응용통계연구 (2011)
24(5), 941–949
DOI: http://dx.doi.org/10.5351/KJAS.2011.24.5.941

세 집단 판별분석 상황에서의 영향함수 유도 및 그 응용

이혜정[1] · 김홍기[2]

[1]충남대학교 정보통계학과, [2]충남대학교 정보통계학과

(2011년 7월 접수, 2011년 8월 채택)

요 약

본 논문에서는 세 집단만을 판별분석 할 경우에 계산되는 오분류확률에 영향을 미치는 이상치 판별을 목적으로 하며, 쉽게 응용 가능한 간단한 영향함수식을 제시하였다. 그리고 제시된 수식을 이용하여 안면 데이터로 세 가지 사상체질을 분류해보고 각 관찰값들의 오분류확률에 대한 영향함수를 계산하였다. 이상치를 제거하고 재 판별분석을 하는데 있어, 오분류확률에 대한 영향함수를 이용하는 것이 효율적인 방법임을 확인하였다.

주요용어: 영향함수, 판별분석, 오분류확률, 이상치.

1. 서론

우리는 자료집단에서 다른 관찰값에 비해 유난히 작거나 큰 값으로 보통의 관찰값과는 다른 관찰값을 이상치(outlier)라 정의한다. 이상치를 발견하고 선택하기 위해 사용되는 방법 중 하나인 영향함수(influence function)는 Hampel (1974)에 의해 처음으로 소개되었으며, Hampel에 의하여 제안된 영향함수는 통계학의 모든 분야에서 응용되고 있다.

Kim (1992)은 이차원 분할표의 대응분석에서 고유치들에 대한 이론적인 영향함수를 유도하였으며 이를 다차원 분할표의 대응분석으로 확장하여 적용하였다. Kim과 Lee (1996), Kim (1998)은 χ^2통계량에 대한 영향함수를 다루었고, Kim 등 (2003)은 허용한계에 대한 영향함수를 그리고 Kim과 Kim (2005)은 t통계량에 대한 영향함수를 유도하였고, Lee (2008)는 변이계수에 대한 영향함수를 유도하였다. 다변량 분석방법 중의 하나인 판별분석(discriminant analysis)에서는 이를 시행하기 위해 계산하는 표본 평균과 분산-공분산 행렬이 이상치에 대해 민감하다. Campbell (1978)은 판별분석에서 이상치 탐지에 영향함수를 최초로 이용하였고, Johnson (1987)은 Bayesian approach에 영향함수를 이용하였다. 그 후에 Critchley와 Vitiello (1991)과 Fung (1992)이 두 그룹을 판별하는 판별분석에 영향함수를 이용하였고, Fung (1996)은 세 집단 이상을 판별하는 판별분석에서 영향함수를 이용하여 이상치를 검색하고 이상치에 따른 오분류확률(misclassification probability)을 탐색하는 연구로 확장하였다.

영향함수를 이용하여 판별분석 시 이상치를 탐색하는 기존 연구에서는 탐색된 이상치를 제거하여 나머지 데이터를 이용하여 판별분석을 하였을 경우 오분류확률이 어느정도 향상되는지는 고려되지 않았다. 하지만 실제 상황에서 이상치를 정의하고 이상치를 제외한 나머지 데이터를 이용하여 그룹을 판별하는 모형을 분석하고자 할 때에는 이상치를 정의하는 기준과 판별분석 결과 오분류확률이 어느정도 나타나는지를 고려할 필요성이 있다.

[2]교신저자: (305-764) 대전광역시 유성구 대학로 99, 충남대학교 정보통계학과, 교수.
E-mail: honggiekim@cnu.ac.kr

사상체질의학은 조선시대 이제마 선생으로부터 창안되어 인간을 사상체질(태양, 태음, 소양, 소음)로 분류해 이에 맞추어 생리, 병리, 진단, 관리 기준을 적용하는 한의학의 한 학문 분야이다. 사상체질은 다양한 방법으로 분류하는데 한의학적 진단방법인 보고 듣고 묻고 만져보는 사진법을 이용하여 체질을 진단하고 있다. 그리고 이러한 과정을 객관화 및 표준화 하여 한의사 진단법을 재현할 수 있도록 다양한 한방의료기기가 개발되고 있으며, 진단에 활용되는 한방의료기기의 품질향상을 위한 신뢰도 증진 연구가 진행되었으며 (이혜정 등, 2010; Ryu 등, 2010), 진단 정확도를 높이기 위한 타당도 연구가 진행되었다 (강재환 등, 2009; 진희정 등, 2010). 본 연구에 사용된 데이터의 체질진단결과는 약진을 통해 확인된 체질결과를 이용하였으며 (진희정 등, 2010), 이 데이터를 기반으로 안면, 체형, 음성, 피부, 성정 등 여러 가지 특성을 모두 종합하여 체질을 진단하는 연구가 진행 중에 있다. 그 중에서도 안면의 특징은 체질진단에 있어서 활용도가 가장 높은 것으로 조사 되어 있으며 (이준희 등, 2007), 안면사진에 나타난 특징점 사이의 거리, 각도, 비율 정보 등을 이용하여 체질을 판별하는 연구들이 진행되고 있다 (이의주 등, 2005; Koo 등, 2009; 도준형 등, 2010). 하지만 체질이 진단된 전체 데이터 중에는 안면의 체질적 특징을 갖고 있지 않는 데이터가 포함되어 있어 실제 임상에서 안면 정보의 활용도는 높지만 데이터 분석결과 모형의 타당성은 높지 않은 편이다. 따라서 전체 데이터를 모두 이용하여 체질을 판별하는 것보다 체질특성이 애매한 케이스를 제외하고 체질특성이 뚜렷한 케이스만 이용하여 체질을 판별하는 분석을 할 필요성이 있다. 이때, 실제로는 네 개의 체질이 존재하나 태양인은 그 수가 무시할 만큼 적어 거의 모든 사상체질 판별분석은 세 그룹간의 판별분석으로 충분해진다.

본 연구는 안면 데이터에서 영향함수를 도입하여 안면 특성으로 체질이 확실히 구분되지 않는 사람을 밝혀내고, 안면에서 나타나는 체질특성이 뚜렷한 사람들을 대상으로 체질을 재 판별 한 경우 오분류확률이 어느정도 감소하는지 검토해 보고자 한다.

2. 오분류확률에 대한 영향함수

2.1. 오분류확률(misclassification probability)

다변량 정규분포를 따르는 m개의 모집단이 있다고 가정하고($N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $i = 1, \ldots, m$), 계산의 간편화를 위해 집단의 사전확률(prior probability)과 비용함수(cost function)는 동일하다고 가정한다. 관찰벡터 $\boldsymbol{x}$는 p개의 판별변수로 구성되어있는 랜덤벡터이며, 이 관찰벡터가 R_i에 속하면 그룹 π_i로 판정한다. 이때, 오분류확률(misclassification probability)이 최소가 되도록 하는 판별분석의 규칙에 따른 R_i는 다음과 같이 주어진다 (Fung, 1996).

$$R_i = \bigcap_{j \neq i} \left[2(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x} + \boldsymbol{\mu}_i^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i - \boldsymbol{\mu}_j^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_j < 0 \right]. \tag{2.1}$$

π_i와 π_j의 그룹간 마할라노비스 거리(Mahalanobis distance)의 제곱은 아래와 같다.

$$\Delta_{ij}^2 = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j).$$

Fung (1996)은 판별분석의 규칙인 식 (2.1)에 의하여 실제 π_i그룹에 속하는 관찰벡터의 오분류확률인 MP_i가 다음과 같이 주어짐을 보였다.

$$\mathrm{MP}_i = 1 - G\left(b_{i1}, \ldots, b_{i,i-1}, b_{i,i+1}, \ldots, b_{im}\right),$$

여기서 $b_{ij} = \Delta_{ij}/2$이며 G는 다변량 표준정규분포함수이다.

위 식은 $1 - P(R_i \,|\, \boldsymbol{x} \in \pi_i)$의 다른 표현이며, 즉 $P(R_i \,|\, \boldsymbol{x} \in \pi_i) = G(b_{i1}, \ldots, b_{i,i-1}, b_{i,i+1}, \ldots, b_{im})$이며, $z_j = (\boldsymbol{\mu}_j - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i)/\Delta_{ij}$라 할 때, 식 (2.1) 내의 부등식 $2(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x} + \boldsymbol{\mu}_i^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i -$

$\boldsymbol{\mu}_j^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j < 0$은 $(\boldsymbol{\mu}_j-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)/\Delta_{ij} < \Delta_{ij}/2$가 되므로 $P(R_i\,|\,\boldsymbol{x}\in\pi_i)$는 $P[\bigcap_{j\neq i} z_j < b_{ij}]$가 된다. 이때 z_j들은 각각 표준정규분포를 따르며 $j\neq k$인 두 z_j, z_k간의 공분산은 다음과 같이 주어진다 (Fung, 1996).

$$\mathrm{cov}(z_j,z_k)=\frac{\Delta_{ij}^2+\Delta_{ik}^2-\Delta_{jk}^2}{2\Delta_{ij}\Delta_{ik}}. \tag{2.2}$$

2.2. 영향함수(influence functions)

T는 분포함수에 대해 실수값을 갖는 범함수(real-valued functional), 즉 일련의 모수이고, F_i, $(i=1,\ldots,m)$는 분포함수라고 하자. 그리고 δ_x는 실수 공간의 한 점인 x에서 확률이 1인 분포함수이다. r번째 그룹의 분포함수에 임의의 관찰값 $\boldsymbol{x}$를 추가함으로써 생기는 F_r과 δ_x의 혼합분포함수는 $\widetilde{F}_r=(1-\epsilon)F_r+\epsilon\delta_x$이며, 이때 $\widetilde{F}_r$을 F_r의 섭동(perturbation)이라 한다.

Hampel (1974)은 범함수에 $T(F)$대한 $\boldsymbol{x}$의 영향함수(influence function)를 다음과 같이 정의하였다.

$$\mathrm{IF}(T,\boldsymbol{x})=\lim_{\epsilon\to 0}\frac{T(F_r)-T(F)}{\epsilon}.$$

그리고 Fung (1996)은 r번째 그룹에 섭동(pertubation)이 된 분포함수를 바탕으로 Δ_{ij}^2의 영향함수가 아래와 같이 주어짐을 보였다.

$$I\left(\boldsymbol{x};\Delta_{ij}^2\right)=\begin{cases} w_r\Delta_{ij}^2-w_r\psi_{ij}^2, & i\neq r,\ j\neq r,\\ w_r\Delta_{ij}^2+2\psi_{ij}-w_r\Delta_{ij}^2, & i=r,\\ w_r\Delta_{ij}^2-2\psi_{ij}-w_r\Delta_{ij}^2, & j=r,\end{cases} \tag{2.3}$$

여기서 $\psi_i=(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y}$이며, $\boldsymbol{y}=\boldsymbol{x}-\boldsymbol{\mu}_r$이다.

한편, 같은 상황에서 Fung (1996)은 오분류확률에 대한 영향함수를 다음과 같은 형태로 유도하였다.

$$I(\boldsymbol{x};\mathrm{MP}_i)=-\sum_{j\neq i}\frac{\partial G(b_{i1},\ldots,b_{i,i-1},b_{i,i+1},\ldots,b_{im})}{\partial b_{ij}}I(\boldsymbol{x};\Delta_{ij}^2)/4\Delta_{ij}. \tag{2.4}$$

위의 식은 m개의 모집단에 대해 판별분석을 시행할 경우에 적용되는 영향함수의 일반적인 형태이다. 본 연구에서는 위의 식을 우리가 관심을 갖는 3개의 모집단에 대한 판별분석을 시행하는 경우로 구체화하여 제시하고 이를 이용하여 오분류확률의 변화를 살펴보았다.

3. 3개 집단 판별분석에서의 오분류확률에 대한 영향함수

3개의 모집단에 대한 오분류확률의 영향함수는 식 (2.4)에서 $i=1,2,3$인 경우이며, $i=1$인 경우의 오분류확률에 대한 영향함수는 아래와 같다.

$$\begin{aligned} I(x;\mathrm{MP}_1)&=-\sum_{j\neq i}\frac{\partial G(b_{12},b_{13})}{\partial b_{1j}}\bullet I(x;\Delta_{ij}^2)/4\Delta_{ij}\\ &=-\frac{\partial G(z_2<b_{12},z_3<b_{13})}{\partial b_{12}}\left(\frac{I(x;\Delta_{12}^2)}{4\Delta_{12}}\right)-\frac{\partial G(z_2<b_{12},z_3<b_{13})}{\partial b_{13}}\left(\frac{I(x;\Delta_{13}^2)}{4\Delta_{13}}\right), \end{aligned} \tag{3.1}$$

여기서 $G(z_2<b_{12},z_3<b_{13})$는 $\int_{-\infty}^{b_{13}}\int_{-\infty}^{b_{12}}f(z_2,z_3)dz_2dz_3$이며, z_2,z_3는 각각 표준정규분포를 따르고 공분산이 식 (2.2)에 의해 주어지므로, 이변량 결합확률밀도함수 $f(z_2,z_3)$는 아래와 같이 주어진다.

$$f(z_2,z_3)=\frac{1}{2\pi|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\bullet\exp\left(-\begin{pmatrix}z_2\\ z_3\end{pmatrix}'\boldsymbol{\Sigma}^{-1}\begin{pmatrix}z_2\\ z_3\end{pmatrix}/2\right),$$

여기서 $\mathbf{\Sigma} = \begin{bmatrix} 1 & \mathrm{cov}(z_2,z_3) \\ \mathrm{cov}(z_2,z_3) & 1 \end{bmatrix}$이다.

그러므로 식 (3.1)의 첫 번째 편미분 $\{\partial G(z_2 < b_{12}, z_3 < b_{13})\}/\partial b_{12}$는 $\partial/\partial b_{12} \int_{-\infty}^{b_{13}} \int_{-\infty}^{b_{12}} f(z_2,z_3)dz_2dz_3 = \int_{-\infty}^{b_{13}} f(b_{12},z_3)dz_3$로 주어지게 된다.

한편 이 식의 $f(b_{12},z_3)$는 아래와 같다.

$$f(b_{12},z_3) = \frac{1}{2\pi|\mathbf{\Sigma}|^{\frac{1}{2}}} \bullet \exp\left(-\begin{pmatrix} b_{12} \\ z_3 \end{pmatrix}' \mathbf{\Sigma}^{-1} \begin{pmatrix} b_{12} \\ z_3 \end{pmatrix}/2\right), \tag{3.2}$$

여기서 $\mathbf{\Sigma} = \begin{bmatrix} 1 & (\Delta_{12}^2+\Delta_{13}^2-\Delta_{23}^2)/2\Delta_{12}\Delta_{13} \\ (\Delta_{12}^2+\Delta_{13}^2-\Delta_{23}^2)/2\Delta_{12}\Delta_{13} & 1 \end{bmatrix}$이다.

식 (3.2)의 $\begin{pmatrix} b_{12} \\ z_3 \end{pmatrix}' \mathbf{\Sigma}^{-1} \begin{pmatrix} b_{12} \\ z_3 \end{pmatrix}$은 아래와 같이 유도된다.

$$\begin{aligned}
\begin{pmatrix} b_{12} \\ z_3 \end{pmatrix}' \mathbf{\Sigma}^{-1} \begin{pmatrix} b_{12} \\ z_3 \end{pmatrix} &= \begin{pmatrix} b_{12} \\ z_3 \end{pmatrix}' \frac{1}{|\mathbf{\Sigma}|} \begin{bmatrix} 1 & -\dfrac{(\Delta_{12}^2+\Delta_{13}^2-\Delta_{23}^2)}{2\Delta_{12}\Delta_{13}} \\ -\dfrac{(\Delta_{12}^2+\Delta_{13}^2-\Delta_{23}^2)}{2\Delta_{12}\Delta_{13}} & 1 \end{bmatrix} \begin{pmatrix} b_{12} \\ z_3 \end{pmatrix} \\
&= \frac{1}{|\mathbf{\Sigma}|}\left(b_{12}^2 - b_{12}z_3\frac{(\Delta_{12}^2+\Delta_{13}^2-\Delta_{23}^2)}{\Delta_{12}\Delta_{13}} + z_3^2\right) \\
&= \frac{1}{|\mathbf{\Sigma}|}\left[z_3 - \frac{b_{12}(\Delta_{12}^2+\Delta_{13}^2-\Delta_{23}^2)}{2\Delta_{12}\Delta_{13}}\right]^2 - \frac{b_{12}^2(\Delta_{12}^2+\Delta_{13}^2-\Delta_{23}^2)^2}{4\Delta_{12}^2\Delta_{13}^2} + b_{12}^2
\end{aligned}$$

그러므로 다시 식 (3.2)를 정리하면

$$\frac{1}{\sqrt{2\pi}\sqrt{2\pi}|\mathbf{\Sigma}|^{\frac{1}{2}}} \bullet \exp\left(-\left[z_3 - \frac{b_{12}(\Delta_{12}^2+\Delta_{13}^2-\Delta_{23}^2)}{2\Delta_{12}\Delta_{13}}\right]^2/2|\mathbf{\Sigma}|\right) \bullet \exp\left(-b_{12}^2\left[1 - \frac{(\Delta_{12}^2+\Delta_{13}^2-\Delta_{23}^2)^2}{4\Delta_{12}^2\Delta_{13}^2}\right]/2|\mathbf{\Sigma}|\right)$$

가 된다. 여기서 $1/(\sqrt{2\pi}|\mathbf{\Sigma}|^{1/2}) \bullet \exp(-[z_3 - \{b_{12}(\Delta_{12}^2+\Delta_{13}^2-\Delta_{23}^2)\}/\{2\Delta_{12}\Delta_{13}\}]^2/2|\mathbf{\Sigma}|)$는 평균이 $\{b_{12}(\Delta_{12}^2+\Delta_{13}^2-\Delta_{23}^2)\}/(2\Delta_{12}\Delta_{13})$이고, 분산이 $1-(\Delta_{12}^2+\Delta_{13}^2-\Delta_{23}^2)^2/(4\Delta_{12}^2\Delta_{13}^2)$인 정규분포의 *pdf*이며, 이를 이용하여 다시 식 (3.1)의 첫 번째 편미분 식을 정리하면 아래와 같다.

$$\frac{1}{\sqrt{2\pi}} \bullet e^{-\frac{b_{12}^2}{2}} \bullet \int_{-\infty}^{b_{13}} f\left(x;\mu_{12},\sigma_{12}^2\right)dx,$$

여기서 $\mu_{12} = \{b_{12}(\Delta_{12}^2+\Delta_{13}^2-\Delta_{23}^2)\}/(2\Delta_{12}\Delta_{13})$, $\sigma_{12}^2 = 1-(\Delta_{12}^2+\Delta_{13}^2-\Delta_{23}^2)^2/(4\Delta_{12}^2\Delta_{13}^2)$이다.

이와 같은 방법으로 식 (3.1)의 두 번째 편미분 $\partial G(z_2 < b_{12}, z_3 < b_{13})/(\partial b_{13})$을 정리하면 아래와 같다.

$$\frac{1}{\sqrt{2\pi}} \bullet e^{-\frac{b_{13}^2}{2}} \bullet \int_{-\infty}^{b_{12}} f\left(x;\mu_{13},\sigma_{13}^2\right)dx,$$

여기서 $\mu_{13} = b_{13}(\Delta_{12}^2+\Delta_{13}^2-\Delta_{23}^2)/(2\Delta_{12}\Delta_{13})$, $\sigma_{13}^2 = 1-(\Delta_{12}^2+\Delta_{13}^2-\Delta_{23}^2)^2/(4\Delta_{12}^2\Delta_{13}^2)$이다.

이 편미분 값들과 식 (2.3)에 주어진 Δ_{ij}^2에 대한 영향함수를 이용하면 $I(x;\mathrm{MP}_1)$를 구할 수 있으며 아래와 같이 나타나고, 동일한 과정으로 $i = 2,3$인 경우의 오분류확률에 대한 영향함수는 아래와 같이 주어진다.

$$\begin{aligned}
&I(x;\mathrm{MP}_1) \\
&= -\sum_{j\neq i} \frac{\partial G(b_{12},b_{13})}{\partial b_{1j}} \bullet I(x;\Delta_{ij}^2)/4\Delta_{ij} \\
&= -\frac{1}{\sqrt{2\pi}} e^{-\frac{b_{12}^2}{2}} \int_{-\infty}^{b_{13}} f(x;\mu_{12},\sigma_{12}^2)dx \bullet \frac{I(x;\Delta_{12}^2)}{4\Delta_{12}} - \frac{1}{\sqrt{2\pi}} e^{-\frac{b_{13}^2}{2}} \int_{-\infty}^{b_{12}} f(x;\mu_{13},\sigma_{13}^2)dx \bullet \frac{I(x;\Delta_{13}^2)}{4\Delta_{13}},
\end{aligned}$$

여기서 $\mu_{12} = b_{12}(\Delta_{12}^2+\Delta_{13}^2-\Delta_{23}^2)/(2\Delta_{12}\Delta_{13})$, $\mu_{13} = b_{13}(\Delta_{12}^2+\Delta_{13}^2-\Delta_{23}^2)/(2\Delta_{12}\Delta_{13})$이며, $\sigma_{12}^2 = \sigma_{13}^2 = 1-(\Delta_{12}^2+\Delta_{13}^2-\Delta_{23}^2)^2/(4\Delta_{12}^2\Delta_{13}^2)$이다.

$$
\begin{aligned}
&I(x;\mathrm{MP}_2)\\
&= -\sum_{j\neq i}\frac{\partial G(b_{12},b_{23})}{\partial b_{2j}} \bullet \frac{I(x;\Delta_{ij}^2)}{4\Delta_{ij}}\\
&= -\frac{1}{\sqrt{2\pi}}e^{-\frac{b_{12}^2}{2}}\int_{-\infty}^{b_{23}} f(x;\mu_{21},\sigma_{21}^2)dx \bullet \frac{I(x;\Delta_{12}^2)}{4\Delta_{12}} - \frac{1}{\sqrt{2\pi}}e^{-\frac{b_{23}^2}{2}}\int_{-\infty}^{b_{12}} f(x;\mu_{23},\sigma_{23}^2)dx \bullet \frac{I(x;\Delta_{23}^2)}{4\Delta_{23}},
\end{aligned}
$$

여기서 $\mu_{21} = b_{12}(\Delta_{12}^2+\Delta_{23}^2-\Delta_{13}^2)/(2\Delta_{12}\Delta_{23})$, $\mu_{23} = b_{23}(\Delta_{12}^2+\Delta_{23}^2-\Delta_{13}^2)/(2\Delta_{12}\Delta_{23})$이며, $\sigma_{21}^2 = \sigma_{22}^2 = 1-(\Delta_{12}^2+\Delta_{23}^2-\Delta_{13}^2)^2/(4\Delta_{12}^2\Delta_{23}^2)$이다.

$$
\begin{aligned}
&I(x;\mathrm{MP}_3)\\
&= -\sum_{j\neq i}\frac{\partial G(b_{13},b_{23})}{\partial b_{3j}} \bullet \frac{I(x;\Delta_{ij}^2)}{4\Delta_{ij}}\\
&= -\frac{1}{\sqrt{2\pi}}e^{-\frac{b_{13}^2}{2}}\int_{-\infty}^{b_{23}} f(x;\mu_{31},\sigma_{31}^2)dx \bullet \frac{I(x;\Delta_{13}^2)}{4\Delta_{13}} - \frac{1}{\sqrt{2\pi}}e^{-\frac{b_{23}^2}{2}}\int_{-\infty}^{b_{13}} f(x;\mu_{32},\sigma_{32}^2)dx \bullet \frac{I(x;\Delta_{23}^2)}{4\Delta_{23}},
\end{aligned}
$$

여기서 $\mu_{31} = b_{13}(\Delta_{13}^2+\Delta_{23}^2-\Delta_{12}^2)/(2\Delta_{13}\Delta_{23})$, $\mu_{32} = b_{23}(\Delta_{13}^2+\Delta_{23}^2-\Delta_{12}^2)/(2\Delta_{13}\Delta_{23})$이며, $\sigma_{31}^2 = \sigma_{32}^2 = 1-(\Delta_{13}^2+\Delta_{23}^2-\Delta_{12}^2)^2/(4\Delta_{13}^2\Delta_{23}^2)$이다.

이와 같이 3가지 그룹에 대한 오분류확률의 영향함수를 구한 후 이를 평균하여 전체 오분류확률의 영향함수 $I(x;\mathrm{MP})$를 구한다. 일반적인 영향함수는 모집단 분포함수 F_i에서 정의되며 $I(x;\mathrm{MP})$로 주어지지만 표본으로부터 추정하는 경험적 분포함수 $\hat{F}_i$에 의하여 경험적 영향함수(empirical influence function; EIF)인 $\mathrm{EI}(x;\mathrm{MP})$가 구해지게 된다.

4. 예제

조선시대에 이제마 선생이 발표한 사상체질의학에서는 사람의 체질을 오장육부의 허와 실의 정도에 따라 4가지 체질로 분류하고 있으며 최근에는 사상체질을 객관화와 표준화된 방법으로 분류하는 연구가 진행되고 있다. 사상체질을 안면정보만을 이용하여 분류할 경우, 체질특성이 애매한 케이스를 제외하고 체질특성이 뚜렷한 케이스만 이용하여 체질을 판별할 필요성이 있다.

따라서 본 장에서는 영향함수를 이용하여 이상치를 제거한 경우에 판별분석을 시행한 결과의 오분류확률이 어떻게 변화하는지 검토하였다. 분석에 사용된 데이터는 약진결과를 이용하여 사상체질이 태음인(TE), 소양인(SY), 소음인(SE)으로 판정된 360명의 남성 데이터이며 7개의 안면변수를 이용하여 판별분석을 시행하고 각 데이터에 대하여 오분류확률의 영향함수를 구하였다. 각 데이터에 대한 $\mathrm{EI}(x;\mathrm{MP})$는 그림 4.1과 같다.

$\mathrm{EI}(x;\mathrm{MP})$분포는 0에 근접하게 나타났으며, $\mathrm{EI}(x;\mathrm{MP})$값이 큰 값은 해당 데이터를 추가한 경우 오분류확률이 커짐을 의미하며 $\mathrm{EI}(x;\mathrm{MP})$값이 작은 값은 해당 데이터를 추가한 경우 오분류확률이 작아짐을 의미한다. 따라서 $\mathrm{EI}(x;\mathrm{MP})$값이 0보다 작은 경우의 케이스들은 판별분석을 시행할 경우 오분류확률이 작아짐을 예상할 수 있다.

$\mathrm{EI}(x;\mathrm{MP})$값이 작은 케이스들만으로 한 판별분석 결과의 오분류확률이 감소하는지 확인하기 위해 $\mathrm{EI}(x;\mathrm{MP})$의 기준을 줄이면서 케이스를 탈락시키고 남은 케이스들만으로 판별분석을 재 시행하였고 그 결과는 표 4.1과 같다.

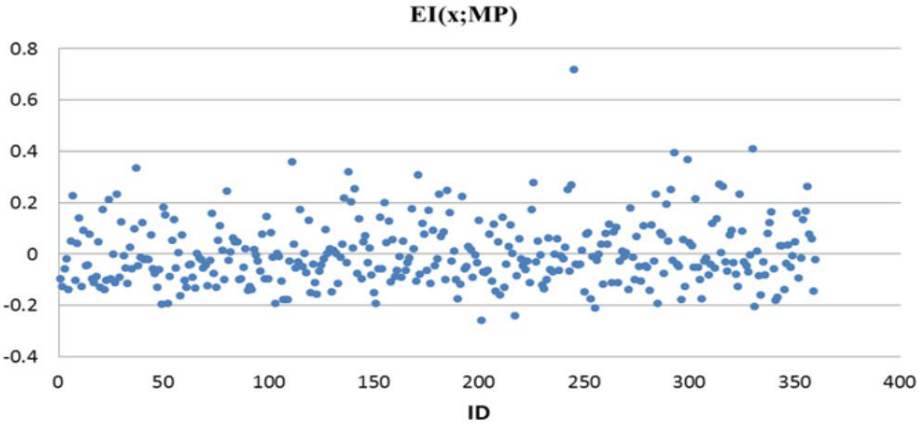

그림 **4.1.** EI(x; MP) 분포

표 **4.1.** EI(x; MP)의 기준에 따른 판별분석 결과

구분			예측 소속집단				N	%N	오분류확률
			TE	SE	SY	전체			
전체	빈도	TE	**100**	27	27	154	360	100%	46.1%
		SE	20	**53**	21	94			
		SY	29	42	**41**	112			
EI(x; MP) < 0.17	빈도	TE	**9**4	11	28	133	324	90.0%	39.8%
		SE	11	**55**	21	87			
		SY	23	35	**46**	104			
EI(x; MP) < 0.087	빈도	TE	**90**	12	20	122	288	80.0%	33.3%
		SE	6	**53**	16	75			
		SY	19	23	**49**	91			
EI(x; MP) < 0.046	빈도	TE	**92**	7	11	110	252	70.0%	24.6%
		SE	5	**51**	9	65			
		SY	12	18	**47**	77			
EI(x; MP) < 0.0026	빈도	TE	**86**	3	5	94	216	60.0%	15.3%
		SE	2	**48**	5	55			
		SY	7	11	**49**	67			
EI(x; MP) < 0	빈도	TE	**85**	3	4	92	211	58.6%	13.7%
		SE	1	**47**	5	53			
		SY	7	9	**50**	66			

EI(x; MP) 분포는 0에 근접하게 나타났으며, EI(x; MP)값이 큰 값은 해당 데이터를 추가한 경우 오분류확률이 커짐을 의미하며 EI(x; MP)값이 작은 값은 해당 데이터를 추가한 경우 오분류확률이 작아짐을 의미한다. 따라서 EI(x; MP)값이 0보다 작은 경우의 케이스들은 판별분석을 시행할 경우 오분류확률이 작아짐을 예상할 수 있다.

EI(x; MP)값이 작은 케이스들만으로 한 판별분석 결과의 오분류확률이 감소하는지 확인하기 위해 EI(x; MP)의 기준을 줄이면서 케이스를 탈락시키고 남은 케이스들만으로 판별분석을 재 시행하였고 그 결과는 표 4.1과 같다.

표 **4.2.** 판별확률의 기준에 따른 판별분석 결과

구분			예측 소속집단				N	$\%N$	오분류확률
			TE	SE	SY	전체			
전체	빈도	TE	100	27	27	154	360	100%	46.1%
		SE	20	53	21	94			
		SY	29	42	41	112			
판별확률 >0.3867	빈도	TE	**94**	26	22	142	324	90.0%	43.2%
		SE	16	**50**	13	79			
		SY	24	39	**40**	103			
판별확률 >0.4197	빈도	TE	**86**	22	22	130	288	80.0%	42.0%
		SE	11	**51**	8	70			
		SY	24	34	**30**	88			
판별확률 >0.443	빈도	TE	**75**	22	16	113	252	70.0%	40.9%
		SE	7	**48**	6	61			
		SY	21	31	**26**	78			
판별확률 >0.472	빈도	TE	**70**	18	11	99	216	60.0%	39.8%
		SE	6	**42**	6	54			
		SY	17	28	**18**	63			

$\mathrm{EI}(x;\mathrm{MP})$값이 작은 케이스들만 이용하여 판별분석을 시행할수록 오분류확률이 줄어듦을 확인하였다. 판별분석을 시행하면 각 집단의 판별점수가 나타나며 판별된 집단의 판별확률값도 계산된다. 이 판별확률값을 기준으로 케이스를 선택한 경우에 오분류확률이 어떻게 나타나는지 살펴보고 $\mathrm{EI}(x;\mathrm{MP})$를 이용하여 분석한 결과와 비교해 보았다. 전체 케이스를 대상으로 판별분석을 시행하였을 경우 나타나는 판별확률값의 기준을 높이면서 선택된 케이스들만으로 판별분석을 재 시행하였고 그 결과는 표 4.2와 같다.

판별확률이 큰 케이스들만 이용하여 판별분석을 시행할수록 오분류확률이 줄어듦을 확인하였다. 하지만 $\mathrm{EI}(x;\mathrm{MP})$를 이용하여 분석한 결과와 비교해보면 오분류확률이 $\mathrm{EI}(x;\mathrm{MP})$를 이용한 경우보다 더 크게 나타났다. 따라서 판별확률을 이용하여 케이스를 제외하는 것보다 $\mathrm{EI}(x;\mathrm{MP})$를 이용하여 케이스를 제외하여 판별분석을 재 시행하는 것이 더 효과적임을 알 수 있었다.

5. 결론

본 논문에서는 일반적으로 다루고 있는 세 집단 이상을 판별하는 판별분석 시 나타나는 오분류확률의 영향함수를 세 집단만을 판별분석 할 경우에 나타나는 오분류확률의 영향함수에 한정시킴으로써 쉽게 응용 가능한 간단한 함수식을 제시하였다. 그리고 영향함수를 이용하여 판별분석 시 이상치를 구분하고 이상치를 제외하고 재판별분석을 하였을 경우 오분류확률이 어떻게 감소하는지 알아보았다. 안면 데이터를 이용하여 세 가지 사상체질을 분류해보고 $\mathrm{EI}(x;\mathrm{MP})$를 계산하였다. $\mathrm{EI}(x;\mathrm{MP})$값이 큰 값이 나타난 케이스는 해당 케이스가 포함되면 오분류확률이 커짐을 의미하고 이 값이 작게 나타난 케이스는 해당 케이스가 포함되면 오분류확률이 작아짐을 의미한다. 따라서 $\mathrm{EI}(x;\mathrm{MP})$값이 작은 케이스들만 판별분석을 시행하였을 경우 판별분석 결과의 오분류확률이 작아질 것을 예측할 수 있고, 실제로 $\mathrm{EI}(x;\mathrm{MP})$값의 기준이 작아질수록 오분류확률이 감소하는 것을 확인하였다.

또한 판별분석 결과 나타나는 판별확률값을 기준으로 판별확률값이 큰 케이스들만 판별분석을 시행하였을 경우의 오분류확률이 어떻게 나타나는지 살펴보았다. 그 결과 판별확률값의 기준이 커질수록 오분류

확률이 감소하는 것을 확인하였으나, EI(x; MP)값을 기준으로 분석한 경우보다 오분류확률이 더 많이 감소하지 않았다. 따라서 판별분석 시 이상치를 정의하고 이상치를 제외하고 재 판별분석을 시행할 경우에 오분류확률의 영향함수값을 이용하는 것이 효율적임을 확인하였다.

참고문헌

강재환, 유종향, 이혜정, 김종열 (2009). 음성을 이용한 사상체질 분류 알고리즘, <말소리와 음성과학>, **1**, 155–161.

도준형, 김근호, 김종열 (2010). 다양한 환경 조건에서의 얼굴 윤곽선 영역 검출을 위한 분할 영역 히스토그램 분석, <전자공학회지>, **47**, 1–8.

이의주, 손은혜, 유정희, 김종원, 김규곤, 고병희 (2005). 四象人의 容貌에 관한 문헌적 연구, <사상체질의학회지>, **17**, 55–68.

이준희, 김윤희, 황민우, 김종열, 이의주, 송일병, 고병희 (2007). 四象人의 안면, 음성, 피부 및 맥진 특성에 관한 설문조사 연구, <사상체질의학회지>, **19**, 126–143.

이혜정, 강남식, 전영주, 김근호, 김홍기, 김종열 (2010). 피부탄성 측정 문제점 개선을 위한 6시그마 프로젝트, <한국한의학연구원논문집>, **16**, 135–140.

진희정, 이혜정, 김명근, 김홍기, 김종열 (2010). 사상체질 판별을 위한 2단계 의사결정 나무 분석, <사상체질의학회지>, **22**, 87–97.

Campbell, N. A. (1978). The influence function as an aid in outlier detection in discriminant analysis, *Applied Statistics*, **27**, 251–258.

Critchley, F. and Vitiello, C. (1991). The influence of observations on misclassification probability estimates in linear discriminant analysis, *Biometrika*, **78**, 677–690.

Fung, W. K. (1992). Some diagnostic measures in discriminant analysis, *Statistics & Probability Letters*, **13**, 279, 285.

Fung, W. K. (1996). The influence of observations on misclassification probability in multiple discriminant analysis, *Communications in Statistics-Theory and Methods*, **25**, 1917–1930.

Hampel, F. R. (1974). The influence curve and its role in robust estimation, J*ournal of the American Statistical Association*, **69**, 383–393.

Johnson, W. (1987). The detection of influential observations for allocation, separation, and determination of probabilities in a Bayesian framework, *Journal of Business & Economic Statistics*, **5**, 369–381.

Kim, H. (1992). Measures of influence in correspondence analysis, *Journal of Statistical Computation and Simulation*, **40**, 201–217.

Kim, H. (1998). A study on cell influence to Chi-square statistic in contingency tables, *The Korean Communications in Statistics*, **5**, 35–42.

Kim, H. and Kim, K. H. (2005). Influence of an observation on the t-statistic, *The Korean Communications in Statistics*, **12**, 453–462.

Kim, H. and Lee, H. S. (1996). Influence functions on χ^2 statistics in contingency tables, *The Korean Communications in Statistics*, **3**, 69–76.

Kim, H., Lee, Y. H., Shin, H. S. and Lee, S. (2003). Influence function on tolerance limit, *The Korean Communications in Statistics*, **10**, 305–317.

Koo, I., Kim, J., Kim, M. and Kim, K. (2009). Feature Selection from a Facial Image for Distinction of Sasang Constitution, *Evidence-Based Complementary and Alternative Medicine*, **6**, 65–71.

Lee, Y. (2008). Influence function on the coefficient of variation, *The Korean Communications in Statistics*, **15**, 509–516.

Ryu, H., Lee, H., Kim, H. and Kim, J. (2010). Reliability and Validity of a Cold-Heat Pattern Questionnaire for Traditional Chinese Medicine, *The Journal of Alternative and Complementary Medicine*, **16**, 663–667.

Derivation and Application of Influence Function in Discriminant Analysis for Three Groups

Haejung Lee[1] · Honggie Kim[2]

[1]Department of Information & Statistics, Chungnam National University
[2]Department of Information & Statistics, Chungnam National University

(Received July 2011; accepted August 2011)

Abstract

The influence function is used to develop criteria to detect outliers in discriminant analysis. We derive the influence function of observations that estimate the the misclassification probability in discriminant analysis for three groups. The proposed measures are applied to the facial image data to define outliers and redo the discriminant analysis excluding the outliers. The study proves that the derived influence function is more efficient than using the discriminant probability approach.

Keywords: Influence function, discriminant analysis, misclassification probability, outlier.

[2]Corresponding author: Professor, Department of Information & Statistics, Chungnam National University, Daejeon 305-764, Korea. E-mail: honggiekim@cnu.ac.kr