The Korean Journal of Applied Statistics
2024, Vol. 37, No. 1, 73–86

DOI: https://doi.org/10.5351/KJAS.2024.37.1.073
Print ISSN1225-066X / Online ISSN 2383-5818

Classification and discrimination of excel radial charts using the statistical shape analysis

Seungeon Lee[a], Jun Hong Kim[a], Yeonseok Choi[a], Yong-Seok Choi[1,a]

[a]Department of Statistics, Pusan National University

Abstract

A radial chart of Excel is very useful graphical method in delivering information for numerical data. However, it is not easy to discriminate or classify many individuals. In this case, after shaping each individual of a radial chart, we need to apply shape analysis. For a radial chart, since landmarks for shaping are formed as many as the number of variables representing the characteristics of the object, we consider a shape that connects them to a line. If the shape becomes complicated due to the large number of variables, it is difficult to easily grasp even if visualized using a radial chart. Principal component analysis (PCA) is performed on variables to create a visually effective shape. The classification table and classification rate are checked by applying the techniques of traditional discriminant analysis, support vector machine (SVM), and artificial neural network (ANN), before and after principal component analysis. In addition, the difference in discrimination between the two coordinates of generalized procrustes analysis (GPA) coordinates and Bookstein coordinates is compared. Bookstein coordinates are obtained by converting the position, rotation, and scale of the shape around the base landmarks, and show higher rate than GPA coordinates for the classification rate.

Keywords: shape analysis, radial chart, GPA, bookstein coordinates

1. 서론

맛과 영양의 균형과 같은 평가지표나 스포츠 선수들의 능력치에 대한 자료는 수치로 나타난 자료의 형태보다 시각적으로 표현하는 것이 정보 전달에 효과적이다. 수치형 자료의 경우 개체들의 차이를 직관적으로 파악하기는 어려우므로 여러 항목에 대한 평가 점수를 엑셀(Excel)의 방사형 차트의 형태로 나타내곤 한다. 차트로 표현하면 각 개체의 특성을 한눈에 파악하기는 좋지만 비교하고자하는 변수 또는 관측치의 개수가 많은 경우 이들의 차이를 구별하기 어렵다. 보다 쉽게 개체를 분류하고 특성을 알아내기 위해 자료를 각각 하나의 형상으로 두고, 형상의 정보를 대표할 수 있는 형상점을 찾아 형상분석을 이용하고자 한다. 형상분석은 의학, 생물학, 정보기술, 유전학, 인류학, 고고학 등 다양한 분야에서 응용되며 이미지분석, 자동인식 등의 기술에서 활용되어질 수 있는 아주 유용한 분석으로 알려져 있다. Dryden과 Mardia (2016)에 의하면 형상분석은 기하적 공간상에서 형상점에 의해 나타낸 개체들의 형상을 측정하고 기술하며 비교하려는 분석 방법이다. 본 연구는 수치형 자료를 방사형 차트로 표현한 것에 대해 이러한 형상분석을 적용하여 형상의 분류 및 판별을 수행하고자 하는 것이 목적이다. 형상분석을 진행하기 위해서는 각 개체의 형상을 잘 대표하는 형상점을 찾고, 그 형상점에 대한 형상좌표를 추출해내야 한다. 방사형 차트는 원점을 기준으로 동일한 간격에 따라 각 개체의

[1]Corresponding author: Department of Statistics, Pusan National University, 2 Busandaehak-ro, 63 beon-gil, Geumjeong-Gu, Busan 46241, Korea. E-mail:yschoi@pusan.ac.kr

Published 29 February 2024 / journal homepage: http://kjas.or.kr

Table 1: Some data of tastes and noses of 86 whiskeys

Distillery	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	C
Aberfeldy	2	2	2	0	0	2	1	2	2	2	2	2	C_1
Aberlour	3	3	1	0	0	4	3	2	2	3	3	2	C_2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Tullibardine	2	3	0	0	1	0	2	1	1	2	2	1	C_1

(a) (b)

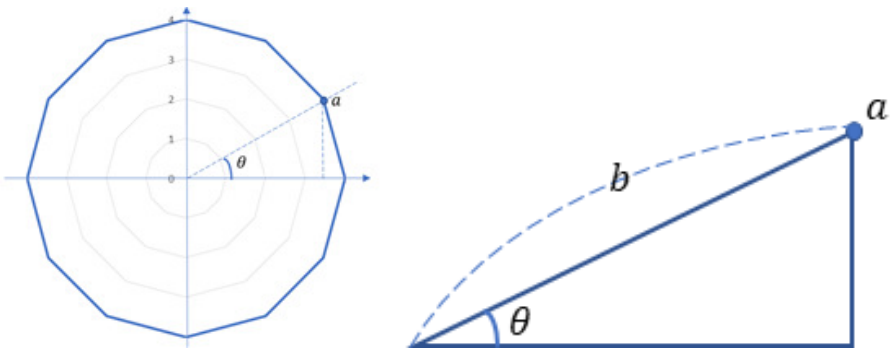
Figure 1: *Excel radial charts of the whiskeys.*

특성이 나타나있다. 이러한 특징을 이용하여 형상좌표를 찾기 위한 방법을 찾고, 형상의 판별을 진행하기 위한 북스테인좌표와 GPA (generalized procrustes analysis)적합 좌표로의 변환 과정을 알아보고자 한다. 변환한 자료를 이용하여 전통적 판별분석과 서포트벡터머신 기법을 적용해 방사형 차트로 표현된 형상을 분류하고 판별하는 방법을 소개할 것이다. 더 나아가 만약 방사형 차트의 각이 많을 경우, 즉 자료의 변수 개수가 많을 경우 형상이 시각적으로 효과적이지 않으므로 주성분 분석(PCA)을 이용하는 것을 제안하고자 한다. 2장에서는 수치로 표현되어있는 자료를 엑셀의 방사형 차트를 활용하여 시각적으로 나타내고 형상분석을 적용해 좌표화하는 방법과 판별에 사용할 기법들을 소개하고자 한다. 3장에서는 2장에서 소개한 방법을 활용한 방사형 차트 분류와 판별의 활용 사례를 보이고 결과를 비교해보고자 한다. 마지막으로 4장에서는 본 논문의 결론을 정리하려 한다.

2. 수치형 자료의 형상분석

2.1. 수치형 자료의 엑셀 방사형 차트

제 3장 활용사례의 Table 1은 스코틀랜드에 위치한 86개 증류소의 위스키들에 대해 12가지 맛과 향의 정도가 수치로 나타나있는 자료이다(https://www.kaggle.com/datasets/ koki25ando/scotch-whisky-dataset). 해당 맛과 향이 약하면 0, 강하면 4의 점수로 표기되어있다. 이런 자료의 경우 변수의 개수가 많아 해당 개체에 대한 특성을 파악하기 힘들고 또 다른 개체와의 비교가 쉽지 않다. 더 나아가 더 많은 변수를 가진 자료일 경우와 많은 개체끼리 비교가 필요한 경우, 수치 자료 그대로를 이용하는 것보다 앞에서 설명한 엑셀의 방사형 차트를 활용하여 도형 형태로 표현하는 것이 자료를 파악하기 쉬울 것이다. 그리고 Table 1에 나타난 수치 자료를 엑셀의 방사형 차트 기법을 이용하여 도형의 형태로 나타낸 것을 Figure 1에서 확인할 수 있다. 특히 Table 1의



(a) (b)

Figure 2: *A regular dodecagon represented by excel radial chart and a right triangle.*

Aberfeldy와 Aberlour 두 개체에 대해 비교할 수 있도록 변수 위치를 동일한 위치에 설정하고 동일한 간격을 갖는 차트의 형태를 만들었다. 각각의 도형을 살펴보았을 때 각 변수에 대한 정도를 한 눈에 쉽게 파악할 수 있으며 Figure 1(a)와 (b)의 도형에서 각 변수 별로 어떤 개체가 더 높고 낮은지에 대한 정도를 비교하여 특징을 구분하기에도 쉬워졌다고 볼 수 있다.

2.2. 방사형 차트의 형상화

엑셀에서 수치형 자료를 방사형 차트로 그렸을 때 각 개체의 특성을 파악하고 개체 별 특성을 비교하기에 용이하지만 시각적으로 표현하더라도 개체 수가 많을 경우는 군집화를 시도하거나 알려진 군집에 대해 판별을 해내기는 어렵기 때문에 형상분석을 활용하여 이를 실행하고자 한다. 형상분석을 적용하기 위해서는 방사형 차트로 나타낸 것을 형상으로 두고 형상점과 그에 대한 형상좌표를 추출해야 한다.

도형으로 형상화 된 자료에 대하여 형상좌표를 구하기 위해서 삼각함수 sin과 cos을 이용한다. Figure 2의 (a)는 Figure 1에서 12개의 변수에 대해 나타낸 것처럼 임의로 표현한 정12각형이며 원점을 기준으로 축과 축을 그어 기준선을 만들어준다. 12개의 꼭짓점 중 하나를 a라고 지정하고, 원점에서부터 형상점 까지 이은 선분과 수평축이 이루는 각을 θ라고 표현한다. 형상점 a를 수평축에 수직이 되도록 선분을 그어 Figure 2의 (b)와 같은 직각삼각형의 형태를 만든다. 원점에서부터 형상점 까지 이은 선분의 길이는 b로 둔다. Figure 2(b)는 12각형의 일부를 나타낸 직각삼각형으로, 형상점 a에 대한 형상좌표를 (x, y)로 두고 식 (2.1)의 피타고라스 정리를 이용하면 다음과 같은 (2.2) 식이 성립한다.

$$\sqrt{x^2 + y^2} = b. \tag{2.1}$$

$$x = b \times \cos\theta, \quad y = b \times \sin\theta. \tag{2.2}$$

2.3. 형상분석을 위한 좌표화

Kendall (1977)의 정의에 의하면 형상은 개체들의 위치, 척도, 회전에 의한 효과들이 제거되었을 때 남아있는 모든 기하적인 정보를 의미한다. 따라서 형상에 대한 좌표체계는 형상점들의 집합에 대해 이동, 척도, 회전의 효과를 가하여도 변하지 않아야 한다. 이를 위해 2차원 평면에 대해 형상점1과 2를 고정된 위치 $(0, 0)$과 $(1, 0)$

또는 $(-1/2, 0)$과 $(1/2, 0)$으로 변환하도록 척도, 이동, 회전을 고려한 후 남은 나머지 좌표를 구하는 북스테인좌표를 알아보고자 한다. 또한 여러 개체의 형태의 비교를 위해 최소제곱법에 기반을 둔 프로크러스티즈 분석으로부터 얻어지는 일반화 프로크러스티즈 좌표를 소개하려 한다. 추가적으로 이들과 상관이 높은 중심크기와 리만거리도 포함하여 형상분석을 활용하기 위한 좌표체계를 Choi (2021)와 같이 생성할 것이다.

2.3.1. 북스테인좌표

형상에 대한 좌표를 확인하면 원 자료의 변수 개수만큼 형상좌표가 생성된다. k (≥ 3)개의 형상점이 존재한다고 가정할 때 형상좌표 (x_j, y_j), $j = 1, \ldots, k$ 가 존재한다. 이때 두 기저 형상점 사이의 제곱 유클리드거리를 식 (2.3)과 같이 정의한다.

$$D_{12}^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2 > 0. \tag{2.3}$$

특히, 식 (2.4)와 (2.5)의 북스테인좌표 $\mathbf{u} = (u_j^B, v_j^B)^T$는 j 번째 형상점 좌표를 $\mathbf{x} = (x_j, y_j)^T$라 하면 변환 $\mathbf{u} = \beta\Gamma(\mathbf{x} - \alpha)$을 통해 얻어진다. 여기서 $\beta > 0$ 는 크기이고 Γ는 시계 방향으로 각 θ만큼 회전시킬 회전행렬이며 $\alpha = (\alpha_1, \alpha_2)^T$는 이동벡터이다. 그리고 기저형상점으로 첫 번째와 두 번째 형상점 $(-1/2, 0)$과 $(1/2, 0)$로 변환하도록 $(\beta, \theta, \alpha_1, \alpha_2)$를 구하고 이들을 $\mathbf{x} = (x_j, y_j)^T$, $j = 3, \ldots, k$와 함께 변환 $\mathbf{u} = \beta\Gamma(\mathbf{x} - \alpha)$에 대입하면 기저가 되는 형상점을 제외한 북스테인좌표 $(u_j^B, v_j^B)^T$, $j = 3, \ldots, k$가 얻어진다.

$$u_j^B = \frac{\left[(x_2 - x_1)\left(x_j - x_1\right) + (y_2 - y_1)\left(y_j - y_1\right)\right]}{D_{12}^2} - \frac{1}{2}. \tag{2.4}$$

$$v_j^B = \frac{\left[(x_2 - x_1)\left(y_j - y_1\right) + (y_2 - y_1)\left(x_j - x_1\right)\right]}{D_{12}^2}. \tag{2.5}$$

일반적으로 k개 형상점을 가진 n개 개체에서 얻어지는 북스테인좌표 $u_i^B = (u_{i3}^B, \ldots, u_{ik}^B, v_{i3}^B, \ldots, v_{ik}^B)^T$, $i = 1, \ldots, n$의 각 형상점에 대한 산술평균은 식 (2.6)과 같고, 이들에 의한 전체 개체들의 평균형상 좌표는 식 (2.7)과 같다.

$$\overline{u_j^B} = \frac{1}{n}\sum_{i=1}^{n} u_{ij}^B, \quad \overline{v_j^B} = \frac{1}{n}\sum_{i=1}^{n} v_{ij}^B, \quad j = 3, \ldots, k. \tag{2.6}$$

$$\overline{u^B} = \left(\overline{u_3^B}, \ldots, \overline{u_k^B}, \overline{v_3^B}, \ldots, \overline{v_k^B}\right)^T. \tag{2.7}$$

이를 활용하여 개체들의 군집 별 평균형상을 2차원에 표현하여 각 군집의 특성을 비교할 수 있다. 더 나아가 알려진 군집에 대한 형상의 특징에 따라 각 개체들이 잘 분류 및 판별이 되는지 확인하기 위해 북스테인좌표를 이용해 판별분석을 진행할 수 있다. 형태를 비교하기 위한 객관적인 크기척도인 형상들의 중심크기는 북스테인좌표와 상관이 높으므로 중심크기를 포함한 새로운 자료를 활용하고자한다. 이 때 형태행렬의 중심크기는 식 (2.8)로 정의된다.

$$S(X) = \|X_c\| = \sqrt{\sum_{i=1}^{k}\sum_{j=1}^{m}\left(x_{ij} - \overline{x_j}\right)^2}, \tag{2.8}$$

여기서, m차원의 데카르트좌표 공간상에서 k개의 형상점들로 이루어지는 $k \times m$의 크기로 이루어진 형태행렬을 X라고 한다. 형태행렬에서 형상의 위치효과를 제거하기 위해 중심화 행렬 $C = I_k - 1_k 1_k^T/k$을 활용하며 이를 적용한 중심화 형태행렬을 $X_c = CX$라고 한다. 이때, C에서 I_k를 k차 항등행렬이라고 하고 1_k를 모든 원소를 1로 가지는 $k \times 1$ 벡터라고 한다. 여기서 $\overline{x_j} = k^{-1} \sum_{i=1}^{k} x_{ij}$는 j 번째 차원의 형상점들에 대한 평균이고, $\|X\| = \sqrt{\mathrm{tr}(X^T X)}$는 행렬 X의 유클리드 놈이다.

2.3.2. GPA (generalized procrustes analysis) 적합좌표

Dryden과 Mardia (2016)에 의하면 앞서 설명한 북스테인좌표는 특정 기준선을 선택할 때 대칭성이 부족하기 때문에 프로크러스티즈 분석을 활용한다. 프로크러스티즈 분석은 두 개체의 형태의 유클리드거리에 따라 가능한 가깝도록 최소제곱법을 활용하여 변환에 의해 유사하도록 하게 하는 매칭 분석기법으로 제안되었으며 이로부터 여러 개체의 형태를 비교하기 위한 일반화 프로크러스티즈 분석(GPA)을 소개하고자 한다. m차원에서 k개의 형상점들에 의해 각각 표현되는 크기가 $k \times m$인 $n\,(\geq 2)$개의 중심화 형태행렬이 있다고 가정한다. 이들 형태행렬을 크기가 $k \times m$인 평균형상의 모집단으로부터 추출한 확률표본이라고 할 때 이로부터 평균형상을 추정하고 각각의 형태행렬에 대한 프로크러스티즈 좌표를 얻으면 이것이 GPA에 의한 적합좌표가 된다. GPA에 의한 적합좌표와 추정된 평균형상은 식 (2.9)과 (2.10)에 나타나있다.

$$\hat{X}_i = \hat{\beta}_i X_i \hat{\Gamma}_i + 1_k \hat{\alpha}_i^T, \quad i = 1, \ldots, n. \tag{2.9}$$

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \hat{X}_i, \tag{2.10}$$

여기서 $\hat{\alpha}_i$는 크기 $m \times 1$의 위치벡터 추정치이고 $\hat{\beta}_i > 0$는 척도모수 추정치, $\hat{\Gamma}_i$는 $m \times m$ 크기의 특수 직교회전행렬, 1_k는 모든 원소를 1로 가지는 $k \times 1$ 벡터이다.

형상좌표에 다중공선성이 존재하는 경우 형상의 형태행렬자료에 PCA를 적용하여 활용하고자한다. Dryden과 Mardia (2016)는 GPA에서 PCA를 이용한 형상변동에 대해 설명하고 있다. Choi (2021)에 의하면 형상분석에서는 GPA 적합에 의해 추정된 평균형상을 구한 뒤 프로크러스티즈잔차에 대한 PCA를 통해 형상변동을 살펴본다. 주성분은 원 변수의 수 p만큼 생성되며, 전체 p개 주성분의 총 분산에 대한 비율인 설명력이 약 70%를 넘게 되는 주성분의 개수만큼 택하면 합리적이다. GPA 적합좌표 각각의 크기가 $k \times m$인 $\hat{X}_1, \ldots, \hat{X}_n$이 구해졌을 때 합좌표의 형상과 추정된 평균형상 간 차이인 프로크러스티즈 잔차는 식 (2.11)에 나타나있다. 이에 PCA를 적용하여 형상변동을 평가할 수 있으며 GPA 적합좌표에 대한 크기가 $km \times km$인 공분산행렬의 고유값-고유벡터를 고려하기로 한다. 공분산행렬은 식 (2.12)에 나타나있으며 여기서 $\overline{X} = n^{-1} \sum_{i=1}^{n} \hat{X}_i$는 추정된 평균형상과 같으며, vec 기호는 행렬을 벡터화하는 것으로 $k \times m$이니 $\hat{X}_i$를 $km \times 1$ 크기의 벡터로 변환한다.

$$R_i = \hat{X}_i - \hat{\mu}, \quad i = 1, \ldots, n. \tag{2.11}$$

$$S = \frac{1}{n-1} \sum_{i=1}^{n} \mathrm{vec}\left(\hat{X}_i - \overline{X}\right) \mathrm{vec}\left(\hat{X}_i - \overline{X}\right)^T. \tag{2.12}$$

공분산행렬 S의 고유값을 크기순으로 $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$라 하고 이에 대응하는 고유벡터를 각각 $\gamma_j, j = 1, \ldots, p$라 한다. 이는 j 번째 주성분을 제공하고 이들의 고유벡터는 $\gamma_j^T \gamma_j = 1$이고 $\gamma_i^T \gamma_j = 0, i \neq j$을 만족한다. 여기서 $p = \mathbf{min}\,(n-1, M)$이고 $M = 2k - 4$는 형상공간의 차원을 나타낸다. 따라서 형상분석에서 j 번째 주성분에 대한 i 번째 개체의 주성분점수는 식 (2.13)과 같다.

$$z_{ij} = \gamma_j^T \mathrm{vec}\left(\hat{X}_i - \overline{X}\right), \quad i = 1, \ldots, n, \;\; j = 1, \ldots, p. \tag{2.13}$$

일반적으로 다변량 자료분석에서 PCA에 의해 차원이 축소된 새로운 변수인 주성분에 의한 주성분점수를 새로운 자료로 활용하듯이 형상분석에서도 이러한 주성분 점수를 이용하여 형상의 중심크기와 리만거리와 함께 행렬의 형태로 만들어 새로운 자료로 활용하고자 한다. 함께 사용되는 리만거리에 대해 살펴보겠다. 중심화 형태행렬에 척도효과를 제거한 것을 $(k-1)\times 2$인 사전형상 Z_1과 Z_2라고 할 때 Z_1과 Z_2는 사전형상공간에 속하는 구의 형상공간의 형상 $[X_1]$과 $[X_2]$에 존재하고 이를 사전형상구라고 한다. 크기가 $k\times 2$인 두 형태행렬 X_1과 X_2의 사전형상구에서 사전형상 Z_1과 Z_2간의 가장 근접하게 큰 원을 리만거리로 나타내고 식 (2.14)를 만족한다. 여기서 비정칙값 $\lambda_k, k = 1,2$는 비정칙값분해 (2.15)를 만족한다. 여기서 $\Lambda = \mathrm{diag}(\lambda_1,\ldots,\lambda_m)$는 크기가 $m\times m$인 대각행렬이며, 대각원소에 해당하는 비정칙값 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{m-1} \geq |\lambda_m|$를 만족하며 U와 V는 직교행렬이다.

$$\rho(X_1, X_2) = \cos^{-1}\left(\sum_{k=1}^{m} \lambda_k\right). \tag{2.14}$$

$$Z_2^T Z_1 = U\Lambda V^T. \tag{2.15}$$

2.4. 형상의 판별분석

2.3절에서 소개한 형상좌표를 사용하여 판별분석을 시도하고 오분류율을 통해 형상의 분류 정도를 평가하려고 한다. 먼저 다변량 정규성을 확인하여 이를 만족하는 경우 분산의 동질성을 확인하고 이에 따라 선형판별함수와 이차판별함수를 활용해 전통적 판별분석을 실시하는 방법을 소개하고자 한다. 또한 다변량 정규성과 등분산성의 가정과 무관하게 자료 분포에 대한 정보를 몰라도 적용이 가능한 머신러닝과 딥러닝의 기법들 중 대표적으로 사용되는 서포트벡터머신과 인공신경망 방법을 하나씩 추가로 소개할 것이다.

2.4.1. 전통적 판별분석

Choi (2018)에 의하면 전통적 판별분석은 이미 알려진 개체들의 군집을 잘 분류해주는 판별기준으로 개체들의 다양한 정보를 측정한 변수들의 선형결합 또는 비선형결합에 의한 함수를 찾는 것이다. 군집 자료의 다변량 정규성과 분산의 동질성에 대한 가정을 확인하여 자료의 분포에 대한 정보를 먼저 파악한다. 가정에 따라 다변량 정규성을 따르고 분산이 동일할 경우 선형판별함수를, 다변량 정규성을 따르고 분산이 동일하지 않은 경우 이차판별함수를 활용하여 분류규칙에 따라 판별분석을 진행한다.

p개 확률변수에 의한 확률벡터 $x = (x_1,\ldots,x_p)^T$가 다변량 정규분포를 따르는 g개의 군집 $C_1,\ldots,C_g$을 가진다고 할 때, 관찰된 $x_o = (x_{o1},\ldots,x_{op})^T$에 대해 판별점수가 아래의 식으로 표현된다. 각 군집의 분산이 동일하지 않으면 식 (2.16)을, 분산이 동일하면 식 (2.17)을 활용하여 판별점수를 구한다.

$$\hat{Q}_j(x) = -\frac{1}{2}\ln|S_j| - \frac{1}{2}\left(x-\overline{x_j}\right)^T S_j^{-1}\left(x-\overline{x_j}\right) + \ln p_j. \tag{2.16}$$

$$\hat{L}_j(x) = \overline{x_j}S_p^{-1}x - \frac{1}{2}\overline{x_j}^T S_p^{-1}\overline{x_j} + \ln p_j, \quad j = 1,\ldots,g. \tag{2.17}$$

이 때, $\overline{x_j}$와 S_j는 j 번째 그룹의 평균벡터와 공분산행렬이고, S_p는 합동공분산행렬과 p_j는 사전확률을 나타낸다. 조건에 따라 식 (2.18)과 식 (2.19)를 적용시켜 선형판별분석 또는 이차판별분석을 진행하게 되며 이들 식을 만족하는 x_0는 군집 C_k에 할당된다.

$$\hat{Q}_k(x_0) = \mathbf{max}\left\{\hat{Q}_1(x_0), \ldots, \hat{Q}_g(x_0)\right\}. \tag{2.18}$$

$$\hat{L}_k(x_0) = \mathbf{max}\left\{\hat{L}_1(x_0), \ldots, \hat{L}_g(x_0)\right\}. \tag{2.19}$$

2.4.2. 서포트벡터머신

전통적 판별분석에서 형상의 군집이 불분명하게 구분이 잘 되지 않는 경우가 많다. 이때 머신러닝 알고리즘은 자료의 분포에 대한 정보를 몰라도 적용가능하며 그 중 서포트벡터머신은 작거나 중간크기의 데이터 셋에 적합하며 전통적 판별분석과 로지스틱 회귀분석보다 예측력이 높은 것으로 알려져 있다. 서포트벡터머신은 Izenman (2008)에 의하면 기본적으로 두 개의 그룹을 분리하는 것으로 두 그룹 사이의 최적의 분리 초평면(hyperplane)을 선택하는 방법이다. 2.4.1절에서는 다변량 정규성과 분산의 동질성에 대한 가정을 확인하고 전통적 판별을 진행했지만 이번 절에서는 그러한 가정을 제외하고 선형 서포트벡터머신을 적용해보고자 한다. 서포트벡터머신의 분류함수는 식 (2.21)과 같으며 커널함수를 이용하여 선형으로 분리할 수 없는 점들을 분류하도록 한다. 커널함수로 대표적으로 많이 사용되는 가우스 RBF커널을 적용할 것이며 식 (2.22)에 나타나있다. 식 (2.20)에서 n쌍의 훈련자료의 $(x_1, y_1), \ldots, (x_n, y_n)$이 다음의 집합을 만족한다고 하자.

$$\mathcal{L} = \{(x_i, y_i) : i = 1, \ldots, n\}, \ \text{s.t} \ x_i \in \mathbb{R}^p, \ y_i \in \{-1, 1\}. \tag{2.20}$$

$$\widehat{f}(x) = \hat{\beta}_0 + \sum_{i=1}^{n} \hat{\alpha}_i y_i K(x_i, x). \tag{2.21}$$

$$K(x_i, x_{i'}) = \mathbf{exp}\left(-\gamma |x_i - x_{i'}|^2\right), \tag{2.22}$$

여기서 두 관측값 $x_i, x_{i'} \in \mathbb{R}^p$이고 $\hat{\beta}_0$는 편차인 절편항이다. $\hat{\alpha}_i$는 라그랑지승수를 의미하며 γ는 커널함수에서 사용되는 감마모수이다.

2.4.3. 인공신경망

딥러닝 알고리즘은 자료의 분포에 대한 정보를 몰라도 적용가능하며 그 중 인공신경망은 예측 변수 사이의 가능한 모든 상호작용을 탐지할 수 있고 대규모 데이터 셋을 효율적으로 훈련시킬 수 있다. Eberhart (2014)에 의하면 인공신경망은 인간의 뇌가 작동하는 방식을 모형화 하려는 컴퓨터 시스템으로 인간과 같이 의사결정을 할 수 있는 것처럼 보인다. 인공신경망은 주어진 데이터 x에 대한 뉴런의 출력 $y = g(x; \theta)$를 계산하여 이를 바탕으로 주어진 x가 어떤 군집에 속하는지 판별한다. 이는 식 (2.23)에 의해 계산된다. 여기서 θ는 매개변수, w_j는 가중치, b는 편의를 나타낸다.

$$g(x; \theta) = f\left(\sum_{j=1}^{M} w_j x_j + b\right). \tag{2.23}$$

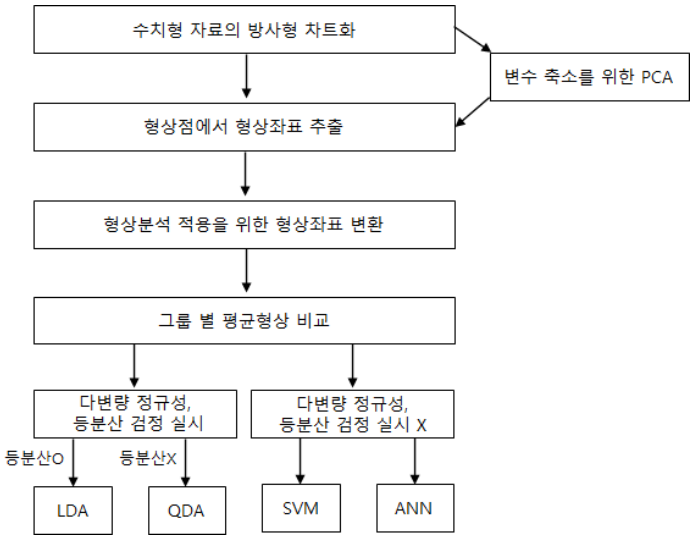


Figure 3: *Discriminant analysis procedure for excel radial chart.*

2.4.4. 모형 평가 방법

자료의 각 군집별 개체 수가 적고 전체 개체 수 또한 적은 편이므로 재대입법(resubstitution method; RSM)과 10-fold cross validation, LOOCV 방법(leave one out cross validation; LOOCV)을 사용하여 비교하기로 한다. 재대입법은 판별함수를 만드는데 사용된 자료를 판별함수의 평가에 그대로 이용하는 방법이다. 이는 표본크기가 큰 경우에도 최적오류율이 실제보다 작게 추정되는 경향이 있다. 10-fold cross validation는 Figure 3과 같이 10개의 데이터 폴드 세트를 만들어서 데이터셋에 있는 9개의 샘플로 모델을 학습시키고 남아있는 하나의 데이터 포인트로 모델을 평가하고 총 10번의 반복이 수행된다. Raschka (2018)에서 LOOCV는 k-fold cross validation 방법 중 $k = n$인 경우, 즉 폴드의 수를 훈련 샘플의 수와 동일하게 한 방법이라고 정의되었다. 작은 데이터 셋에서 유용하게 사용되므로 데이터 셋이 작은 편인 해당 자료의 경우 교차타당성 방법으로 LOOCV를 택하여 판별분석을 진행하고자 한다. Jain 등 (1987)에 의하면 RSM보다 LOOCV의 경우 치우침이 작으며 작은 표본 크기 상황에서 널리 사용된다고 언급하였다.

3. 활용사례

Table 1의 실제 자료를 이용하여 수치 형태를 방사형 차트로 나타내고 형상분석을 적용하여 분류 및 판별을 진행하고자한다. 2장에서 알아본 과정에 따라 이미 알려진 군집에 잘 판별이 되는지 확인해보고자 한다. 활용사례로 스코틀랜드의 Highland, Speyside, Islay & Islands, Lowland 네 지역에 위치한 86개 증류소의 위스키에 대하여 12가지 맛과 향의 정도를 0 ~ 4까지 점수로 표현한 자료를 이용하고자 한다.

Table 1에는 Highland, Speyside, Islay & Islands, Lowland 네 지역에 위치한 86개 증류소의 위스키에 대하여 12가지 맛의 정도를 표현한 데이터 중 일부가 나타나 있다. C는 군집을 나타내며 Highland, Speyside, Islay & Islands, Lowland 지역에 대해 각각 군집 C_1, C_2, C_3, C_4로 정의했고, 12가지의 변수는 x_1 = Body, x_2 = Sweetness, x_3 = Smoky, x_4 = Medicinal, x_5 = Tobacco, x_6 = Honey, x_7 = Spicy, x_8 = Winey, x_9 = Nutty, x_{10} = Malty, x_{11} = Fruity, x_{12} = Floral를 나타내고 있으며 12가지 맛과 향의 정도를 0부터 4까지 점수로 측정되었다. 각 지역별 관측치 수는 총 86개 중 Highland 31개, Speyside 41개, Islay & Islands 11개, Lowland 3개로 나타

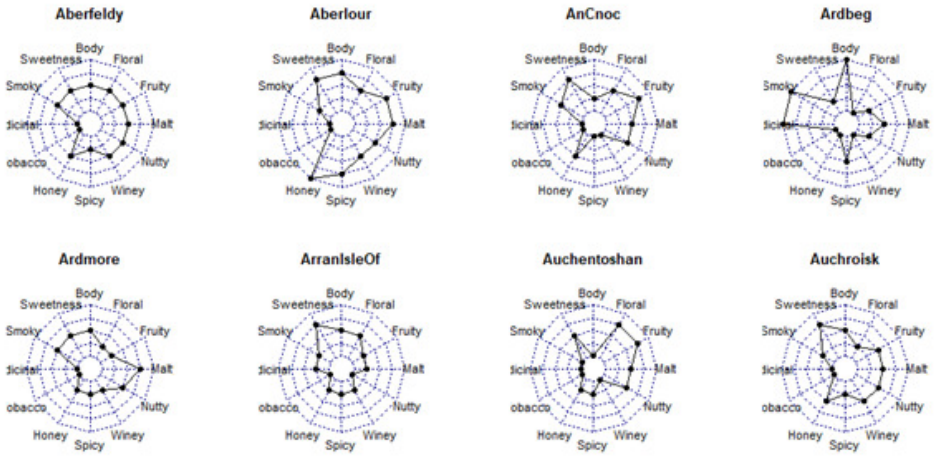


Figure 4: *Some of the shapes of the data represented in the Table 1.*

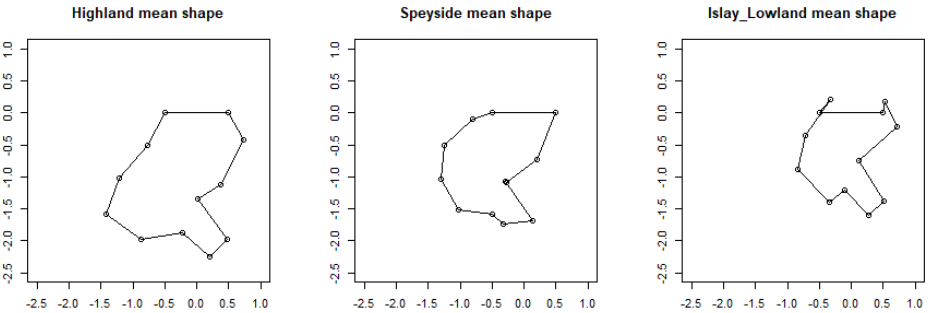


Figure 5: *Bookstein mean shapes of three regions.*

나있다. 일반적으로 Highland에 위치한 증류소의 위스키는 풍부한 과일과 약간의 향신료의 풍미, Speyside는 부드러운 달콤한 과일 풍미, Lowland는 꽃 향이 나는 부드럽고 가벼운 풍미, Islay & Islands는 스모키한 풍미를 가지는 특징이 있는 것으로 알려져 있다. 86개의 증류소 중 임의로 선정한 8개의 증류소 위스키의 맛과 향에 대한 수치형 데이터를 엑셀의 방사형 그래프를 이용하여 나타낸 것이 Figure 4에 나타나있다.

각 변수에 해당하는 점수가 원점 0을 기준으로 1의 간격으로 점으로 표현되었고, 그 점들은 12가지 맛과 향에 대한 형상점이므로 대체적으로 12각형 도형으로 형상화 되었다. 이들 형상에 대한 Figure 5는 북스테인좌표를 구한 것을 바탕으로 세 지역의 북스테인 평균형상을 나타낸 것이다. Highland, Speyside, Islay & Islands & Lowland의 경우 다른 지역보다 형상이 작은 특징을 지니고 있는 것으로 보인다.

먼저 GPA 적합좌표를 이용하여 분류 및 판별의 결과를 살펴본다. Table 2는 판별분석을 위해 생성한 새로운 자료를 활용한 분석 결과를 나타내고 있다. 이미 알려진 군집이 행에 나타나있고 판별함수에 의해 나누어진 분류군집이 열로 나타나있으며 Highland 31개, Speyside 41개, Islay & Islands & Lowland 14개의 총 관측치가 존재한다. 먼저 전통적 판별방법 중 이차판별함수에 의한 판별분석 결과를 보면 RSM으로 분석한 것이 왼쪽에 LOOCV 방법으로 분석한 것이 오른쪽에 나타나있다. RSM의 경우 Highland로 잘 분류된 개체 수가 25개, Speyside로 잘 분류된 개체 수 36개, Islay & Islands & Lowland로 잘 분류된 개체 수 11개로 나타났으며

Table 2: Confusion tables of QDA, SVM and ANN with 12 landmarks

		QDA						SVM						ANN					
		RSM			LOOCV			10-fold CV			LOOCV			10-fold CV			LOOCV		
		C_1	C_2	C_3	C_1	C_2	C_3	C_1	C_2	C_3	C_1	C_2	C_3	C_1	C_2	C_3	C_1	C_2	C_3
Cluster	C_1	25	6	0	20	10	1	31	0	0	31	0	0	19	12	0	15	12	4
	C_2	3	36	2	5	34	2	0	41	0	0	41	0	0	41	0	0	41	0
	C_3	2	1	11	6	1	7	0	1	13	0	1	13	0	4	10	0	4	10
Error Rate		16.28%			29.07%			1.16%			1.16%			18.60%			23.26%		

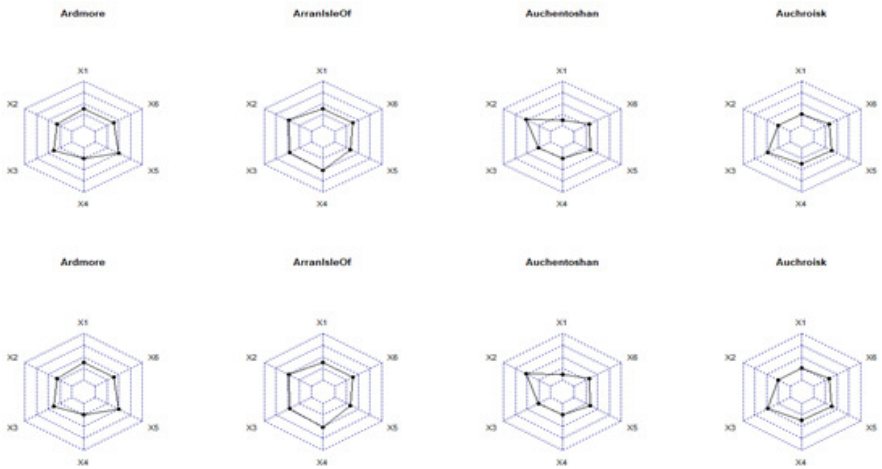


Figure 6: *Some of the shapes of the data using PC scores.*

오분류율은 16.28%이다. 동일한 자료에 대하여 LOOCV로 판별분석을 진행한 결과, Highland로 잘 분류된 개체 수가 20개, Speyside로 잘 분류된 개체 수 34개, Islay & Islands & Lowland로 잘 분류된 개체 수 7개로 나타났으며 오분류율은 29.07%이다. 서포트머신의 경우 10-fold CV 방법에서 오분류율은 1.16%이고 동일한 자료에 대하여 LOOCV로 판별분석을 진행한 결과, 10-fold CV 방법과 동일한 결과를 얻었다. 인공신경망의 경우 10-fold CV 방법에서 오분류율이 18.60%이고 LOOCV 방법의 오분류율은 23.26%로 나타났다.

Figure 5와 같이 12가지 맛과 향에 대한 특성을 모두 사용하여 방사형 차트로 나타낼 경우 한 눈에 개체별 특징을 파악하기는 쉽지 않다. 방사형 차트에서 사각형부터 육각형까지의 형태의 경우 가장 시각적인 효과가 뛰어나다. 따라서 원자료에 대해 보디감, 달콤함, 스모키함, 약내음, 담배, 꿀, 스파이스, 와인, 견과류, 몰트, 과일, 꽃의 맛과 향을 나타내는 12가지 변수들을 수학적 선형결합을 시켜 주성분이라는 새로운 변수들을 만들어 시각적으로 효과적인 형상을 만들고자 한다. 상호 독립성을 보장할 수 없는 원 변수들 12가지에 대해 PCA를 실시하여 생성된 12개의 주성분 중 주성분의 총 분산의 비율인 설명력이 76.56%가 되는 6개의 주성분을 선택하였다. 6개의 주성분이라는 새로운 변수들을 살펴보면 제 1주성분은 보디감이 깊고 달콤함은 적고, 제 2주성분은 보디감과 와인의 향이 적으며 제 3주성분은 달콤하고 와인 및 견과류의 향이 강하다. 제 4주성분은 달콤하고 스파이스한 향이 나고 제 5주성분은 몰트, 담배 향이 강하며 제 6주성분은 꿀과 꽃의 향이 깊다. 따라서 제 1주성분은 무거운 맛, 제 2주성분은 다채롭지 않고 가벼운 맛, 제 3주성분은 고소하고 달달한 맛, 제 4주성분은 계피향이 나는 달달한 맛, 제 5주성분은 숙성된 맛, 제 6주성분은 허브향과 묵직한 꿀맛을 나타내고 있다. 주성분에 중심화 값을 대입하여 86×6 크기의 주성분 점수로 구성된 새로운 자료행렬을 얻게

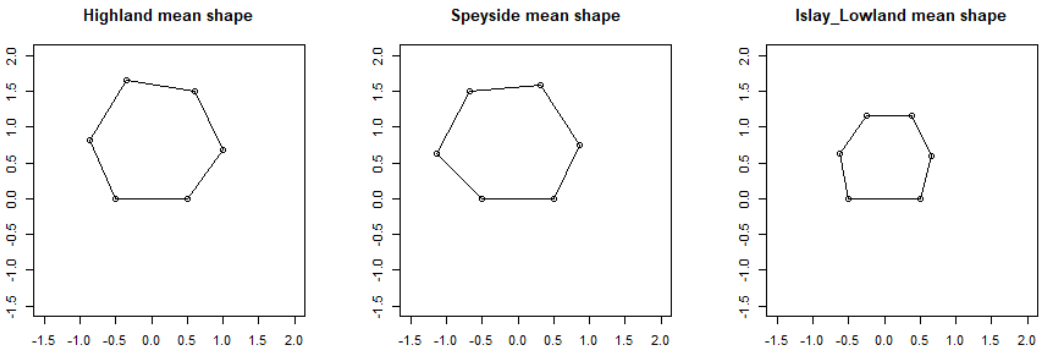


Figure 7: *Bookstein mean shapes of three regions using PC scores.*

Table 3: Confusion tables of QDA, SVM and ANN with 6 bookstein landmarks

		QDA						SVM						ANN					
		RSM			LOOCV			10-fold CV			LOOCV			10-fold CV			LOOVE		
		C_1	C_2	C_3	C_1	C_2	C_3	C_1	C_2	C_3	C_1	C_2	C_3	C_1	C_2	C_3	C_1	C_2	C_3
Cluster	C_1	12	9	10	4	15	12	31	0	0	31	0	0	22	8	1	19	10	2
	C_2	5	27	9	7	23	11	1	40	0	1	40	0	0	41	0	3	37	1
	C_3	1	1	12	7	5	2	0	0	14	0	0	14	5	4	5	3	4	7
Error rate		40.70%			66.28%			1.16%			1.16%			20.93%			26.73%		

되며, 이렇게 생성된 자료행렬을 새로운 자료로 활용한다.

PCA를 통해 구한 주성분 점수를 이용하여 6개의 새로운 변수에 대한 형상을 생성시킨다. 형상은 육각형의 형태로 나타나며 원 형상의 일부가 Figure 6에 나타나있다. Figure 6와 같이 만들어진 형상에 대해 형상분석을 실시하기 위해 식 (2.2)를 적용해 해당 형상점에 대한 형상좌표를 구한다. 형상의 각 꼭짓점은 주성분에 의해 어떤 맛과 향을 나타내는지 알 수 있다. x_1은 무거운 맛, x_2는 다채롭지 않고 가벼운 맛, x_3은 고소하고 달달한 맛, x_4는 계피향이 나는 달달한 맛, x_5는 숙성된 맛, x_6은 허브향과 묵직한 꿀맛을 나타낸다.

형상에 대한 좌표체계는 형상점들의 집합에 대해 이동, 척도, 회전의 효과를 가하여도 변하지 않아야하므로 2차원 평면에 대해 형상점 1과 2를 고정된 위치 $(-1/2, 0)$과 $(1/2, 0)$으로 변환하도록 북스테인좌표를 구하고, 북스테인 평균형상을 이용하여 지역 별 위스키의 맛과 향의 형태를 비교하고자 한다.

Figure 7은 북스테인좌표를 구한 것을 바탕으로 세 지역의 북스테인 평균형상을 나타낸 것이다. Highland, Speyside는 비슷한 형상을 보이고 있으며, Islay & Islands & Lowland의 경우 다른 지역보다 형상이 줄어든 특징을 지니고 있는 것으로 보인다. 이에 대해 판별 분석을 진행하여 각 개체들이 해당 지역에 잘 판별되어지는지 확인하고자 한다.

Table 3은 주성분점수를 통해 새롭게 형성된 6개의 형상점에 대한 판별분석 결과이다. Table 2에서 12개의 형상점을 모두 이용하여 판별분석한 결과와 비교했을 때 GPA 적합좌표를 활용한 경우 QDA에서는 RSM, LOOCV 방법으로는 각각 16.28%에서 40.70%와 29.09%에서 66.28%로 증가하였음을 확인할 수 있다. ANN에서는 10-fold CV와 LOOCV 방법에서 오분류율이 18.60%에서 20.93%와 23.26%에서 26.73%로 소폭 증가하였다. 그리고 SVM에서는 똑같이 1.16%를 기록하였다. SVM과 ANN은 변수 축소로 인한 형상점 갯수의 감소가 예측률에 큰 영향을 주지 않았지만, QDA는 크게 영향을 받은 것으로 확인할 수 있다. 특히, QDA의 LOOCV에서는 예측 모형으로서의 사용이 힘들다고 판단할 수 있다.

4. 결론

형상이라는 용어는 자연적이거나 인위적으로 존재하는 개체를 나타낼 때 주로 사용된다. 효과적인 정보 전달을 위해 수치로 나타난 자료를 인위적으로 방사형 차트로 표현하였고 이를 하나의 형상으로 생각하였다. 나타낸 형상으로 각각의 개체의 특성을 파악할 뿐 아니라 개체 별 형태 차이를 구분하고 이미 알려진 군집으로 잘 분류되어지는지 확인하고자 하였다. 형상분석을 적용하여 각 형상들이 속한 이미 알려진 군집에 대해 알맞게 판별되어지고 분류되는지 확인하였다. 분석을 진행하기 위해 개체의 형상을 잘 대표하는 형상점을 파악하고, 해당 형상점에 대한 형상좌표를 구해야하는데 일반적인 경우 임의의 좌표 체계를 만들어 추출하곤 한다. 그러나 엑셀의 방사형 차트로 형상을 나타냈고 이는 원점을 기준으로 동일한 간격으로 형상점이 생성되어 정다각형 모양이 되므로 피타고라스정리를 이용한 삼각함수를 적용하여 쉽게 형상좌표를 구할 수 있었다. 만약 자료의 변수 개수가 많은 경우는 시각적으로 직관적이라는 방사형 차트의 장점의 의미가 줄어든다. 따라서 형상분석을 적용하기 전, 방사형 차트의 형태를 파악하여 일반적으로 방사형 차트에서 사용되는 육각형이하의 모양으로 만들기 위해 원자료에 대해 PCA를 적용한 후 활용하는 것을 제안하였다.

형상은 위치, 척도, 회전에 의한 효과들이 제거되었을 때 남아있는 모든 기하적인 정보를 의미하므로 이러한 효과들을 고려한 GPA 적합좌표와 북스테인좌표로의 변환을 통해 자료를 만들고 이를 판별분석에 이용하였다. GPA 적합좌표의 경우 적합좌표의 형상과 추정된 평균형상 간 차이인 프로크러스티즈 잔차의 PCA를 적용하여 사용하였으며, 북스테인좌표는 두 기저 형상점을 고정된 위치 $(-1/2, 0)$과 $(1/2, 0)$로 변환하도록 효과를 고려한 후 기저 형상점을 제외한 나머지 좌표를 사용하였다. 추가적으로 리만거리와 중심크기와 같이 각 형상좌표와 상관성이 높은 변수를 추가하여 형상의 정보를 더 많이 담을 수 있는 새로운 자료를 활용하였다.

활용 사례에서 원자료와 PCA를 적용한 자료에 대해 GPA 적합좌표로 세 종류의 방법을 활용하여 판별과 분류를 진행하여 오분류율을 확인하고 PCA를 적용한 자료에 대해 GPA 적합좌표와 북스테인좌표로 판별을 진행하고 비교하였다. 12개의 변수를 6개로 축소시켜 진행하였을 때 오분류율은 높아졌지만 개체의 특성을 시각적으로 파악하기 좋은 형태로 변형된 것을 고려하면 어느정도 분류를 해낸 것으로 해석할 수 있었다. 형상좌표의 관점에서 보면 GPA 적합좌표는 다중공선성을 고려하여 형상 PCA를 적용한 것이므로 기저 형상점 두 개를 제외하고 사용하는 북스테인좌표보다 자료의 손실이 있어 오분류율이 상대적으로 높았다. 그러나 무조건 북스테인좌표의 판별이 더 잘 일어난다고 결론내릴 수는 없으며 자료의 특성에 따라 다를 것이다.

본 연구에서는 관측치 개수가 적은 자료를 활용했고 각 군집 별 개체 수가 불균형이므로 오분류율이 높은 경우도 생겼지만 크기가 큰 데이터 셋을 활용하고 군집 별로 충분히 훈련 가능할 정도의 개체 수가 주어진다면 어떤 성능을 보일지 확인해볼 필요가 있다. 또한 머신러닝과 딥러닝의 다른 이점을 가지는 다른 기법들을 활용하여 형상의 분류와 판별의 정도를 비교해보는 연구가 추가적으로 이어질 수 있을 것이다.

References

Choi YS (2021). *Multivariate Statistical Shape Analysis with R*, Pusan National University Press, Busan.

Choi YS (2018). *Multivariate Data Analysis with R*, Kyungmoon, Seoul.

Dryden IL and Mardia KV (2016). *Statistical Shape Analysis: With Applications in R*, John Wiley & Sons Ltd, Chichester, UK.

Eberhart RC (2014). *Neural Network PC Tools: A Practical Guide*, Academic Press, Inc., New York.

Jain AK, Dubes RC, and Chen CC (1987). Bootstrap techniques for error estimation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **5**, 628–633.

Izenman AJ (2008). *Modern Multivariate Statistical Techniques*, Springer, New York.

Kendall DG (1977). The diffusion of shape, *Advances in Applied Probability*, **9**, 428–430.

Raschka S (2018). Model evaluation, model selection, and algorithm selection in machine learning, Available

from: *arXiv preprint arXiv:1811.12808*

Received July 19, 2023; Revised August 29, 2023; Accepted September 15, 2023

통계적 형상분석을 이용한 엑셀 방사형 차트의 분류와 판별

이승언[a], 김준홍[a], 최연석[a], 최용석[1,a]

[a]부산대학교 통계학과

요 약

평가지표와 같은 수치형 자료의 경우 수치 형태보다 엑셀(Excel)의 방사형 차트 형태로 나타내 시각적으로 표현하면 정보 전달에 더욱 효과적일 것이다. 그러나 개체가 많은 경우 시각적으로 판별하거나 분류하는 것이 쉽지 않다. 이럴 경우 각 개체에 대해 방사형 차트를 이용하여 형상화 시킨 후, 형상의 정보를 대표할 수 있는 형상점을 찾고 형상좌표로 변환해 형상분석을 적용하여 분류 및 판별하는 방법을 알아보고자 한다.
형상분석을 이용하기 위해 주로 분석자의 주관으로 형상점을 얻고 임의의 좌표공간을 생성시켜 좌표를 얻곤 했다. 방사형 차트는 해당 개체의 특징을 나타내는 변수의 개수만큼 형상점이 생기게 되고 이를 선으로 이은 것은 하나의 형상으로 여겨진다. 따라서 중심을 원점으로 두고 2차원 공간으로 정의를 내린 후, X축과 각 특징을 나타내는 축이 이루는 각에 대해 삼각함수를 적용해 형상좌표를 추출해낸다. 변수의 개수가 많아 형상의 모양이 복잡해질 경우 방사형 차트를 이용해 시각화하더라도 쉽게 파악하기 어렵다. 독립성을 보장할 수 없는 변수들에 대해 주성분 분석(PCA)을 실시하여 시각적으로 효과적인 형상을 만든다. PCA를 실시하기 전과 후의 형상에 대해 전통적 판별분석, 서포트벡터머신(support vector machine; SVM), 인공신경망(artificial neural network; ANN)의 기법을 적용시켜 분류표와 분류율을 확인한다. 또한 GPA (generalized procrustes analysis) 적합좌표, 북스테인좌표 2가지 좌표에 대한 판별의 차이를 비교한다. 북스테인좌표의 경우 기저 형상점을 중심으로 형상의 위치와 회전, 척도를 변환한 좌표로써, 분류율에 대해 GPA 형상좌표보다 더 높은 결과를 보이고 있다. 북스테인좌표의 경우 여러 군집 간의 형상을 비교하는데 유용하게 활용된다.

주요용어: 형상분석, 방사형 차트, GPA, 북스테인좌표

[1]교신저자: (46241) 부산광역시 금정구 부산대학로 63번길 2, 부산대학교 통계학과. E-mail: yschoi@pusan.ac.kr