김일성종합대학학보 수 학

주체108(2019)년 제65권 제3호

일반화된 네이만-피어슨판별규칙과 동등한 판별규칙

정현성, 림창호

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

《과학연구부문에서는 나라의 경제발전과 인민생활향상에서 전망적으로 풀어야 할 문제들과 현실에서 제기되는 과학기술적문제들을 풀고 첨단을 돌파하여 지식경제건설의 지름길을 열어놓아야 합니다.》

판별분석문제들은 패턴인식과 종양진단, 금융 등 많은 현실문제들에서 제기되며 모집단은 보통 여러개로 주어진다. 론문에서는 판별능력이 최량인 일반화된 네이만-피어슨판별규칙과 ROC곡면의 견지에서 동등한 판별규칙들에 대하여 론의한다.

1. 선행연구결과와 문제설정

선행연구[1]에서는 $m=2$인 경우에 제1종의 오판별확률이 α를 넘지 않는 조건하에서 제2종의 오판별확률이 최소로 되는 네이만-피어슨판별규칙과 그에 대한 추론문제들을 론의하였으며 선행연구[2]에서는 베이즈판별규칙과 우도비판별규칙에 대하여 고찰하였다. 또한 선행연구[3]에서는 $m\geq 3$인 경우에 ROC곡면구성과 그에 대한 추정문제들을 론의하였으며 선행연구[4]에서는 $m=2$인 경우에 선형판별함수에 의한 중소기업들사이의 국내외의 금융선택권의 판별분석문제를 취급하였다.

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 밀도함수를 $f_r(\boldsymbol{x})\ (r=1,\ \cdots,\ m)$라고 하자.

판별규칙으로서는 표본공간 $(R^p,\ B^p)$의 어떤 분할규칙 $R^p=R_1^p+\cdots+R_m^p\quad (R_r^p\in B^p,\ r=1,\ \cdots,\ m)$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 $\alpha_r=\int_{\overline{R}_r^p} f_r(\boldsymbol{x})d\boldsymbol{x}$이며 $\boldsymbol{x}\in G_r$라고 정확히 판별하는 정판별확률은 $\beta_r=\int_{R_r^p} f_r(\boldsymbol{x})d\boldsymbol{x}$이다. 여기서 $\overline{R}_r^p=R^p\setminus R_r^p$이다.

ROC곡면은 $m\geq 3$개의 모집단에 대한 판별규칙에서 모집단을 식별하는 여러가지 선별값(판별경계값)들에 대하여 그 선별값들에 따르는 G_1의 정판별확률을 1축으로, G_2의 정판별확률을 2축으로, $\cdots$, G_m의 정판별확률을 m축으로 배치한 점들의 곡면이다.

두 판별규칙의 ROC곡면이 같으면 동등한 판별규칙이라고 부른다.

$m\geq 3$인 여러 모집단 $G_r\ (r=1,\cdots,m)$의 밀도함수 $f_r(\boldsymbol{x})\ (r=1,\cdots,\ m)$에 대하여

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

$$\Phi(\alpha_1,\ \cdots,\ \alpha_{m-1})=$$
$$=\left\{(R_1^p,\ \cdots,\ R_{m-1}^p)\left|\int\limits_{\overline{R}_1^p} f_1(\mathbf{x})d\mathbf{x}\le\alpha_1,\ \cdots,\ \int\limits_{\overline{R}_{m-1}^p} f_{m-1}(\mathbf{x})d\mathbf{x}\le\alpha_{m-1},\ R_1^p,\ \cdots,\ R_{m-1}^p\in B^p\right.\right\}$$
$$\overline{R}_r^p=R^p\setminus R_r^p\ \ (r=1,\ \cdots,\ m)$$

로 되는 일반화된 네이만-피어슨판별규칙은

$$(R_1^P)^*=\{\mathbf{x}\,|\,f_2(\mathbf{x})\le c_{\alpha_1}f_1(\mathbf{x}),\ \cdots,\ f_m(\mathbf{x})\le c_{\alpha_{m-1}}f_1(\mathbf{x})\}$$
$$(R_2^P)^*=\{\mathbf{x}\,|\,f_2(\mathbf{x})> c_{\alpha_1}f_1(\mathbf{x}),\ \cdots,\ c_{\alpha_1}f_m(\mathbf{x})\le c_{\alpha_{m-1}}f_2(\mathbf{x})\}$$
$$\vdots$$
$$(R_m^P)^*=\{\mathbf{x}\,|\,f_m(\mathbf{x})> c_{\alpha_{m-1}}f_1(\mathbf{x}),\cdots,c_{\alpha_{m-2}}f_m(\mathbf{x})> c_{\alpha_{m-1}}f_{m-1}(\mathbf{x})\}\tag{1}$$
$$c_{\alpha_1}\ge 0,\ \cdots,\ c_{\alpha_{m-1}}\ge 0:\ \int\limits_{(\overline{R}_1^p)^*} f_1(\mathbf{x})d\mathbf{x}=\alpha_1,\ \cdots,\ \int\limits_{(\overline{R}_{m-1}^p)^*} f_{m-1}(\mathbf{x})d\mathbf{x}=\alpha_{m-1}$$
$$(\overline{R}_r^p)^*=R^p\setminus(R_r^p)^*,\ r=1,\ \cdots,\ m$$
$$0\le\alpha_1\le 1,\ \int\limits_{(R_1^p)^*} f_2(\mathbf{x})d\mathbf{x}\le\alpha_2\le 1,\cdots,\ \int\limits_{(R_1^p)^*+\cdots+(R_{m-2}^p)^*} f_{m-1}\ (\mathbf{x})d\mathbf{x}\le\alpha_{m-1}\le 1$$

론문에서는 일반화된 네이만-피어슨판별규칙과 동등한 판별규칙들에 대하여 론의한다.

2. 일반화된 네이만-피어슨판별규칙과 동등한 판별규칙

보조정리 1 모집단 $G_r\ (r=1,\ \cdots,\ m)$의 밀도함수를 $f_r(\mathbf{x})\ (r=1,\ \cdots,\ m)$라고 하면 일반화된 네이만-피어슨판별규칙에서 여러가지 선별값 $c_{\alpha_1},\ \cdots,\ c_{\alpha_{m-1}}$에 따르는 판별규칙은

$$(R_1^P)^*=\{\mathbf{x}\,|\,q_2f_2(\mathbf{x})\le q_1f_1(\mathbf{x}),\ \cdots,\ q_mf_m(\mathbf{x})\le q_1f_1(\mathbf{x})\}$$
$$(R_2^P)^*=\{\mathbf{x}\,|\,q_2f_2(\mathbf{x})> q_1f_1(\mathbf{x}),\ \cdots,\ q_mf_m(\mathbf{x})\le q_2f_2(\mathbf{x})\}$$
$$\vdots\tag{2}$$
$$(R_m^P)^*=\{\mathbf{x}\,|\,q_mf_m(\mathbf{x})> q_1f_1(\mathbf{x}),\ \cdots,\ q_mf_m(\mathbf{x})> q_{m-1}f_{m-1}(\mathbf{x})\}$$
$$0\le q_i\le 1,\ i=1,\ \cdots,\ m,\ q_1+\cdots+q_m=1$$

과 같다. 여기서

$$q_1=c_{\alpha_1}q_2,\ \cdots,\ q_1=c_{\alpha_{m-1}}q_m,\ c_{\alpha_1}q_2=c_{\alpha_2}q_3,\ \cdots,\ c_{\alpha_{m-2}}q_{m-1}=c_{\alpha_{m-1}}q_m$$

이다.

모집단 $G_r\ (r=1,\ \cdots,\ m)$의 밀도함수 $f_r(\mathbf{x})=f(\mathbf{x}\,|\,G_r)\ (r=1,\ \cdots,\ m)$와 $\mathbf{x}$에 대응하는 모집단 G_r의 우도 $L(G_r\,|\,\mathbf{x})=f(\mathbf{x}\,|\,G_r)$에 대하여

$$r^*=\arg\max_r\{L(G_r\,|\,\mathbf{x})\}\Rightarrow\mathbf{x}\in G_{r^*}$$

로 되는 판별규칙을 우도비판별규칙이라고 부른다.

보조정리 2 모집단 $G_r\ (r=1,\ \cdots,\ m)$의 밀도함수 $f_r(\mathbf{x})\ (r=1,\ \cdots,\ m)$에 대한 우도비판별규칙에서 여러가지 선별값 $c_1\ge 0,\ \cdots,\ c_{m-1}\ge 0$에 따르는 판별규칙은

$$(R_1^P)^* = \{\boldsymbol{x} \mid q_2 f_2(\boldsymbol{x}) \le q_1 f_1(\boldsymbol{x}),\ \cdots,\ q_m f_m(\boldsymbol{x}) \le q_1 f_1(\boldsymbol{x})\}$$
$$(R_2^P)^* = \{\boldsymbol{x} \mid q_2 f_2(\boldsymbol{x}) > q_1 f_1(\boldsymbol{x}),\ \cdots,\ q_m f_m(\boldsymbol{x}) \le q_2 f_2(\boldsymbol{x})\}$$
$$\vdots$$
$$(R_m^P)^* = \{\boldsymbol{x} \mid q_m f_m(\boldsymbol{x}) > q_1 f_1(\boldsymbol{x}),\ \cdots,\ q_m f_m(\boldsymbol{x}) > q_{m-1} f_{m-1}(\boldsymbol{x})\}$$
$$0 \le q_i \le 1,\ i=1,\ \cdots,\ m,\ q_1+\cdots+q_m=1$$

과 같다. 여기서

$$q_1 = c_1 q_2,\ \cdots,\ q_1 = c_{m-1} q_m,\ c_1 q_2 = c_2 q_3,\ \cdots,\ c_{m-2} q_{m-1} = c_{m-1} q_m$$

이다.

모집단 G_r $(r=1,\ \cdots,\ m)$의 밀도함수 $f_r(\boldsymbol{x})$ $(r=1,\ \cdots,\ m)$와 사전확률 π_r $(r=1,\ \cdots,\ m)$에 대하여

$$((R_1^P)^*,\ \cdots,\ (R_m^P)^*) \in \arg\min_{\Phi(R_1^p,\ \cdots,\ R_{m-1}^p)} \pi_1 \int_{\overline{R}_1^P} f_1(\boldsymbol{x})d\boldsymbol{x} + \cdots + \pi_m \int_{\overline{R}_m^P} f_m(\boldsymbol{x})d\boldsymbol{x}$$
$$\Phi(R_1^p,\ \cdots,\ R_{m-1}^p) = \{(R_1^p,\ \cdots,\ R_{m-1}^p) \mid R^p = R_1^p + \cdots + R_m^p,\ R_1^p,\ \cdots,\ R_{m-1}^p \in B^p\}$$
$$\overline{R}_r^p = R^p \setminus R_r^p \quad (r=1,\ \cdots,\ m)$$

로 되는 판별규칙을 베이즈판별규칙이라고 부른다.

보조정리 3 모집단 G_r $(r=1,\ \cdots,\ m)$의 밀도함수 $f_r(\boldsymbol{x})$ $(r=1,\ \cdots,\ m)$와 사전확률 π_r $(r=1,\ \cdots,\ m)$에 대한 베이즈판별규칙에서 여러가지 선별값 $c_1' \ge 0,\ \cdots,\ c_{m-1}' \ge 0$에 따르는 판별규칙은

$$(R_1^P)^* = \{\boldsymbol{x} \mid q_2 f_2(\boldsymbol{x}) \le q_1 f_1(\boldsymbol{x}), \cdots, q_m f_m(\boldsymbol{x}) \le q_1 f_1(\boldsymbol{x})\}$$
$$(R_2^P)^* = \{\boldsymbol{x} \mid q_2 f_2(\boldsymbol{x}) > q_1 f_1(\boldsymbol{x}),\ \cdots,\ q_m f_m(\boldsymbol{x}) \le q_2 f_2(\boldsymbol{x})\}$$
$$\vdots$$
$$(R_m^P)^* = \{\boldsymbol{x} \mid q_m f_m(\boldsymbol{x}) > q_1 f_1(\boldsymbol{x}),\ \cdots,\ q_m f_m(\boldsymbol{x}) > q_{m-1} f_{m-1}(\boldsymbol{x})\}$$
$$0 \le q_i \le 1\ (i=1,\ \cdots,\ m,\ q_1+\cdots+q_m=1)$$

과 같다. 여기서

$$q_1 = c_1' \frac{\pi_1}{\pi_2} q_2,\ \cdots,\ q_1 = c_{m-1}' \frac{\pi_1}{\pi_m} q_m,\ c_1' \pi_3 q_2 = c_2' \pi_2 q_3,\ \cdots,\ c_{m-2}' \pi_m q_{m-1} = c_{m-1}' \pi_{m-1} q_m$$

이다.

정리 1 모집단 G_r $(r=1,\ \cdots,\ m)$의 밀도함수를 $f_r(\boldsymbol{x})$ $(r=1,\ \cdots,\ m)$라고 하면 일반화된 네이만-피어슨판별규칙과 우도비판별규칙, 베이즈판별규칙은 모두 동등한 판별규칙이다.

주의 모집단 G_r $(r=1,\ \cdots,\ m)$의 밀도함수와 사전확률을 각각 $f_r(\boldsymbol{x})$, π_r $(r=1,\ \cdots,\ m)$라고 하고 r째 모집단의 표본을 s째 모집단의 표본으로 잘못 판별했을 때의 손실을 $L(s \mid r)$, $L(r \mid r)=0$이라고 할 때 평균손실이 최소로 되는 판별규칙도 우의 판별규칙들과 동등한 판별규칙이라는것을 마찬가지로 증명할수 있다.

정리 2 모집단 G_r $(r=1,\ \cdots,\ m)$의 밀도함수와 사전확률을 각각 $f_r(\boldsymbol{x})$, π_r $(r=1,\ \cdots,\ m)$라고 하자. 이때 일반화된 네이만-피어슨판별규칙의 선별값 $c_{\alpha_1},\ \cdots,\ c_{\alpha_{m-1}}$과 우도비판별규칙의 선별값 $c_1,\ \cdots,\ c_{m-1}$, 베이즈판별규칙의 선별값 $c_1',\ \cdots,\ c_{m-1}'$들사이에는 관계식

$$c_{\alpha_1} = c_1 = c_1' \frac{\pi_1}{\pi_2},\ \cdots,\ c_{\alpha_{m-1}} = c_{m-1} = c_{m-1}' \frac{\pi_1}{\pi_m}$$

이 성립한다.

참 고 문 헌

[1] A. Zhao et al.; Journal of Machine Learning Research, 17, 1, 2016.
[2] S. Theodoridis et al.; Pattern Recognition, Elsevier, 13～86, 2009.
[3] Y. Zhang et al.; Biometrical Journal, 58, 6, 1338, 2016.
[4] O. O. Nto Philips et al.; International Journal of Research in Business Studies and Management, 6, 2, 8, 2015.

주체108(2019)년 3월 15일 원고접수

The Generalized Neyman-Pearson Discriminant Rule and Equivalent Discriminant Rules

Jong Hyon Song, Rim Chang Ho

In this paper, we study the generalized Neyman-Pearson discriminant rule, Likelihood ratio discriminant rule and Bayes discriminant rule for the several populations, and prove the equivalence of the discriminant rules by ROC surface. And we obtain a relation of their thresholds.

Key words: discriminant analysis, discriminant rule, ROC surface