김일성종합대학학보 주체109(2020)년
수 학 제66권 제4호

제한된 오판별확률을 가지는 최량판별규칙에 대하여

림 창 호

경애하는 최고령도자 **김정은**동지께서는 다음과 같이 말씀하시였다.

《이미 일정한 토대가 있고 전망이 확고한 연구대상들에 힘을 넣어 세계패권을 쥐며 그 성과를 확대하는 방법으로 과학기술을 빨리 발전시켜야 합니다.》

판별분석문제들은 패턴인식과 의학진단, 금융 등 많은 현실문제들에서 제기된다.

선행연구[2]에서는 2개의 모집단에 대하여 한 모집단의 오판별확률이 α를 넘지 않게 하는 조건밑에서 다른 모집단의 오판별확률이 최소로 되는 네이만-피어슨판별규칙과 그에 대한 추론문제를 론의하였으며 선행연구[3]에서는 네이만-피어슨분류에 대한 연구를 하면서 모집단의 개수 m이 $m\geq 3$인 경우에 네이만-피어슨판별의 확장문제를 론의하였다. 또한 선행연구[1]에서는 일반화된 네이만-피어슨판별규칙과 그에 대한 ROC분석문제를 론의하였다. 그리고 선행연구[4]에서는 오판별확률에 대한 제한이 없는 최량판별규칙(베이스판별규칙)에 대하여 론의하였다.

론문에서는 제한된 오판별확률을 가지는 한가지 최량판별규칙에 대하여 연구하였다.

표본 $\boldsymbol{x}=(x_1,\ \cdots,\ x_p)^{\mathrm{T}}$가 모집단 $G=\{G_1,\ \cdots,\ G_m\}$에 속한다고 하고 $G_r\ (r=1,\ \cdots,\ m)$의 밀도함수와 사전확률을 각각 $f_r(\boldsymbol{x}),\ \pi_r\ (r=1,\ \cdots,\ m)$라고 하자.

표본 $\boldsymbol{x}\in R^p$와 모집단 $G_r\ (r=1,\ \cdots,\ m)$에 관한 자료에 기초하여 $\boldsymbol{x}$가 모집단 $G_r\ (r=1,\ \cdots,\ m)$에 속하는가 또는 속하지 않는가를 판단하는것을 판별분석이라고 부른다.

판별규칙으로서는 표본공간 $(R^p,\ B^p)$의 어떤 분할 $R^p=R_1^p+\cdots+R_m^p\ (R_r^p\in B^p\ (r=1,\ \cdots,\ m))$을 생각하고 $\boldsymbol{x}\in R_r^p \Rightarrow \boldsymbol{x}\in G_r$로 판별한다. $\boldsymbol{x}\in G_r$일 때 $\boldsymbol{x}\notin G_r$라고 잘못 판별하는 오판별확률은 $\int_{\overline{R}_r^p} f_r(\boldsymbol{x})d\boldsymbol{x}$이며 총오판별확률은

$$p=\sum_{r=1}^m q_r p_r = q_1\int_{\overline{R}_1^p} f_1(\boldsymbol{x})d\boldsymbol{x}+\cdots+q_m\int_{\overline{R}_m^p} f_m(\boldsymbol{x})d\boldsymbol{x}$$

이다. 여기서 $\overline{R}_r^p+R_r^p=R^p\ (r=1,\ \cdots,\ m)$이다.

두 모집단에 대한 네이만-피어슨판별규칙은

$$((R_1^P)^*,\ (R_2^P)^*)_\alpha \in \arg\min_{\Phi(\alpha)} \int_{R_1^P} f_2(\boldsymbol{x})d\boldsymbol{x}$$

이며 $\Phi(\alpha)=\left\{(R_1^p,\ R_2^p)\left|\int_{\overline{R}_1^p} f_1(\boldsymbol{x})d\boldsymbol{x}\leq\alpha,\ R_1^p,\ R_2^p\in B^p\right.\right\}$, $\overline{R}_r^p=R^p-R_r^p\ (r=1,\ 2)$이다. 여기서 α는 사용자가 제한하여주는 수준으로서 $0\leq\alpha\leq 1$이다.[2]

m개의 모집단에 대한 일반화된 네이만-피어슨판별규칙은

$$((R_1^P)^*,\ \cdots,\ (R_m^P)^*)_{(\alpha_1,\ \cdots,\ \alpha_{m-1})} \in \arg\min_{\Phi(\alpha_1,\ \cdots,\ \alpha_{m-1})} \int_{\overline{R}_m^P} f_m(\boldsymbol{x})d\boldsymbol{x}$$

$$\Phi(\alpha_1,\ \cdots,\ \alpha_{m-1}) = \left\{ (R_1^p,\ \cdots,\ R_{m-1}^p) \left| \int_{\overline{R}_1^p} f_1(\boldsymbol{x})d\boldsymbol{x} \le \alpha_1,\ \cdots,\ \int_{\overline{R}_{m-1}^p} f_{m-1}(\boldsymbol{x})d\boldsymbol{x} \le \alpha_{m-1}\ \ (R_1^p,\ \cdots,\ R_{m-1}^p \in B^p) \right. \right\}$$

$\overline{R}_r^p = R^p - R_r^p\ \ (r=1,\ \cdots,\ m)$이다. $\alpha_1,\ \cdots,\ \alpha_{m-1}$은 사용자가 제한하여주는 수준이다.[1]

그리고 최량판별규칙은

$$((R_1^P)^*,\ \cdots,\ (R_m^P)^*) \in \arg\min_{\Phi(R_1^p,\ \cdots,\ R_{m-1}^p)} \left(\pi_1 \int_{\overline{R}_1^P} f_1(\boldsymbol{x})d\boldsymbol{x} + \cdots + \pi_m \int_{\overline{R}_m^P} f_m(\boldsymbol{x})d\boldsymbol{x} \right)$$

$$\Phi(R_1^p,\ \cdots,\ R_{m-1}^p) = \{(R_1^p,\ \cdots,\ R_{m-1}^p) \mid R^p = R_1^p + \cdots + R_m^p\ \ (R_1^p,\ \cdots,\ R_{m-1}^p \in B^p)\}$$

$$\overline{R}_r^p = R^p \setminus R_r^p\ \ (r=1,\ \cdots,\ m)$$

이다.[4]

우리는 $m=3$개의 모집단에 대하여 다음과 같이 제한된 오판별확률을 가지는 최량판별규칙

$$((R_1^P)^*,\ (R_2^P)^*,\ (R_3^P)^*)_\alpha \in \arg\min_{\Phi(\alpha)} \left(\pi_2 \int_{\overline{R}_2^P} f_2(\boldsymbol{x})d\boldsymbol{x} + \pi_3 \int_{\overline{R}_3^P} f_3(\boldsymbol{x})d\boldsymbol{x} \right)$$

$$\Phi(\alpha) = \left\{ (R_1^p,\ R_2^p,\ R_3^p) \left| \int_{\overline{R}_1^p} f_1(\boldsymbol{x})d\boldsymbol{x} \le \alpha\ \ (R_1^p,\ R_2^p,\ R_3^p \in B^p) \right. \right\},\ \overline{R}_r^p = R^p - R_r^p\ \ (r=1,\ 2,\ 3)$$

에 대하여 론의하려고 한다. 여기서 α는 사용자가 제한하여주는 수준으로서 $0 \le \alpha \le 1$이다.

보조정리 1 모집단 $G_r\ (r=1,\ 2,\ 3)$의 밀도함수를 $f_r(\boldsymbol{x})\ (r=1,\ 2,\ 3)$라고 하면 일반화된 네이만—피어슨판별규칙은

$$\begin{aligned} (R_1^P)^* &= \{\boldsymbol{x} \mid f_2(\boldsymbol{x}) \le c_{\alpha_1} f_1(\boldsymbol{x}),\ f_3(\boldsymbol{x}) \le c_{\alpha_2} f_1(\boldsymbol{x})\} \\ (R_2^P)^* &= \{\boldsymbol{x} \mid f_2(\boldsymbol{x}) > c_{\alpha_1} f_1(\boldsymbol{x}),\ c_{\alpha_1} f_3(\boldsymbol{x}) \le c_{\alpha_2} f_2(\boldsymbol{x})\} \\ (R_3^P)^* &= \{\boldsymbol{x} \mid f_3(\boldsymbol{x}) > c_{\alpha_2} f_1(\boldsymbol{x}),\ c_{\alpha_1} f_3(\boldsymbol{x}) > c_{\alpha_2} f_2(\boldsymbol{x})\} \end{aligned} \tag{1}$$

이다. 여기서

$$c_{\alpha_1} \ge 0,\ c_{\alpha_2} \ge 0,\ \int_{(\overline{R}_1^p)^*} f_1(\boldsymbol{x})d\boldsymbol{x} = \alpha_1,\ \int_{(\overline{R}_2^p)^*} f_2(\boldsymbol{x})d\boldsymbol{x} = \alpha_2,\ (\overline{R}_r^p)^* = R^p - (R_r^p)^*\ \ (r=1,\ 2,\ 3)$$

이며 $\alpha_1,\ \alpha_2$는 조건 $0 \le \alpha_1 \le 1,\ \int_{(R_1^p)^*} f_2(\boldsymbol{x})d\boldsymbol{x} \le \alpha_2 \le 1$을 만족시킨다.[1]

보조정리 2 모집단 $G_r\ (r=1,\ 2,\ 3)$의 밀도함수와 사전확률을 $f_r(\boldsymbol{x}),\ \pi_r\ \ (r=1,\ 2,\ 3)$라고 하면 최량판별규칙은

$$(R_1^P)^* = \{\boldsymbol{x} \mid \pi_2 f_2(\boldsymbol{x}) \le \pi_1 f_1(\boldsymbol{x}),\ \pi_3 f_3(\boldsymbol{x}) \le \pi_1 f_1(\boldsymbol{x})\}$$
$$(R_2^P)^* = \{\boldsymbol{x} \mid \pi_2 f_2(\boldsymbol{x}) > \pi_1 f_1(\boldsymbol{x}),\ \pi_3 f_3(\boldsymbol{x}) \le \pi_2 f_2(\boldsymbol{x})\} \tag{2}$$
$$(R_3^P)^* = \{\boldsymbol{x} \mid \pi_3 f_3(\boldsymbol{x}) > \pi_1 f_1(\boldsymbol{x}),\ \pi_3 f_3(\boldsymbol{x}) > \pi_2 f_2(\boldsymbol{x})\}$$

이다.[4]

정리 모집단 G_r $(r=1,\ 2,\ 3)$의 밀도함수와 사전확률을 각각 $f_r(\boldsymbol{x}),\ \pi_r$ $(r=1,\ 2,\ 3)$라고 하면 제한된 오판별확률을 가지는 최량판별규칙은

$$(R_1^P)^* = \{\boldsymbol{x} \mid f_2(\boldsymbol{x}) \le c_\alpha f_1(\boldsymbol{x}),\ \pi_3 f_3(\boldsymbol{x}) \le \pi_2 c_\alpha f_1(\boldsymbol{x})\}$$
$$(R_2^P)^* = \{\boldsymbol{x} \mid f_2(\boldsymbol{x}) > c_\alpha f_1(\boldsymbol{x}),\ \pi_3 f_3(\boldsymbol{x}) \le \pi_2 f_2(\boldsymbol{x})\} \tag{3}$$
$$(R_3^P)^* = \{\boldsymbol{x} \mid \pi_3 f_3(\boldsymbol{x}) > \pi_2 c_\alpha f_1(\boldsymbol{x}),\ \pi_3 f_3(\boldsymbol{x}) > \pi_2 f_2(\boldsymbol{x})\}$$

이다. 여기서 $c_\alpha \ge 0,\ \int_{(\overline{R}_1^p)^*} f_1(\boldsymbol{x})d\boldsymbol{x} = \alpha,\ (\overline{R}_1^p)^* = R^p - (R_1^p)^*$이며 $0 \le \alpha \le 1$이다.

참 고 문 헌

[1] 림창호, 정현성; 조선민주주의인민공화국 과학원통보, 2, 11, 주체108(2019).

[2] A. Zhao et al.; Journal of Machine Learning Research, 17, 213, 1, 2016.

[3] Xin Tong et al.; Advanced Review, 8, 64, 2016.

[4] J. M. Geoffry; Discriminant Analysis and Statistical Pattern Recognition, John Wiley & Sons, 1～524, 2004.

주체109(2020)년 6월 5일 원고접수

On an Optimal Discriminant Rule with Restricted Probability of Misclassification

Rim Chang Ho

In this paper, a discriminant rule with restricted probability of misclassification

$$((R_1^P)^*,\ (R_2^P)^*,\ (R_3^P)^*)_\alpha \in \arg\min_{\Phi(\alpha)} \left(\pi_2 \int_{\overline{R}_2^P} f_2(\boldsymbol{x})d\boldsymbol{x} + \pi_3 \int_{\overline{R}_3^P} f_3(\boldsymbol{x})d\boldsymbol{x} \right)$$

$$\Phi(\alpha) = \left\{ (R_1^p,\ R_2^p,\ R_3^p) \left| \int_{\overline{R}_1^p} f_1(\boldsymbol{x})d\boldsymbol{x} \le \alpha \ \ (R_1^p,\ R_2^p,\ R_3^p \in B^p) \right. \right\},\ \overline{R}_r^p = R^p - R_r^p \ \ (r=1,\ 2,\ 3)$$

is considered, here $f_r(\boldsymbol{x})$ and π_r for $r=1, 2, 3$ is respectively the density function and prior probability of the population G_r. And has been obtained the optimal discriminant rule with restricted probability of misclassification.

Keyword: optimal discriminant rule