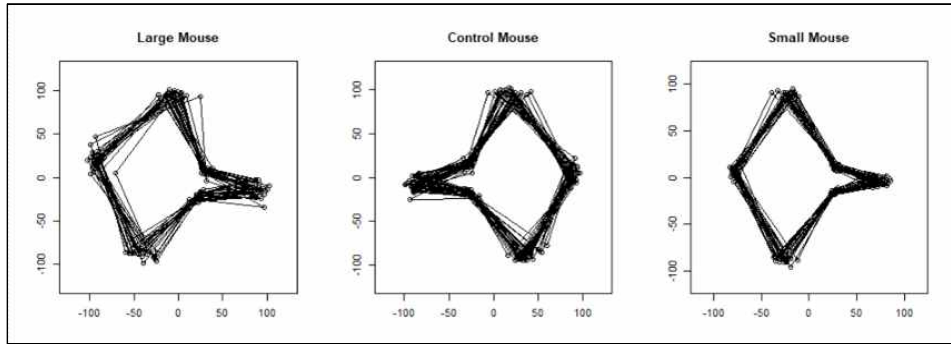


Figure 3. three shape diagrams of centered data

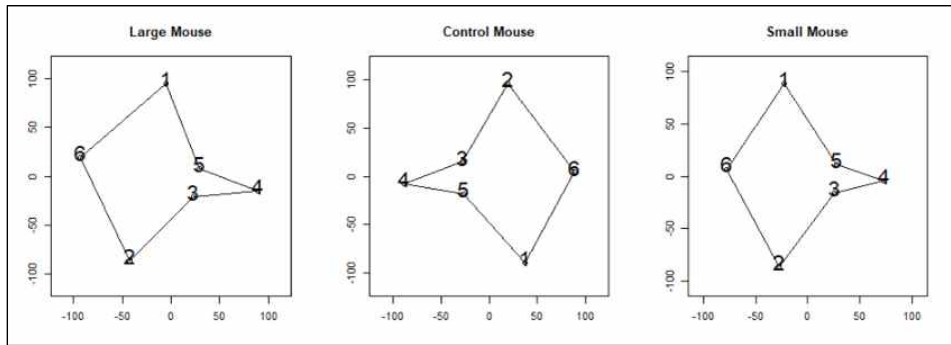


Figure 4. mean shapes of tree cluster in the centered data

Table 2. Summary of coordinates values by cluster (centered data)

| Large Mouse (n=23)   |       |       |       |       |       |       |       |       |       |       |       |       |
|----------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
|                      | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ | $y_6$ |
| Avg                  | -5.7  | -42.0 | 22.7  | 88.9  | 29.1  | -93.0 | 94.6  | -86.5 | -21.3 | -14.7 | 8.1   | 19.7  |
| SD                   | 10.3  | 10.9  | 4.1   | 7.9   | 3.0   | 6.3   | 3.7   | 5.7   | 4     | 7.7   | 4.1   | 10.0  |
| Control Mouse (n=30) |       |       |       |       |       |       |       |       |       |       |       |       |
|                      | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ | $y_6$ |
| Avg                  | 36.5  | 18.7  | -28.3 | -88.5 | -26.5 | 88.0  | -89.2 | 95.7  | 15.0  | -7.8  | -18.2 | 4.5   |
| SD                   | 9.4   | 10.8  | 3.3   | 4.9   | 4.8   | 4.0   | 4.0   | 3.6   | 4.0   | 7.0   | 3.5   | 7.6   |
| Small Mouse (n=23)   |       |       |       |       |       |       |       |       |       |       |       |       |
|                      | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ | $y_6$ |
| Avg                  | -22.2 | -27.5 | 26.2  | 73.5  | 28.0  | -78.0 | 88.6  | -86.9 | -15.6 | -4.4  | 11.3  | 6.9   |
| SD                   | 6.3   | 6.4   | 2.1   | 8.5   | 1.9   | 2.8   | 2.8   | 3.3   | 2.0   | 3.7   | 2.5   | 4.8   |

Table 2는 중심화 자료의 군집별 좌푯값을 요약한 것으로,  $x_i$ 와  $y_i$ 는 각각  $i$ 번째 좌표의  $x$ 와  $y$ 값을 의미한다. 전체적으로 좌푯값들의 편차가 원자료에 비해 감소하였음을 확인할 수 있고, 특히 전체적으로 조절 군집의 편차가 크게 형성되어 있음을 확인할 수 있다. 예시로 원자료의 조절 군집에서  $x_2$ 의 편차가 21.1이었으나, 중심화 자료의 조절 군집에서의  $x_2$ 의 편차는 10.8로 크게 감소하였다. 또한, 원자료의 조절 군집에서  $y_2$ 의 편차가 15.5였으나, 중심화 자료의 조절 군집에서의  $y_4$ 의 편차는 7.0으로 크게 감소하였다. 이는 원자료에 대하여 중심화 전처리를 통해 형상점 자료들이 위치적으로 적합되어 정리가 되었다고 판단할 수 있다. 한편 중심화 자료의 좌푯값을 변수로 채택한 세 가지 분류모형 결과를 보면 랜덤 포레스트와 SVM에서는 2.63%, 로지스틱 회귀분석에서는 7.80%의 오분류율을 기록하였는데 이는 앞의 원자료와 똑같이 전체적으로 좋은 분류율을 보이고 있으며 전체적으로 로지스틱 회귀분석을 제외하고 원자료보다 더 좋은 결과를 형성하였다.

앞의 원자료의 경우 적합이 이루어지지 않았고, 중심화 자료는 위치에 대한 적합이 시행되었다. 이번에는 북스테인 좌표 적합을 통해 위치, 척도 및 회전에 대하여 적합을 진행한다. Figure 6은 북스테인 자료에 대한 세 가지 군집의 형상 그림이며, 기존 원자료 및 중심화 자료와 다르게 방향이 일치됨을 확인할 수 있다. 이는 첫 번째 좌표와 두 번째 좌표를 각각  $(-0.5, 0)$ 과  $(0.5, 0)$ 으로 고정하였기 때문이다. 또한, 전체적인 좌푯값이 첫 번째 및 두 번째 좌표에 영향을 받아 형상의 크기가 크게 작아졌음을 확인할 수 있다. 이는 북스테인 자료에 대한 군집별 평균 형상 그림과 이에 대한 형상점 위치를 표기한 Figure 7에서도 확인할 수 있다.

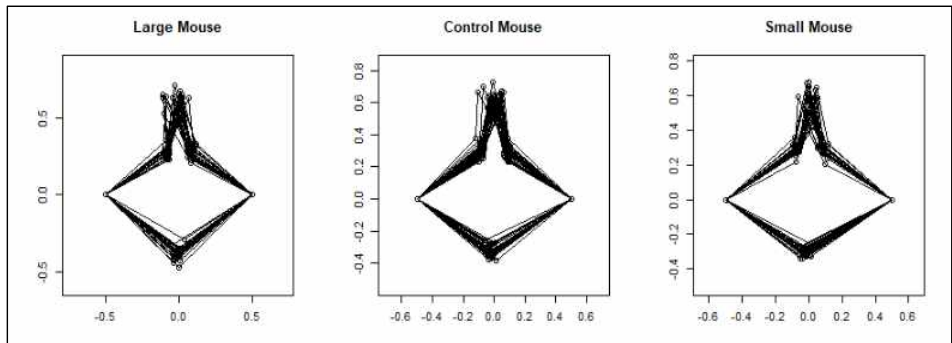


Figure 6. three shape diagrams of Bookstein data

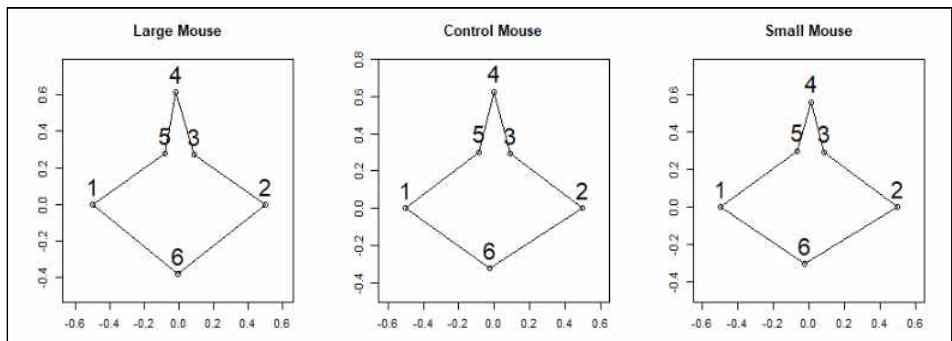


Figure 7. mean shapes of tree cluster in the Bookstein data

Table 3는 북스테인 자료의 군집별 좌푯값을 요약한 것으로,  $x_i$ 와  $y_i$ 는 각각  $i$ 번째 좌표의  $x$ 와  $y$  값을 의미한다. 전체적으로 좌푯값들의 편차가 앞의 두 자료에 비해 크게 감소하였음을 확인할 수 있다. 첫 번째 좌표와 두 번째 좌표가 각각  $(-0.5, 0)$ ,  $(0.5, 0)$ 으로 고정되었기 때문에  $x_1$ 과  $x_2$ ,  $y_1$ 과  $y_2$ 의 편차는 존재하지 않는다. 또한, 이러한 고정을 통해 조절 군집의 좌푯값이 큰 쥐 군집과 작은 쥐 군집과 비슷한 값을 형성하게 되었다. 북스테인 자료의 좌푯값을 변수로 채택한 세 가지 분류모형 결과를 보면 랜덤 포레스트에서는 26.32%, 로지스틱 회귀분석에서는 19.23%, SVM에서는 25.00%의 오분류율을 기록하였다. 앞의 두 자료보다 오분류율이 급격히 높아지는 결과가 나타났고 이는 회전모수가 북스테인 적합과정으로 인해 사라지게 되면서 발생한 결과로 확인할 수 있다.

Table 3. Summary of coordinates values by cluster (Bookstein data)

| Large Mouse (n=23)   |       |       |       |       |       |       |       |       |       |       |       |       |
|----------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
|                      | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ | $y_6$ |
| Avg                  | -0.5  | 0.5   | 0.08  | -0.02 | -0.08 | -0.01 | 0     | 0     | 0.27  | 0.62  | 0.28  | -0.38 |
| SD                   | 0     | 0     | 0.02  | 0.05  | 0.01  | 0.02  | 0     | 0     | 0.03  | 0.05  | 0.03  | 0.04  |
| Control Mouse (n=30) |       |       |       |       |       |       |       |       |       |       |       |       |
|                      | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ | $y_6$ |
| Avg                  | -0.5  | 0.5   | 0.09  | 0.00  | -0.09 | -0.02 | 0     | 0     | 0.29  | 0.63  | 0.30  | -0.32 |
| SD                   | 0     | 0     | 0.02  | 0.04  | 0.02  | 0.02  | 0     | 0     | 0.04  | 0.04  | 0.04  | 0.03  |
| Small Mouse (n=23)   |       |       |       |       |       |       |       |       |       |       |       |       |
|                      | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ | $y_6$ |
| Avg                  | -0.5  | 0.5   | 0.08  | 0.01  | -0.07 | -0.03 | 0     | 0     | 0.29  | 0.56  | 0.30  | -0.30 |
| SD                   | 0     | 0     | 0.02  | 0.03  | 0.01  | 0.02  | 0     | 0     | 0.03  | 0.08  | 0.03  | 0.02  |

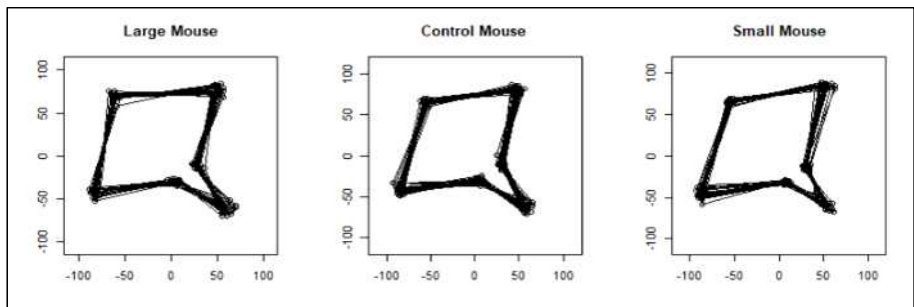


Figure 8. three shape diagrams of GPA data

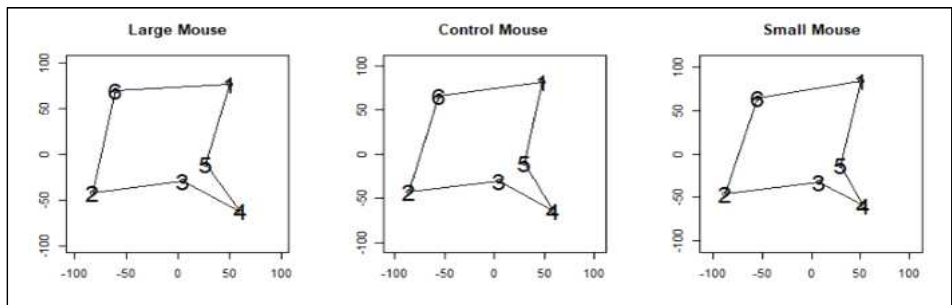


Figure 9. mean shapes of tree cluster in the GPA data

Table 4. Summary of shape variables by cluster (GPA data)

| Large Mouse (n=23)   |       |       |        |        |        |
|----------------------|-------|-------|--------|--------|--------|
|                      | rho   | csize | PC1    | PC2    | PC3    |
| Avg                  | 0.072 | 194.3 | 4.794  | 5.992  | 0.155  |
| SD                   | 0.026 | 10.1  | 5.805  | 6.849  | 5.434  |
| Control Mouse (n=30) |       |       |        |        |        |
|                      | rho   | csize | PC1    | PC2    | PC3    |
| Avg                  | 0.058 | 192.1 | 2.104  | -2.301 | 0.663  |
| SD                   | 0.020 | 5.7   | 3.953  | 5.702  | 6.103  |
| Small Mouse (n=23)   |       |       |        |        |        |
|                      | rho   | csize | PC1    | PC2    | PC3    |
| Avg                  | 0.075 | 173.8 | -7.537 | -2.990 | -1.020 |
| SD                   | 0.032 | 6.0   | 9.032  | 5.578  | 2.920  |

지금까지는 좌푯값의 변수화를 활용한 분류가 진행되었다. 이번에는 GPA 적합을 통해 생성되는 형상 특성변수들을 통해 분류 과정을 진행할 것이다. GPA 적합은 앞의 북스테인과 비슷하게 위치, 척도 및 회전에 대하여 적합을 시행한 후, 이에 대한 리만 거리 및 중심크기, 그리고 형상 PC 점수들을 추출한다. Figure 8은 GPA 적합 자료에 대한 세 가지 군집의 형상 그림이며, 북스테인 그림과 마찬가지로 세 군집들의 형상의 방향이 비슷하게 변환되었음을 확인할 수 있다. 이는 GPA 적합 자료에 대한 군집별 평균 형상 그림과 이에 대한 형상점 위치를 표기한 Figure 9에서도 확인할 수 있다.

Table 4는 GPA 적합 자료의 군집별 형상 특성변수들을 요약한 것으로, rho, csize, PC1~PC3은 각각 리만거리, 중심크기, 제1~3 형상주성분을 뜻한다. 전체적으로 리만거리의 경우 군집별 큰 차이는 없는 것으로 확인되며, 주로 형상주성분에서 차이가 보이는 것으로 확인할 수 있다.

Figure 10은 GPA 적합 자료에서 제1~제3 형상주성분의 역할을 보여주는 그림이다. 각 행은 제1~제3 형상주성분에 대하여 표준 주성분점수의 값  $c$ 를 -3, 0, 3으로 설정하여 얻어지는 전체 데이터에 대한 척추 뼈 형상이다. 제1 형상주성분과 제2 형상주성분, 제3 형상주성분은 각각 36.4%, 28.3%, 14.4%의 설명비율을 차지하고 있으며, 각 형상주성분의 변화에 따른 전체적인 형상모양의 변동을 확인할 수 있다. 제1 형상주성분에서는 증가 혹은 감소에 따라서 4번째 좌푯점의 변화가 가장 크게 나타났고, 제2 형상주성분에서는 1번째와 3번째, 그리고 5번째 좌푯점의 변화가 크게 나타났다. 마지막으로 제3 형상주성분에서는 3번째 좌푯점과 4번째 좌푯점의 변화가 가장 크게 나타났다.

GPA 적합 자료의 형상 특성변수를 변수로 채택한 세 가지 분류모형 결과를 보면 랜덤 포레스트에서는 23.68%, 로지스틱 회귀분석에서는 23.08%, SVM에서는 22.37%의 오분류율을 기록하였다.

Table 5는 자료별 분류모형 결과에 대한 오분류율을 정리한 표이다. 전체적으로 결과를 정리해 보면, 중심화 자료의 변수화가 가장 낮은 오분류율을 보였다. 또한, 원자료의 변수화도 10% 미만의 오분류율을 보여 좋은 분류결과를 보였다. 반면, 북스테인 자료의 변수화와 GPA 적합자료의 형상 특성변수는 상대적으로 좋지 못한 분류율을 보였다. 이에 대해서는 적합과정으로 인해  $C_2$ 의 가장 큰 특성이었던 방향이 회전 모수의 적합과정으로 인하여 다른 군집들과 비슷하게 변환되어 군집 간의 형상적 차이가 상대적으로 상쇄되었다고 판단할 수 있다. 분석 기법끼리 비교해보면, 로지스틱 회귀분석이 가장 작은 편차를 보였으며, SVM과 랜덤 포레스트는 비슷한 결과를 보였다고 판단할 수 있다.

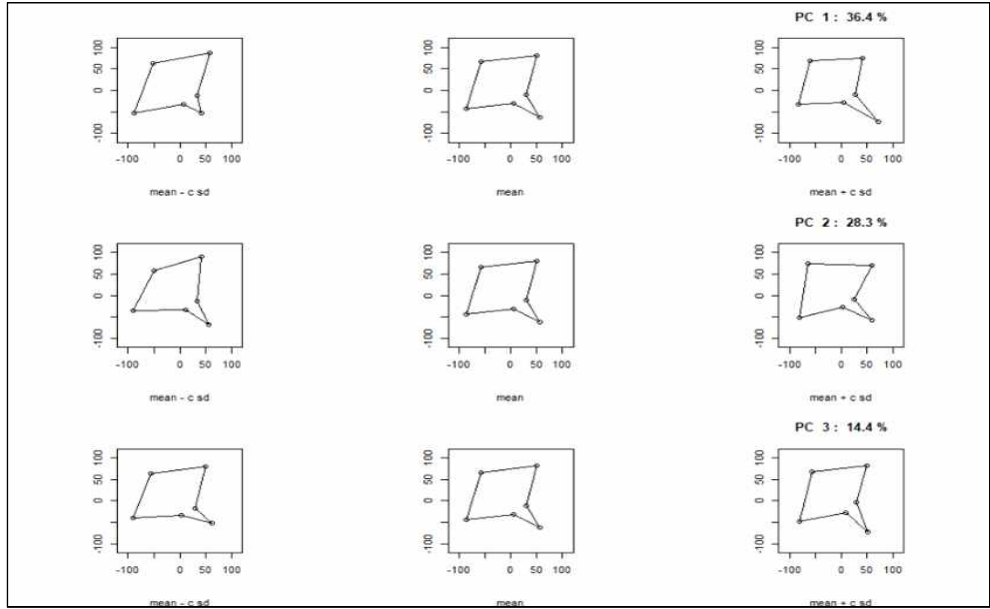


Figure 10. Characteristics of shape variation by 1~3 principal components

Table 5. Result of classification by data types

| content        |                | Random Forest | Logistic Regression | SVM    |
|----------------|----------------|---------------|---------------------|--------|
| coordinates    | raw data       | 3.95%         | 6.41%               | 5.26%  |
|                | centered data  | 2.63%         | 7.70%               | 2.63%  |
|                | Bookstein data | 26.32%        | 19.23%              | 25.00% |
| shape variable | GPA data       | 23.68%        | 23.08%              | 22.37% |

#### 4. 결론

지금까지 배열로 구성된 형상점 자료들을 분류모형에 사용하기 적합한 행렬 형태로 변환하기 위하여 GPA 적합 및 좌푯값의 변수화를 사용하였다. GPA 적합의 경우, 각 형상 표본들에 대하여 위치, 크기 및 방향에 대하여 적합하는 과정을 진행한 후 리만거리, 중심크기 및 형상 PCA를 통한 형상 PC 점수를 변수로 채택하였다. 그리고 좌푯값의 변수화는 2차원 기준으로  $n$ 개의 좌표가 있다고 가정할 시  $x_1, \dots, x_n, y_1, \dots, y_n$  등의 좌푯값들을 변수로 채택하였다. 그리고 좌푯값 데이터들에 대하여 각각 원자료, 중심화 자료, 그리고 2차원 자료 한정으로 북스테인 적합 자료를 채택하여 실제 자료에 대하여 분류 과정을 실시하였다. 위의 GPA 적합을 활용한 방식과 좌푯값의 변수화에서는 단순히 변수 형식에 대한 차이만 있는 것이 아닌, 형상의 적합 실시 여부와도 차이가 존재한다. GPA 적합을 통해 얻어지는 형상 특성변수나 북스테인 적합을 통한 좌푯값의 변수화의 경우, 각각의 적합과정을 통해 표본들에 대하여 위치, 척도 및 회전이 비슷하게 조정이 되는 과정을 거쳤고, 중심화 자료의 경우 각 표본들의 형상 중심점을 (0, 0)으로 통일시켜 위치에 대한 적합을 진행하였다. 반면, 원자료의 경우 아무런 위치, 척도 및 회전에 대한 적합과정이 이루어지지 않았다.

적은 형상점을 보유한 2차원 mice 자료를 활용한 사례분석에서는 회전 및 크기에 대한 적합과정이 없었던 원자료 및 중심화 자료의 좌푯값의 변수화가 좋은 결과를 나타냈다. 반면, 북스테인

적합 자료의 좌푯값의 변수화와 GPA 적합을 통한 형상 특성변수의 경우 상대적으로 좋지 못한 분류율을 보였는데 이는 회전 및 척도에 대한 적합으로 인해 군집 간의 특성이 일부 상쇄되었다고 판단할 수 있었다. 한편 형상점의 수가 많은 3차원 자료에 대해선 의미있는 결과를 얻지 못하여 본 연구에 포함하지 않았다.

## References

- Bookstein, F. L. (1986). Size and shape spaces for landmark data in two dimensions (with discussion), *Statistical Science*, 1(2), 181-242. DOI: 10.1214/ss/1177013696
- Choi, Y. S. (2018). *Multivariate Data Analysis with R*, Kyungmoon.(in Korean).
- Choi, Y. S. (2021). *Multivariate Statistical Shape Analysis with R*, Pusan National University Press.(in Korean).
- Chun, H. (2023). A study on the sufficiency of old-age measures for reverse mortgage users, *Journal of the Korean Data Analysis Society*, 25(4), 1285-1297.(in Korean). DOI : 10.37727/jkdas.2023.25.4.1285
- Dryden, I. L., Mardia, K. V. (2016). *Statistical Shape Analysis, with applications in R*, John Wiley & Sons. DOI:10.1002/9781119072492
- Kim T. H., Jeong, H. J. Song, J. Y. Kim, N. N., Lee, E. M. (2023). Analysis of influencing factors of suicide ideation using random forest model : Focusing on the national health and nutrition examination survey, *Journal of The Korean Data Analysis Society*, 1121-1132. (in Korean). DOI: <https://doi.org/10.37727/jkdas.2023.25.3.1121>
- Ko, S., Kim, I. (2024). Prediction of netflix content viewing time using a machine learning classification mode, *Journal of the Korean Data Analysis Society*, 26(5), 1357-1373. DOI : 10.37727/jkdas.2024.26.5.1357
- Lee, K., Yeo, D. (2024). Support vector machines with reject option for unequal costs and imbalanced dataset, *Journal of the Korean Data Analysis Society*, 26(6), 1689-1705. DOI : 10.37727/jkdas.2024.26.6.1689

Classification models using the conversion of shape landmark data<sup>\*</sup>*Jun Hong Kim<sup>1</sup>, Changwan Kang<sup>2</sup>, Yong-Seok Choi<sup>3</sup>*

## Abstract

Dryden, Mardia (2016) define shape analysis as an analysis that measures and describes the shapes of objects represented by landmarks in geometric space and compares them. The shape point data used for shape analysis has the form of an array, which is a form that is difficult to apply to the shape classification model. This study presents two methods for applying shape point data to the classification model. The first method is to apply the Generalized Procrustes Analysis (GPA) fit and then obtain Riemannian distance, centroid size, and shape PC scores to select them as variables, and the second method is to vectorize the configuration matrix of each sample to use all coordinate values as variables. The data obtained by these two methods are put into three classification models: Random Forest, Logistic Regression Analysis, and SVM to compare the misclassification rates for each data. To this end, the method above was applied to the mice data from Dryden, Mardia (2016), and the results were examined. The comparison revealed that the mice data of three clusters, which is a 2-dimensional data set with a small number of shape points, showed that the rotation parameter had a significant impact on cluster determination. Additionally, vectorization of the original data or centered data, which had not undergone the fitting process, yielded better results.

**Keywords :** Shape analysis, GPA fit, Configuration matrix, Random Forest, Logistic regression, SVM.

---

<sup>\*</sup>This Work was based on the master's thesis of first author(Jun Hong Kim).

<sup>1</sup>Graduate student, Dept. of Statistics, Pusan National University, 2 Busandaehak-ro, 63beon-gil, Geumjeong-Gu, Busan, 46241, Korea. E-mail: wnsghd4685@naver.com

<sup>2</sup>Professor, Dept. of Industrial Management & Big Data Engineering, Dongeui University, Busan, 614-714, Korea. E-mail: cwkang@deu.ac.kr

<sup>3</sup>(Corresponding author) Professor, Dept. of Statistics, Pusan National University, 2 Busandaehak-ro, 63beon-gil, Geumjeong-Gu, Busan, 46241, Korea. E-mail: yschoi@pusan.ac.kr

[Received 25 February 2025; Revised 20 March 2025; Accepted 23 March 2025]