行政院國家科學委員會專題研究計畫 成果報告

接受者操作特徵函數線下面積之無母數迴歸分析
研究成果報告(精簡版)

計 畫 類 別 ：個別型
計 畫 編 號 ：NSC 98-2118-M-004-004-
執 行 期 間 ：98 年 08 月 01 日至 99 年 07 月 31 日
執 行 單 位 ：國立政治大學統計學系

計 畫 主 持 人：薛慧敏
共 同 主 持 人：張源俊
計畫參與人員：碩士班研究生-兼任助理人員：劉世凰
博士班研究生-兼任助理人員：許嫚荏

報 告 附 件 ：出席國際會議研究心得報告及發表論文

處 理 方 式 ：本計畫涉及專利或其他智慧財產權，2 年後可公開查詢

中 華 民 國 99 年 10 月 28 日

THE BEST LINEAR COMBINATION OF MARKERS THAT MAXIMIZES THE PARTIAL AREA UNDER THE ROC CURVE(PAUC)

ABSTRACT. When there are multiple markers associated with a disease, except the trivial case, a combination of these markers often performs better than individual ones. Here we focus on the class of linear combinations for an easy and clear interpretation. The AUC (area under the receiver operating characteristic curve) criterion proposed for evaluation of medical diagnostic tools nowadays becomes more and more popular in assessing the discriminating power of a binary classification rule with continuous-scale. However, in some real applications, only a limited region of specificity is of research interest, and hence the partial AUC(pAUC) is a more adequate criterion. The goal of this study is to find the best linear combination that maximizes the pAUC. Under the normality assumption, the partial derivative of the pAUC is obtained and has a complex form. Hence the finding of the maximizer(s) is a difficult task. In this study, an algorithm is developed. Intensive numerical studies are conducted for assessment of the algorithm. It's found that the algorithm has adequate empirical performance.

1. METHOD

Notation and Basic Results

Consider p biomarkers as diagnostic tools for a specific disease and a large value of the biomarker favors a positive diagnosis. Let X and Y be the vectors of p variables for non-diseased and diseased population, respectively. Suppose

$$X \sim MN(\mu_x, \Sigma_x), \;\; Y \sim MN(\mu_y, \Sigma_y),$$

where μ_x, $\mu_y \in R^p$ and Σ_x and Σ_y are $p \times p$ matrices. Let $a \in R^p$ be a vector of coefficients. Then, given a, the linear combinations $a^T Y, a^T Y$ follow

$$V_{\bar{D}} \equiv a^T X \sim N(a^T \mu_x, a^T \Sigma_x a) \tag{1.1}$$

$$V_D \equiv a^T Y \sim N(a^T \mu_y, a^T \Sigma_y a). \tag{1.2}$$

Define $\Delta_\mu = \mu_y - \mu_x$, $Q_x = a^T \Sigma_x a$ and $Q_y = a^T \Sigma_y a$. Let $F_{\bar{D}}$ and F_D be the distribution functions of $V_{\bar{D}}$ and V_D, and $S = 1 - F$. It follows that at fixed a, the ROC curve can be derived as

$$ROC(u; a) = S_D[S_{\bar{D}}^{-1}(u)] = 1 - \Phi\left[\frac{S_{\bar{D}}^{-1}(u) - a^T \mu_y}{\sqrt{Q_y}}\right] = 1 - \Phi\left[\frac{c(u)\sqrt{Q_x} - a^T \Delta_\mu}{\sqrt{Q_y}}\right] \tag{1.3}$$

1991 *Mathematics Subject Classification.*

Key words and phrases.

where u is some level of false positive rate and $c(u) = \Phi^{-1}(1-u)$. Su and Liu(1993) showed that the coefficients for the best linear combination are

$$a^* \propto \Sigma_x^{-1/2}(I + \Sigma_x^{-1/2}\Sigma_y\Sigma_x^{-1/2})^{-1}\Sigma_x^{-1/2}\Delta_\mu = (\Sigma_x + \Sigma_y)^{-1}\Delta_\mu,$$

and the maximal AUC is

$$\Phi\left(\sqrt{\Delta_m u^T(\Sigma_x + \Sigma_y)^{-1}\Delta_m u}\right).$$

Further, the partial area under ROC curve for a given $t \in (0,1)$ of the linear combination is given by

$$pAUC_t(a) = \int_0^t ROC(u;a)du = \int_0^t \left(1 - \Phi\left[\frac{c(u)\sqrt{Q_x} - a^T\Delta_\mu}{\sqrt{Q_y}}\right]\right)du. \tag{1.4}$$

Optimal Linear Combination

To find the best linear combination of biomarkers that maximizes the pAUC for a given t, it suffices to find a, such that

$$\frac{\partial pAUC(a)}{\partial a} = 0. \tag{1.5}$$

Theorem 1.1. *The coefficient vector of the best linear combination of the p biomarkers, a_0, is proportional to*

$$(w_1\Sigma_x + w_2\Sigma_y)^{-1}\Delta_\mu, \tag{1.6}$$

where

$$w_1 = c_1\frac{a_0^T\Delta_\mu}{Q_x + Q_y} + c_2Q_y, \quad w_2 = c_1\frac{a_0^T\Delta_\mu}{Q_x + Q_y} - c_2Q_x.$$

In which,

$$c_1 = \sqrt{2\pi}\sigma\Phi\left(\frac{\nu - c(t)}{\sigma}\right), \quad c_2 = \sigma^2\exp\left[-\frac{(c(t)-\nu)^2}{2\sigma^2}\right]\cdot\frac{1}{\sqrt{Q_x}Q_y},$$

further, $\nu = a_0^T\Delta_\mu\sqrt{Q_x}/(Q_x+Q_y)$, $\sigma^2 = Q_y/(Q_x+Q_y)$.

Proof. Let

$$A = \frac{a^T\Delta_\mu}{Q_y}\Sigma_y a - \Delta_\mu \text{ and } B = \frac{\Sigma_x a}{\sqrt{Q_x}} - \frac{\sqrt{Q_x}}{Q_y}\Sigma_y a,$$

and $\nu = a^T\Delta_\mu\sqrt{Q_x}/(Q_x+Q_y)$, $\sigma^2 = Q_y/(Q_x+Q_y)$. Then Eq. (1.5) can be shown to have the following form,

$$A\int_{c(t)}^\infty \exp\left[-\frac{(y-\nu)^2}{2\sigma^2}\right]dy + B\int_{c(t)}^\infty y\cdot\exp\left[-\frac{(y-\nu)^2}{2\sigma^2}\right]dy = 0. \tag{1.7}$$

Note that

$$\int_{c(t)}^{\infty} \exp\left[-\frac{(y-\nu)^2}{2\sigma^2}\right] dy = \sqrt{2\pi}\sigma\Phi\left(\frac{\nu - c(t)}{\sigma}\right),$$

and

$$\int_{c(t)}^{\infty} y \cdot \exp\left[-\frac{(y-\nu)^2}{2\sigma^2}\right] dy = \sigma^2 \exp\left[-\frac{(c(t)-\nu)^2}{2\sigma^2}\right] + \sqrt{2\pi}\nu\sigma\Phi\left(\frac{\nu - c(t)}{\sigma}\right).$$

It follows that (1.7) becomes

$$c_1(A + B\nu) + c_2 B\sqrt{Q_x}Q_y = 0, \tag{1.8}$$

where

$$c_1 = \sqrt{2\pi}\sigma\Phi\left(\frac{\nu - c(t)}{\sigma}\right), \quad c_2 = \sigma^2 \exp\left[-\frac{(c(t)-\nu)^2}{2\sigma^2}\right] \cdot \frac{1}{\sqrt{Q_x}Q_y}.$$

Because

$$A + B\nu = \frac{a^T\Delta_\mu}{Q_x + Q_y}(\Sigma_x + \Sigma_y)a - \Delta_\mu, \quad B\sqrt{Q_x}Q_y = Q_y\Sigma_x a - Q_x\Sigma_y a,$$

(1.8) becomes

$$c_1\left[\frac{a^T\Delta_\mu}{Q_x + Q_y}(\Sigma_x + \Sigma_y)a - \Delta_\mu\right] + c_2[Q_y\Sigma_x a - Q_x\Sigma_y a] = 0,$$

which implies that

$$c_1(\mu_y - \mu_x) = (w_1\Sigma_x + w_2\Sigma_y)a \tag{1.9}$$

where

$$w_1 = c_1\frac{a^T\Delta_\mu}{Q_x + Q_y} + c_2 Q_y, \quad w_2 = c_1\frac{a^T\Delta_\mu}{Q_x + Q_y} - c_2 Q_x.$$

It is known that the pAUC is invariant to the scale, so the best linear combination is

$$a_0 \propto (w_1\Sigma_x + w_2\Sigma_y)^{-1}(\mu_y - \mu_x).$$

For simplicity, the coefficient vector is restricted to have unit norm.

2. Algorithm

From Theorem 1.1, solving for the optimal coefficient vector is equivalent to a fixed-point problem,

$$a_0 = f(a_0).$$

We consider the following "naive" iterated algorithm:

Step 0. Calculate the coefficients a^* of the best linear combination wrt AUC,

$$a^* = (\Sigma_x + \Sigma_y)^{-1}\Delta_\mu.$$

Step 1. Use a^* as the initial $a_0^{(0)} = a^*$, compute the corresponding pAUC by (1.4). Denote the pAUC by $pAUC^{(0)}$.

Step 2. Calculate $f(a_0^{(0)})$. If $f(a_0^{(0)}) \cdot \Delta_\mu > 0$, then $a_0^{(1)} = f(a_0^{(0)})$. Otherwise, $a_0^{(1)} = -f(a_0^{(0)})$

Step 3. Normalized $a_0^{(1)}$ to have unit norm and compute the corresponding pAUC, $pAUC^{(1)}$.

Step 4. Calculate the increment $\delta^{(1)} = pAUC^{(1)} - pAUC^{(0)}$. Then

a. If $\delta^{(1)} < -\epsilon$, find the first two significant biomarkers according to their absolute magnitudes in $a_0^{(1)}$ and record as $b_{2\times 1}^{(1)}$. Also define $b^{(0)}$ from $a_0^{(0)}$. Find the angle between $b^{(0)}$ and $b^{(1)}$. Rotate $b^{(1)}$ from $b^{(0)}$ by the same angle in reverse direction. Renew $a_0^{(1)}$ by combining the rotated vector with the original $a_0^{(1)}$. Moreover, recalculate $pAUC^{(1)}$ and $\delta^{(1)}$.

b. If at the first stage, $|\delta^{(1)}| < \epsilon$, again find $b^{(1)}$ and $b^{(0)}$. Rotate $b^{(0)}$ counterclockwise by some θ_1. Renew $a_0^{(1)}$ by combine the rotated vector with the original $a_0^{(0)}$. Moreover, recalculate $pAUC^{(1)}, \delta^{(1)}$. Repeat 2-4.

- If $\delta^{(1)} > \epsilon$, repeat Step 2-4. Otherwise, stop and the let the last a_0 be the final solution.
- Note: Here $\epsilon = 10^{-8}, \theta_1 = \pi/8$.

For a simulated data set with an equi-correlation structure, the algorithm is robust to the selection of the initial value and all the convergence is monotone upward. The algorithm performs well with p up to 100. See Case 1. However, for the two examples in Liu et al.(2005), the convergent point of the algorithm is sensitive to the initial value and the monotone convergence is no longer present. Practically, we suggest using the coefficients a^* of the best linear combination for AUC by Su and Liu(1993) as the initial value. See the results of Case 2-3.

Case 1: Slightly positive equi-correlation

Data: $p = 100$

- Mean: $\mu_x = 0$, $\mu_y = (0, \cdots, 0, 1.02, 1.04, \cdots, 2)^T$. (50 effective biomarkers)
- Variance: $\Sigma_x = I_p$. In Σ_y, $\sigma_{i,i} = 1$, and $\sigma_{i,j} = \rho$, for $i \neq j$. Here $\rho = \pm 0.2$.
- The cutoff for pAUC, $t = 0.2$.
- $\epsilon = 10^{-6}$.

Results:

Initial	Iterations for convergence	pAUC
$V_{(1)}$	3	0.2
$V_{(2)}$	3	0.2
$V_{(p)}$	3	0.2

Case 2: Liu et al.(SIM, 2005)

Data: $p = 4$

- Mean: $\mu_x = (14.46, 23.89, 7.29, 17.68)^T, \mu_y = (15, 61, 25.83, 8.2, 19.21)^T$.
- Variance:

$$\Sigma_x = \begin{pmatrix} 1.88 & -3.20 & -2.10 & -1.85 \\ -3.20 & 13.33 & 3.31 & 7.39 \\ -2.10 & 3.31 & 4.67 & 4.10 \\ -1.85 & 7.39 & 4.10 & 10.33 \end{pmatrix}, \quad \Sigma_y = \begin{pmatrix} 13.56 & -7.28 & -5.79 & -6.95 \\ -7.28 & 23.23 & 7.12 & 6.60 \\ -5.79 & 7.12 & 5.86 & 3.85 \\ -6.95 & 6.60 & 3.85 & 13.71 \end{pmatrix}$$

- The cutoff for pAUC, $t = 0.2$.
- $\epsilon = 10^{-6}$.

Results:

1. Marginal pAUC: 0.0838, 0.0537,0.0421,0.0459.
2. Not robust to the initial value and the convergence is not stable.

Initial	Iterations	pAUC	a_0
$V_{(1)}$	3	0.1150	(0.884,0.217,0.373,-0.175)
$V_{(2)}$	2	0.0537	(0.000,1.000,0.000,0.000)
$V_{(4)}$	3	0.0454	(-0.096,0.081,-0.052,0.991)
a^*	3	0.1194	(0.880,0.217,0.417,-0.067)

Case 3: Coronary Heart Disease Example.(Liu et al(2005))

Data: $p = 4$

- Mean: $\mu_x = (0.1275, 0.8845, 4.0776, 6.7724)^T, \mu_y = (0.1402, 0.9337, 4.1225, 6.9112)^T$.

- Variance:

$$\Sigma_x = \begin{pmatrix} 0.0034 & -0.0004 & -0.0002 & -0.0051 \\ -0.0004 & 0.0285 & 0.0029 & 0.0417 \\ -0.0002 & 0.0039 & 0.0488 & 0.0268 \\ -0.0051 & 0.0417 & 0.0268 & 0.2846 \end{pmatrix}, \quad \Sigma_y = \begin{pmatrix} 0.0043 & -0.0004 & -0.0002 & -0.0051 \\ 0.0033 & 0.0415 & 0.0019 & 0.0426 \\ 0.0006 & 0.0019 & 0.0389 & 0.0010 \\ 0.0067 & 0.0426 & 0.0010 & 0.1504 \end{pmatrix}$$

- The cutoff for pAUC, $t = 0.2$.
- $\epsilon = 10^{-6}$.

Results:

1. Marginal pAUC: 0.0331,0.0392,0.0230,0.0176
2. Not robust to the initial value.

Initial	Iterations	pAUC	a_0
$V_{(1)}$	4	0.0480	(0.9475,0.3150,0.0431,0.0320)
$V_{(2)}$	3	0.0440	(0.8927,0.4395,-0.0992,-0.0029)
$V_{(4)}$	2	0.0230	(0.0000,0.0000,1.0000,0.0000)
a^*	3	0.0480	(0.9476,0.3150,0.0432,0.0321)
Su and Liu		0.0384	(1.4600,0.3400,0.4117,0.2216)
Su and Liu(scaled)			(0.9298,0.2165,0.2622,0.1411)
Liu et al.		0.0470(0.019?)	(-0.8436,-3.2269,0.2918,-0.1181)
Liu et al.(scaled)			(-0.2518,-0.9632,0.0871,-0.0352)
Liu et al.(change sign)		0.0402	

References

[1] Su, J. Q. and Liu, J. S.(1993) Linear combinations of multiple diagnostic markers, *Journal of the American Statistical Association*, **88**, 1350-1355.

[2] Liu, A., Schisterman, E.F. and Zhu, Y.(1995) On linear combinations of biomarkers to improve diagnostic accuracy, *Statistics in Medicine*, **24**, 37-47.

行政院國家科學委員會補助國內專家學者出席國際學術會議報告

98 年 8 月 24 日

附件三

報告人姓名	薛慧敏	服務機構及職稱	政治大學統計系
時間 會議 地點	98 年 8 月 2 日至 5 日 美國華盛頓特區	本會核定補助文號	NSC 98-2118-M-004-004
會議名稱	(中文)美國統計研討會 (英文)2009 Joint Statistical Meetings		
發表論文題目	(中文) 針對兩卜瓦松分佈平均數的比較問題之非條件確實統計檢定方法 (英文) Unconditional exact test for comparison of two Poisson means		

報告內容應包括下列各項：

一、參加會議經過

第一天為報到程序。之後三天則為參加研討會議。每一天有四個場次，各有多個演講廳同時進行會議。另外在每天的上午與下午，在安排的場地上皆有海報發表。本人在8/4 下午發表海報文章，在其他時間則至各演講廳，聆聽與會學者的文章發表，學習目前最新學術發展，並適時參與討論。

二、與會心得

本屆大會有約 5000 人與會參加。有來自統計學門與資訊工程學門專家、學者踴躍參與。此會議提供很好的機會讓不同領域的學者能夠在統計計算與計算統計上的理論、方法與實際運用上交流與分享。本人在此次會議上有豐富收穫。

三、考察參觀活動(無是項活動者省略)

無。

四、建議

近年來，由於經濟因素，至歐美參加會議之旅費與生活費逐年高漲，國科會補助通常不敷使用，建議能因應客觀環境，適當提高補助經費，或增加計畫經費流用的彈性，才能提高參加會議意願。

五、攜回資料名稱及內容

包括書面資料一冊與光碟片一份。其中書面資料為會議議程，光碟片則為發表文章摘要。

六、其他

無。

無衍生研發成果推廣資料

98 年度專題研究計畫研究成果彙整表

計畫主持人：薛慧敏	計畫編號：98-2118-M-004-004-
計畫名稱：接受者操作特徵函數線下面積之無母數迴歸分析	

成果項目			量化			單位	備註（質化說明：如數個計畫共同成果、成果列為該期刊之封面故事...等）
			實際已達成數（被接受或已發表）	預期總達成數(含實際已達成數)	本計畫實際貢獻百分比		
國內	論文著作	期刊論文	1	1	100%	篇	
		研究報告/技術報告	0	0	100%		
		研討會論文	0	0	100%		
		專書	0	0	100%		
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（本國籍）	碩士生	1	1	100%	人次	
		博士生	1	1	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		
國外	論文著作	期刊論文	1	1	20%	篇	
		研究報告/技術報告	0	0	100%		
		研討會論文	1	1	100%		
		專書	0	0	100%	章/本	
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（外國籍）	碩士生	0	0	100%	人次	
		博士生	0	0	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		

其他成果 (無法以量化表達之成果如辦理學術活動、獲得獎項、重要國際合作、研究成果國際影響力及其他協助產業技術發展之具體效益事項等，請以文字敘述填列。)	無。

	成果項目	量化	名稱或內容性質簡述
科教處計畫加填項目	測驗工具(含質性與量性)	0	
	課程/模組	0	
	電腦及網路系統或工具	0	
	教材	0	
	舉辦之活動/競賽	0	
	研討會/工作坊	0	
	電子報、網站	0	
	計畫成果推廣之參與（閱聽）人數	0	

國科會補助專題研究計畫成果報告自評表

請就研究內容與原計畫相符程度、達成預期目標情況、研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）、是否適合在學術期刊發表或申請專利、主要發現或其他有關價值等，作一綜合評估。

1. 請就研究內容與原計畫相符程度、達成預期目標情況作一綜合評估

■達成目標

□未達成目標（請說明，以 100 字為限）

□實驗失敗

□因故實驗中斷

□其他原因

說明：

2. 研究成果在學術期刊發表或申請專利等情形：

論文：□已發表 □未發表之文稿 ■撰寫中 □無

專利：□已獲得 □申請中 ■無

技轉：□已技轉 □洽談中 ■無

其他：(以 100 字為限)

3. 請依學術成就、技術創新、社會影響等方面，評估研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）(以 500 字為限)

學術成就:提供適當統計方法於合併數個特徵變數在診斷或分類問題上。

技術創新:可應用於臨床實驗診斷資料的統計分析上。

社會影響:有利於國家生物科技相關產業發展。