مجلهٔ پژوهشی دانشگاه اصفهان (علوم پایه)
جلد سی و سوم ـ شماره ٤ ـ سال ١٣٨٧
صص ٢٦-١٥

تحلیل ممیزی غیرخطی انعطاف‌پذیر با استفاده از رگرسیون ناپارامتری

محسن محمدزاده* و علیرضا هومان**

*** گروه آمار، دانشگاه تربیت مدرس**

**** گروه ریاضی، دانشگاه ولی عصر (عج) رفسنجان**

چکیده

تحلیل ممیزی برای رده‌بندی یک شی یا گروهی از اشیاء به یکی از دو یا چند گروه متمایز معلوم یا نامعلوم مورد استفاده قرار می‌گیرد. در پژوهش‌های علمی برای رده‌بندی، اغلب از توابع ممیز خطی یا درجه دوم فیشر استفاده می‌شود. اما در این مقاله روشهایی غیرخطی براساس دو روش رگرسیون ناپارامتری تحت عناوین اسپلاین رگرسیونی انطباقی چند متغیره و مدل انطباقی جمعی برای رده‌بندی گروه‌ها معرفی شده و با استفاده از یک مطالعه شبیه‌سازی نحوه بکارگیری آنها بررسی و متوسط نرخ خطای آنها با روشهای متداول مورد مقایسه قرار گرفته است.

واژه‌های کلیدی: تحلیل ممیزی، تحلیل ممیزی انعطاف‌پذیر، رگرسیون ناپارامتری.

Flexible Discriminant Analysis via Nonparametric Regression

M. Mohammadzadeh* and A. Hooman**

*** Statistics Department ,Tarbiat Modares University**

**** Mathematic Department, Vali-Asr University**

Abstract

Discriminant analysis is a way for classification of one object or a group to one or more separate groups that are known or unknown. In scientific researches we often use linear or quadratic functions for classification. But in this paper, we suggest a nonlinear discrimination method that uses two nonparametric regression methods, namely multivariate adaptive regression splines and adaptive additive model. In a simulation study, we investigate the application way of proposed methods and comparing them with the ordinary nonlinear discrimination methods via their means of error rates.

Keywords: Discriminate Analysis, Flexible Discriminate Analysis, Nonparametric Regression.

مقدمه

پیدایش تحلیل رده‌بندی سابقه‌ای طولانی دارد، اما اولین تعریف روشن از این مسئله توسط فیشر(۱۹۳۶) ارائه شده است، که تابعی خطی برای رده‌بندی معرفی کرد. همچنین عامل به کارگیری تابع غیرخطی در جمعیت‌های چند متغیره نیز از اندیشه وی سرچشمه گرفته است. بعد از آن کارهای نیمن- پیرسن(۱۹۳۳) در استنباط آماری، شیوه‌های رده‌بندی را تحت تاثیر زیادی قرار داد. ولچ(۱۹۳۹) و والد(۱۹۴۴) رهیافت بیزی را برای این مساله مورد مطالعه قرار دادند. اخیراً تحولاتی در شیوه‌های رده‌بندی غیرخطی توسط هیستی، تیب شیرانی و بوجا(۱۹۹۴) و فریدمن(۱۹۹۱) انجام پذیرفته که منجر به ارائه شیوه‌هایی نوین در این مساله تحت عنوان رده‌بندی انعطاف‌پذیر شده است.

برای رده‌بندی مجموعه‌ای از داده‌ها به گروه‌های مختلف، معیاری مناسب مورد نیاز است. فرض کنید فضای اندازه پذیر χ شامل C گروه مجزای $R_1, R_2, \mathbf{L}, R_C$ باشد و داده‌ها به صورت زوج (j, X) که در آن $j \in \Im = \{1,2, \mathbf{L}, C\}$ برچسب هر گروه و X بردار متغیرهای Mبعدی اندازه‌گیری شده در گروه jام را نشان دهد. یک معیار مناسب می‌تواند تابعی از χ به $\{R_1, R_2, \mathbf{L}, R_C\}$ باشد، که احتمال رده‌بندی نادرست را کمینه کند. در تحلیل ممیزی خطی کلاسیک، برای رده‌بندی فضای نمونه معمولاً فرض بر آن است که متغیرهای مورد مطالعه نمونه‌های مستقلی به صورت (J, X) هستند، که در آن برای $j \in \Im$، توزیع X به شرط $J = j$ نرمال چند متغیرهٔ $N(m_j, \Sigma)$ است. در این صورت با فرض احتمالات یکسان $P(J = j) = \frac{1}{C}, j \in \Im$ برای هر گروه قاعدهٔ رده‌بندی X به گروه R_j با شرط کمینه کردن احتمال رده‌بندی نادرست به صورت:

$$X \in R_j \Leftrightarrow (X - m_j)^T \sum^{-1} (X - m_j) = \min_i (X - m_i)^T \sum^{-1} (X - m_i)$$

بیان می‌شود، که در آن اگر Σ نامعلوم باشد، ماتریس کواریانس ادغام شده[1] درون گروه‌ها یعنی Σ_P جایگزین آن می‌شود(ماردیا و همکاران، ۱۹۷۹). در این صورت با تبدیل $X' = \sum_P^{-\frac{1}{2}} X$، معیار بالا به صورت:

$$X' \in R_j \Leftrightarrow \left\| X' - m_j' \right\|^2 = \min_i \left\| X' - m_i' \right\|^2 \quad (۱)$$

بیان می‌شود، که در آن $\|\cdot\|$ فاصله در فضای اقلیدسی $E^{(M)}$، M تعداد متغیرها و $m_j' = \sum_P^{-\frac{1}{2}} m_j$

1- Pooled Covariance

است. بدون آنکه به کلیت مساله خدشه‌ای وارد شود، می‌توان $\sum_{j=1}^{C} m_j$ را برابر با صفر قرار داد. در این صورت بردارهای نرمال متعامد $a_1, a_2, \mathbf{L}, a_{C-1}$ در فضای $E^{(M)}$ را می‌توان به گونه‌ای یافت که بسط فضای خطی تولید شده توسط $\{m'_1, m'_1, \mathbf{L}, m'_C\}$ باشند. در این صورت قاعدهٔ رده‌بندی(۱) با کمینه کردن $\sum_{i=1}^{C-1}[a_i X' - a_i m'_i]$ هم ارز می‌شود. بنابراین تابع $C-1$ بعدی $Y(X) = (a_1 X', a_2 X', \mathbf{L}, a_{C-1} X')$ یک تصویر خطی از $E^{(M)}$ به فضای $E^{(C-1)}$ است و قاعدهٔ رده‌بندی بصورت

$$X \in R_j \Leftrightarrow \left\| Y'(X) - \bar{Y}_j \right\|^2 = \min_i \left\| Y(X) - \bar{Y}_i \right\|^2$$

حاصل می‌شود، که در آن $\bar{Y}_j$ میانگین گروه jام است. اما به دلیل آنکه تصاویر $Y(X)$ فقط به حالتهای خطی محدود می‌شوند، حالتهایی را که توزیع داده‌ها بر شرایط کلاسیک منطبق نیستند را شامل نمی‌شوند. برای رفع این محدودیت بریمن و ایهاکا(۱۹۸۴) تبدیلات کلی

$$Y_k(X) = \sum_{m=1}^{M} y_{km}(X_m) \quad k \geq 0$$

را در نظر گرفتند، که در آن y_{km} به تابعی خاص محدود نشده و با یک روش هموارسازی و بکاربردن رگرسیون امیدهای شرطی متناوب[1](ACE) به دست می‌آید. در این روش با انجام رگرسیون خطی چند متغیره برای ماتریس مشاهدات، بردار متغیر پاسخ تخمین زده می‌شود، سپس تحلیل ممیزی خطی بر تخمین بردارهای متغیر پاسخ به روش مینیمم فاصله(۱) انجام می‌پذیرد.

هیستی و همکاران(۱۹۹۳) این روش ممیزی را تعمیم داده و با استفاده از روشهای مختلف رگرسیون ناپارامتری مانند ACE و تعاقب تصویر[2](PPR) شیوه‌هایی غیرخطی برای رده‌بندی ارائه و آنها را تحلیل ممیزی انعطاف پذیر[3] نامیدند. فریدمن(۱۹۹۱) دو روش رگرسیون ناپارامتری تحت عناوین اسپلاین رگرسیونی انطباقی چند متغیره[4](MARS) و مدل انطباقی جمعی[5](AAM) را ارائه نمود، که برای مدل‌سازی‌های غیرخطی از آنها استفاده می‌شود. در مقاله حاضر، به منظور توسعه روش‌های تحلیل ممیزی غیرخطی، بکارگیری دو روش MARS و AAM بجای رگرسیون‌های ناپارمتری ACE و PPR برای رده‌بندی گروه‌ها پیشنهاد می‌شود. سپس در یک مطالعه شبیه‌سازی، نحوه استفاده از این روشها و همچنین میزان دقت رده‌بندی‌های حاصل از آنها با دیگر روش‌های معمول رده‌بندی خطی، درجه دو، ACE و PPR مورد مقایسه قرار گرفته و موارد برتری روش‌های پیشنهاد شده در این مقاله تعیین می‌شوند. برای این منظور دو روش رگرسیون ناپارامتری MARS و AAM بطور مختصر در بخش ۲ معرفی شده، سپس نحوه مقیاس گذاری بهینه برای ارتباط بین رگرسیون و تحلیل ممیزی در بخش ۳ و روش تحلیل ممیزی غیرخطی در بخش ۴ ارائه شده‌اند. بخش ۵ شامل مطالعه شبیه‌سازی برای مقایسه دقت رده‌بندیهای

1- Alternating Conditional Expectation

2- Projection Pursuit Regression
3- Flexible Discriminate Analysis
4- Multivariate Adaptive Regression Spline
5- Adaptive Additive Model

مختلف است و نهایتاً بخش آخر به بحث و نتیجه‌گیری اختصاص یافته است.

رگرسیون ناپارامتری چند متغیره

در رگرسیون ناپارامتری چند متغیره، مجموعه‌ای از n بردار اندازه پذیر $p \geq 1$ بعدی $\{x_i \in R^p : i = 1,2,\mathbf{L},n\}$ به همراه اندازه‌های مربوط به متغیر پاسخ $\{Y_i : i = 1,2,\mathbf{L},n\}$ در اختیار است که از مدل

$$Y_i = f(x_i) + e_i \quad \text{i=1,2,}\mathbf{L}\text{,n} \qquad (۲)$$

پیروی می‌کنند، بطوریکه در آن e_iها عبارت باقیمانده و f تابعی نامعلوم است و با استفاده از مشاهدات نمونه‌ای برآورد می‌شود. بریمن و همکاران(۱۹۸۴) رگرسیون افراز بازگشتی[1] را بعنوان یک تقریب برای تابع f به صورت

$$\hat{f}(x) = \sum_{m=1}^{M} a_m B_m(x) \qquad (۳)$$

معرفی کردند، که در آن تابع پایه‌ای $B_m(\cdot)$ به صورت $B_m(x) = I_{u_m}(x)$، I تابع نشانگر و $\{u_m ; m = 1,2,\mathbf{L},M\}$ افرازی از فضای R^p است. یعنی u_mها زیر فضاهایی از R^p هستند که برای هر $u_i \cap u_j = f,\ \text{i} \neq \text{j}$ و $\bigcup_{m=1}^{M} u_m = R^p$. مجموعهٔ $\{a_m ; m = 1,2,\mathbf{L},M\}$ در مدل(۳) ضرایبی هستند که مقادیر آنها به روش کمترین مربعات برآورد می‌شوند. بدلیل ناپیوستگی مدل افراز بازگشتی در مرزهای نواحی u_m، توابع پایه‌ای $B_m(\cdot)$ پله‌ای هستند. لذا(واهبا، ۱۹۹۰) به جای آنها از اسپلاین مرتبه q با توابع پایه‌ای توانی به صورت

$$b_q(x - t) = [x - t]_+^q \qquad (۴)$$

استفاده نمود، که در آن $[a]_+$ بیانگر قسمت مثبت a است. بدیهی است تابع پله‌ای حالت خاصی از این توابع پایه‌ای است، که به ازاء $q = 0$ حاصل می‌شود. در روش MARS توابع پایه‌ای برای(۳) به شکل

$$B_m^q(x) = \prod_{k=1}^{K_m} \left[S_{km} \left(x_{v(k,m)} - t_{km} \right) \right]_+^q$$

انتخاب می‌شوند، که در آن $S_{km} \in \{-1,1\}$، K_m تعداد برشها، $x_{v(k,m)}$متغیر برش و t_{km} آستانه آن است و مقدار $q \geq 1$ بگونه‌ای تعیین می‌شود که توابع پایه‌ای پیوسته شوند. سپس توابع پایه‌ای اضافی به کمک فرآیند زدودن یک به یک حذف می‌شوند تا باعث بهتر شدن مدل و انتخاب یک مجموعهٔ معتبر نهایی برای برآورد مدل انتخابی شود.

بوجا و همکاران(۱۹۸۹) تقریب f در مدل(۲) را به صورت

$$\hat{f}(x) = \sum_{j=1}^{p} b_j^T h_j(x_{ij}) \qquad (۵)$$

در نظر گرفتند، که در آن h_j یک بردار Mتایی از توابع غیرخطی معلوم، x_{ij} iامین مشاهده برای jامین متغیر و b_jها بردارهای ضرایب رگرسیونی هستند که با کمینه کردن مجموع مربعات خطای جریمه‌ای شده[2]

$$\sum_{i=1}^{M} (y_i - \sum_{j=1}^{p} b_j^T h_j(x_{ij}))^2 + \sum_{j=1}^{p} l_j b_j^T d_j b_j \qquad (۶)$$

1- Recursive Partitioning Model

2- Penalized Sum of Square Errors

برآورد می‌شوند، بطوریکه در آن l_j ها پارامترهای هموواری و d_j ستون jام ماتریسی به نام ماتریس جریمه است. معمولا برای برآورد پارامترهای l_j، از معیار اعتبار متقابل تعمیم یافته[1](GCV) استفاده می‌شود. گا و واهبا(۱۹۸۴) الگوریتمی برای این منظور ارائه نموده‌اند، که در آن برآورد هر یک از l_j ها نیاز به محاسبه عبارتهایی از مرتبه $O(n^3)$ دارد. هیستی(۱۹۸۹) معیار GCV را تعدیل نموده و عبارتی از مرتبه $O(n)$ برای آن ارائه کرد و آنرا AAM نامید. محمدزاده(۱۹۹۸) بر اساس معیار هیستی پارامترهای هموواری λ_j را برای اسپلاین هموواری[2] برآورد کرد. چون عبارت مربوط به معیار GCV اغلب چند مدی است، کنت و محمدزاده(۲۰۰۰) نیز الگوریتمی برای بهینه‌سازی آن ارائه نمودند که قادر است پارامتر هموواری را در ماکسیمم مطلق معیار GCV برآورد نماید.

مقیاس‌گذاری بهینه

فیشر(۱۹۳۶) با استفاده از توابع $g_i(x)=w_i^T x+w_{i0},\ \text{i=1,2,}\mathbf{L}\text{,C}$ تحلیل ممیزی را با قاعده

$$x\in R_j \Leftrightarrow g_i(x)\geq g_j(x)\quad \forall \text{i}\neq\text{j}\in\Im$$

معرفی کرد، که در آن خطاها نسبت به بردارهای وزن w_i^T و آستانه‌های w_{i0} به روش حداقل مربعات کمینه می‌شوند. توابع ممیزی g_i را می‌توان با کدگذاری متغیر پاسخ و استفاده از رگرسیون خطی نیز برآورد کرد. برای این کار لازم است مربع خطاها نسبت به مقادیر نشانگر گروه و بردارهای وزن w_i کمینه شوند. اگر q_i تابع نشانگر گروه iام باشد، می‌توان با استفاده از مینیمم مجذور خطا، $K<C$ پاسخ مستقل برای q_iها، بصورت $q_1^*,q_2^*,\mathbf{L},q_K^*$ بدست آورد. اگر Θ یک ماتریس $C\times K$ بعدی با ستون‌های متشکل از q_iها و T ماتریس $n\times C$ بعدی نشانگر گروه‌ها با درایه‌های $T_{ij}=1$ برای $x_i\in R_j$ و صفر در جاهای دیگر تعریف شوند، آنگاه ماتریس $n\times K$ بعدی پاسخ‌های تبدیل یافته گروهها به فرم $\Theta^*=T\Theta$ خواهد بود. بدون آنکه خللی به کلیت مساله وارد شود می‌توان فرض کرد میانگین داده‌ها صفر است و X ماتریس $n\times p$ بعدی داده‌ها با iامین سطر x_i^T است. بنابراین لازم است عبارت

$$E=\|\Theta-XW^T\|^2 \qquad (۷)$$

نسبت به مقیاس‌های Θ و ماتریس وزن $K\times p$ بعدی W کمینه شود. جواب(۷) بصورت $W^T=X^-\Theta^*$ است، که در آن X^- معکوس تعمیم یافته پن‌روز[3] X است. اکنون، با جایگذاری W در(۷) داریم:

$$E=tr[\Theta^T T^T(I-XX^-)T\Theta] \qquad (۸)$$

که کمینه کردن آن نسبت به Θ با اعمال قید $\Theta^T D\Theta=I_K$ یک ماتریس همانی $K\times K$ بعدی و $D=T^TT/n$) به روش لاگرانژ چندگانه، باعث می-

1- Generalized Cross Validation
2- Smoothing Spline
3- Penrose Generalized Inverse

شود که ستون‌های ماتریس Θ در معادله بردار ویژه متقارن عمومی

$$\frac{1}{n}T^T(XX^-)Tq = l^T Dq \qquad (۹)$$

صدق کند. ماتریس حاصلضرب $(XX^-)T$ در(۹)، مقادیر برازش شده $\hat{T}$، در یک رگرسیون T روی X است. بنابراین، ماتریس طرف چپ(۹) حاصلضرب مقادیر هدف و برازش شده، یعنی $T^T\hat{T}$ می‌باشد. برای داده‌های مرکزی شده، ماتریس کواریانس بین گروه ها به صورت

$$S_B = \sum_{i=1}^{C}\frac{n_i}{n}m_i m_i^T = \frac{1}{n}X^T T(T^T T)^- T^T X \quad (۱۰)$$

است، که در آن n_i و m_i به ترتیب حجم و میانگین گروه iام هستند. با استفاده از(۱۰) می‌توان نشان داد که ماتریس کواریانس بین گروه‌ها در یک فضای تبدیل‌یافته به صورت

$$WS_B W^T = diag\{l_1^2, l_2^2, \mathbf{L}, l_K^2\}$$

است، که در آن $l_1, l_2, \mathbf{L}, l_K$ ویژه مقدارهای متناظر با $q_1, q_2, \mathbf{L}, q_K$ هستند که به صورت صعودی مرتب شده‌اند. به طور مشابه ماتریس کواریانس درون گروه‌ها نیز به صورت

$$WS_W W^T = D_{l(1-l)} =$$
$$diag\{l_1(1-l_1), l_2(1-l_2),..., l_K(1-l_K)\}$$

تبدیل می‌شود. بنابراین W هر دو ماتریس کواریانس S_B و S_W را قطری می‌کند. به ویژه، در تحلیل ممیزی خطی می‌توان به جای W، عبارت $W_{LDA} = D_{l(1-l)}^{\frac{1}{2}}W$ را نیز بکار برد. در اینصورت با استفاده از ارتباط بین پاسخ تحلیل ممیزی خطی و پاسخ مقیاس گذاری بهینه، تحلیل ممیزی به صورت

$$x \in R_j \Leftrightarrow g_i \geq g_j \quad \forall i \neq j \in \Im$$

می‌باشد، که در آن

$$g_i = \log(p(R_i)) - \frac{1}{2}(y_i - y_{OS}(x))^T D_{l(1-l)}^{-1}(y_i - y_{OS}(x)),$$

است، بطوریکه $p(R_i)$ احتمال تعلق x به iامین گروه، y_i بردار تبدیل‌یافته میانگین iامین گروه و $y_{OS}(x) = Wx$ می‌باشد. هیستی و همکاران(۱۹۹۵) فرم دیگری برای g_i بصورت

$$g_i = \log(p(R_i)) - \frac{1}{2}(q^i - y_{OS}(x))^T D_{1-l}^{-1}(q^i - y_{OS}(x)) - \|q^i\|^2$$

ارائه کرده‌اند، که در آن

$D_{1-l} = diag\{1-l_1, 1-l_2, \mathbf{L}, 1-l_K\}$ و q^i بردار مقیاسها روی گروه R_i، یعنی iامین ستون Θ^i است. بطور معادل می‌توان g_i را بصورت

$$g_i = \log(p(R_i)) - \frac{1}{2}(q^i - y_{OS}(x))^T D_{e^2}^{-1}(q^i - y_{OS}(x)) - \|q^i\|^2$$

نیز نوشت، که در آن $D_{e^2} = diag(e_1^2, e_2^2, \mathbf{L}, e_K^2)$ و $e_i^2 = 1-\lambda_i$ سهم مقیاس θ_i در میانگین توان دوم خطاها است. بنابراین روش رگرسیون بر پایه مقیاسها به همراه یک تبدیل خطی منجر به قاعده ممیزی هم ارز تحلیل ممیزی خطی می‌شود.

تحلیل ممیزی غیرخطی

در صورتی که در مسأله تحلیل ممیزی برای C گروه، $q(Y)$ متغیر پاسخ تبدیل شده توسط رگرسیون ناپارامتری باشد، بریمن و ایهاکا(۱۹۸۴) نشان دادند که مانند حالت خطی تعداد $K \leq C-1$ جواب مستقل برای مقیاس‌ها وجود دارد که از طریق حل یک معادله

بردار ویژه تعیین می‌شوند. بعلاوه برای هر ستون Θ^* یک مدل جمعی برازش داده‌اند، به طوری که عبارت

$$e_k^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\Theta_{ik}^* - \sum_{j=1}^{p} f_{kj}(x_{ij})\right]^2, \qquad k = 1,2,\mathbf{L},K$$

$(k = 1,2,\mathbf{L},K;\ j = 1,2,\ \mathbf{L},\ p)$ f_{kj} را نسبت به توابع با اجرای رگرسیون ACE تا رسیدن به همگرایی کمینه می‌کند. در این صورت i امین عنصر بردار $y_{OS}(x)$ توسط عناصر بردار

$$(\sum_{j=1}^{p} f_{1j}(x_{ij}),\ \sum_{j=1}^{p} f_{2j}(x_{ij}),\ \mathbf{L},\ \sum_{j=1}^{p} f_{kj}(x_{ij}))$$

برآورد می‌شود و قاعده رده‌بندی

$$x \in R_j \Leftrightarrow d_i \geq d_j \qquad \forall j \neq i$$

را با

$$d_i = \log(p(R_i)) - \frac{1}{2}(q^i - \hat{f}(x))^T D_{e^2}^{-1}(q^i - \hat{f}(x)) - \| q^i \|^2$$

بدست آوردند، که در آن $\hat{f}$ تبدیلی غیرخطی با k امین عنصر $\sum_{j=1}^{p} f_{kj}(x_{ij})$ $(k = 1,2,\mathbf{L},K)$ و q^i بردار مقیاس‌ها برای گروه R_i است.

باید توجه نمود که برای هر مقیاس، مدل جمعی متفاوتی وجود دارد. این امر ممکن است برای داده‌های با تعداد متغیر زیاد، با محاسبات زیاد همراه بوده یا باعث برازش نادرست شود، مگر آنکه مدل‌های جمعی متغیرهای مورد استفاده به نوعی خود افراز شده باشند. اما در عمل با موارد زیادی مواجه می‌شویم که تعداد متغیرها زیاد بوده و لزوماً هیچگونه شواهدی مبنی بر افراز شده بودن آنها در اختیار نمی‌باشد. لذا در این مقاله روشهای ناپارامتری MARS و AAM معرفی شده در بخش ۲ برای تحلیل ممیزی غیرخطی مورد بررسی قرار می‌گیرند.

تحلیل ممیزی غیرخطی انعطاف‌پذیر

در این بخش یک روش تحلیل ممیزی غیرخطی با استفاده از روشهای رگرسیونی MARS و AAM ارائه می گردد که مراحل اجرای آن به شرح زیر است:

۱- به ماتریس Θ_0 یک مقدار اولیه با قید $K < C$ و $D = Y^T Y / n$ داده شود، که در آن $\Theta^T D \Theta = I$ است.

۲- با قراردادن $\Theta_0^* = Y^T \Theta_0$ و استفاده از روشهای رگرسیونی ناپارامتری AAM و MARS مدل را برازش داده و مقادیر برازش شده $\hat{\Theta}_0^*$ محاسبه شوند.

۳- با فرض آنکه $S(\hat{l})$ عملگری باشد که مدل انتخابی نهایی را برازش می‌دهد، ماتریس ویژه بردارهای $\Theta_0^{*T}\hat{\Theta}_0^* = \Theta_0^{*T} S(\hat{l})\Theta_0^*$ را تعیین و با $\Phi = (f_{ij})$ نمایش دهید. در اینصورت مقیاسهای بهینه به فرم $\Theta = \Theta_0 \Phi$ خواهند بود. توجه شود که عملگر $S(\hat{l})$ بسته به انتخاب پارامتر همواری $\hat{\lambda}$ می‌تواند خطی یا غیرخطی باشد و $h(x) = W^T x$ بردار توابع رگرسیون برازش شده است.

۴- با استفاده از مقیاس‌های بهینه، مدل نهایی مرحله دوم بصورت $h(x) \equiv \Phi^T h(x)$ بهنگام شود.

۵- برای C گروه از داده‌ها، بردار توابع $h(x)$

۶- می‌تواند حداکثر $K = C - 1$ مولفه داشته باشد. در این صورت تحلیل ممیزی به صورت

$$x \in R_j \Leftrightarrow d(x,j) = \| D(h(x) - \bar{h}^j) \|^2 = \min_i \| D(h(x) - \bar{h}^i) \|^2$$

تبدیل می‌شود، که در آن $\bar{h}^j = \sum_{g_i = j} h(x_i)/n_j$ میانگین گروه jام است و ماتریس قطری D برازشهای مقیاس‌گذاری بهینه را به متغیرهای تحلیل ممیزی تبدیل می‌کند.

مطالعه شبیه‌سازی

بریمن و همکاران(۱۹۸۴) روشهای تحلیل ممیزی خود را بر اساس رگرسیون خطی(LDA) و درجه دو(QDA)، همچنین دو روش رگرسیون ناپارامتری ACE و PPR بر روی سه نوع داده شبیه سازی شده بکار گرفته و آنها را مورد بررسی قرار دادند، سپس هیستی و همکاران(۱۹۹۳) نیز روش تحلیل ممیزی تعمیم‌یافته خود را با استفاده از همین سه نوع داده مورد ارزیابی قرار دادند. در این بخش ما نیز برای تحلیل ممیزی بر اساس رگرسیون های ناپارامتری AAM و MARS از همین داده‌ها استفاده می کنیم تا بتوان کارایی آنها را با یکدیگر و با روشهای دیگر LDA، QDA، ACE و PPR مورد مقایسه قرار داد.

داده‌های موجی: این داده‌ها سه گروه مستقل، هر یک دارای ۲۱ متغیر تصادفی بصورت:

$$\begin{cases} X_i = Uh_1(i) + (1-U)h_2(i) + e_i \\ X_i = Uh_1(i) + (1-U)h_3(i) + e_i \quad \text{i=1,K,21} \\ X_i = Uh_2(i) + (1-U)h_3(i) + e_i \end{cases}$$

هستند، که در آنها U یک متغیر تصادفی با توزیع یکنواخت $U(0,1)$، e_iها مستقل و دارای توزیع نرمال استاندارد و h_iها نیز بصورت زیر در نظر گرفته شده‌اند:

$$h_1(i) = \max(6 - |i - 11|, 0), \quad h_2(i) = h_1(i-4), \quad h_3(i) = h_1(i+4)$$

داده‌های با اثر متقابل: این داده‌ها دو گروه مستقل از توزیع یکنواخت دو متغیره بر مربع $[-1,1]\times[-1,1]$، انتخاب شده‌اند. گروه اول شامل مشاهداتی است که از گوشه‌های شمال شرقی و جنوب غربی و گروه دوم شامل مشاهداتی است که از گوشه‌های شمال غربی و جنوب شرقی مربع انتخاب شده‌اند.

داده‌های خوشه‌ای: دراین حالت از دو گروه، هر کدام شامل چهار متغیر استفاده شده است. متغیرهای گروه اول مستقل و دارای توزیع نرمال $N(3.5,1)$ هستند، متغیرهای گروه دوم نیز مستقل از یکدیگر ولی دارای توزیع نرمال استاندارد و مستقل از متغیرهای گروه اول هستند.

برای ارزیابی روشهای تحلیل ممیزی خطی LDA، درجه دوم QDA، تحلیل‌های ممیزی غیرخطی ACE و PPR و همچنین تحلیل‌های ممیزی غیرخطی انعطاف پذیر AAM و MARS، از هر سه نوع داده موجی، با اثر متقابل و خوشه‌ای، نمونه‌های آموزشی به حجم‌های ۵۰، ۱۵۰ و ۳۰۰ با احتمال‌های یکسان برای هر گروه تولید شده‌اند.

نمونه تست که به صورت تصادفی به هر یک از گروه‌ها تعلق دارد تولید و صحت رده‌بندی آن بررسی شده است. این کار ۱۰۰۰ بار تکرار شده و متوسط احتمال خطای رده‌بندی روشهای مختلف محاسبه و در جدول ۱ خلاصه شده‌اند. مقایسه احتمال‌های خطای رده‌بندی روش‌های مختلف، برای داده‌های موجی و

حجم‌های مختلف نمونه‌های آموزشی، بیانگر آنست که روش خطی از تمام روشها بجز روش های PPR و AAM عملکرد بهتری دارد.هرچند این دو روش از احتمال خطای رده‌بندی کوچکتری نسبت به روش خطی برخوردار هستند اما اختلاف احتمال خطای رده‌بندی آنها ناچیز و در حد چند هزارم است. اما برای داده‌های با اثر متقابل و خوشه‌ای، همواره روش‌های غیرخطی از احتمال خطای رده‌بندی کوچکتری نسبت به روش خطی برخوردارند و بخصوص احتمال خطای رده‌بندی تحلیل های ممیزی غیرخطی انعطاف پذیر AAM و MARS، نه تنها تفاوت فاحشی(بیشتر از ۰/۴) با احتمال‌های خطای رده‌بندی روش خطی دارند بلکه از هر دو روش غیرخطی ACE و PPR نیز عملکرد بسیار بهتری دارند.

جدول ۱: احتمال خطای رده‌بندی روش‌های مختلف تحلیل ممیزی برای سه نوع داده

حجم نمونه آموزشی	روش‌های تحلیل ممیزی	نوع داده‌ها		
		موجی	با اثر متقابل	خوشه‌ای
50	LDA	0.195	0.475	0.568
	QDA	0.216	0.377	0.133
	ACE	0.236	0.083	0.122
	PPR	0.189	0.060	0.096
	MARS	0.215	0.059	0.071
	AAM	0.192	0.052	0.061
100	LDA	0.192	0.473	0.538
	QDA	0.215	0.357	0.133
	ACE	0.234	0.081	0.124
	PPR	0.187	0.057	0.097
	MARS	0.213	0.052	0.073
	AAM	0.190	0.050	0.059
300	LDA	0.191	0.472	0.501
	QDA	0.216	0.337	0.133
	ACE	0.232	0.081	0.122
	PPR	0.186	0.055	0.096
	MARS	0.211	0.050	0.072
	AAM	0.188	0.047	0.056

برای بررسی تأثیر توأم حجم نمونه‌های آموزشی و تست بر دقت تحلیل ممیزی برای نمونه آموزشی ۵۰تایی نمونه‌های تست به حجم‌های ۵، ۱۰ و ۲۰ که به صورت تصادفی به هر یک از گروه‌ها تعلق دارند، تولید شده‌اند. برای هر یک از نمونه‌ها نرخ خطا، یعنی حاصل نسبت تعداد رده‌بندی‌های نادرست به تعداد کل داده‌ها، برای تمام روش‌های تحلیل ممیزی محاسبه شده‌اند.

این کار را ۱۰۰۰ بار تکرار نموده و متوسط نرخ خطای روش‌های مختلف در جدول ۲ نمایش داده شده‌اند. همین کار برای نمونه آموزشی ۱۵۰ تایی با

نمونه‌های تست به حجم‌های ۳۰، ۵۰ و ۱۰۰ و نمونه آموزشی ۳۰۰ تایی با نمونه‌های تست به حجم‌های ۵۰، ۱۰۰ و ۱۵۰ انجام شده و متوسط نرخ خطای روش‌های مختلف به جدول ۲ اضافه شده‌اند.

همانگونه که در جدول ۲ ملاحظه می‌شود، افزایش حجم نمونه‌های آموزشی و تست همواره موجب کاهش نرخ خطای رده‌بندی و در نتیجه افزایش دقت تمام روش‌های تحلیل ممیزی برای انواع داده‌ها می‌شود که نتیجه‌ای قابل انتظار است.

بعلاوه برای داده‌های موجی متوسط نرخ خطای روش تحلیل ممیزی خطی از متوسط نرخ خطای روشهای غیرخطی دیگر بیشتر نیست و با متوسط نرخ خطای روش غیرخطی انعطاف پذیر AMM یکسان می‌باشد.اما برای داده‌های با اثر متقابل و خوشه‌ای همواره متوسط نرخ خطای روش‌های خطی تفاوت بسیار زیادی با متوسط نرخ خطای سایر روشها به خصوص روشهای غیرخطی دارد.

روش تحلیل ممیزی درجه دوم برای داده‌های با اثر متقابل کمترین متوسط نرخ خطا را در بین روش های تحلیل ممیزی خطی دارا است. برای داده‌های خوشه‌ای همواره روش‌های تحلیل ممیزی انعطاف پذیر نسبت به روش‌های خطی و درجه دوم از متوسط نرخ خطای کمتری برخوردار هستند، اما همانطور که در جدول ۲ ملاحظه می‌شود روش‌های تحلیل ممیزی انعطاف پذیر AAM و MARS که در این مقاله معرفی شده‌اند کمترین متوسط نرخ خطا را در مقایسه با سایر روش‌ها دارند. بعلاوه متوسط نرخ خطای روش AAM برای دو نوع داده موجی و خوشه‌ای حتی از روش MARS به ازای تمام حجم نمونه‌های آموزشی و تست که مورد بررسی قرار گرفته‌اند کمتر است.

بحث و نتیجه‌گیری

برای داده‌های موجی نرخ خطای روش تحلیل ممیزی خطی LDA از نرخ خطای روش‌های غیرخطی دیگر بیشتر نیست و عملکرد آن با روش غیرخطی انعطاف پذیر AMM یکسان می‌باشد. اما برای داده‌های با اثر متقابل و خوشه‌ای همواره روش‌های خطی عملکرد نامناسبی در مقایسه با سایر روش‌ها دارند. روش تحلیل ممیزی درجه دوم برای داده‌های با اثر متقابل بهترین عملکرد را نسبت به روش‌های تحلیل ممیزی خطی از خود نشان می‌دهد.

برای داده‌های خوشه‌ای همواره روش‌های تحلیل ممیزی انعطاف‌پذیر نسبت به روش‌های خطی و درجه دوم از عملکرد بهتری برخوردار هستند، اما روش‌های تحلیل ممیزی انعطاف پذیر AAM و MARS که در این مقاله معرفی شده‌اند عمکرد بسیار بهتری در مقایسه با سایر روش‌های بررسی شده دارند.

جدول ۲: متوسط نرخ خطای روشهای مختلف تحلیل ممیزی برای سه نوع داده مختلف

حجم نمونه آموزشی	حجم نمونه تست	روشهای تحلیل ممیزی	نوع داده‌ها		
			موجی	با اثر متقابل	خوشه‌ای
50	5	LDA	0.207	0.490	0.488
		QDA	0.225	0.390	0.390
		ACE	0.248	0.093	0.092
		PPR	0.203	0.072	0.071
		MARS	0.226	0.071	0.070
		AAM	0.205	0.066	0.065
	10	LDA	0.203	0.487	0.486
		QDA	0.224	0.370	0.369
		ACE	0.243	0.090	0.090
		PPR	0.198	0.070	0.069
		MARS	0.223	0.063	0.062
		AAM	0.202	0.061	0.060
	20	LDA	0.202	0.484	0.483
		QDA	0.224	0.353	0.352
		ACE	0.242	0.090	0.089
		PPR	0.195	0.064	0.063
		MARS	0.223	0.061	0.060
		AAM	0.201	0.058	0.057
150	30	LDA	0.201	0.478	0.572
		QDA	0.222	0.380	0.137
		ACE	0.242	0.086	0.126
		PPR	0.195	0.063	0.100
		MARS	0.223	0.062	0.075
		AAM	0.201	0.055	0.061
	50	LDA	0.198	0.479	0.542
		QDA	0.221	0.363	0.137
		ACE	0.240	0.087	0.128
		PPR	0.193	0.062	0.101
		MARS	0.221	0.058	0.077
		AAM	0.198	0.056	0.063
	100	LDA	0.197	0.478	0.505
		QDA	0.222	0.343	0.137
		ACE	0.238	0.087	0.126
		PPR	0.192	0.061	0.100
		MARS	0.218	0.057	0.075
		AAM	0.197	0.053	0.061
300	50	LDA	0.197	0.478	0.570
		QDA	0.218	0.380	0.135
		ACE	0.238	0.086	0.124
		PPR	0.191	0.063	0.098
		MARS	0.219	0.062	0.073
		AAM	0.197	0.055	0.061
	100	LDA	0.194	0.476	0.540
		QDA	0.217	0.360	0.135
		ACE	0.236	0.084	0.126
		PPR	0.189	0.059	0.099
		MARS	0.217	0.055	0.075
		AAM	0.194	0.053	0.061
	150	LDA	0.193	0.475	0.503
		QDA	0.218	0.340	0.135
		ACE	0.234	0.084	0.124
		PPR	0.188	0.058	0.098
		MARS	0.214	0.054	0.073
		AAM	0.193	0.050	0.059

بعلاوه روش AAM برای دو نوع داده موجی و خوشه‌ای حتی از روش MARS نیز بهتر تحلیل ممیزی می‌کند. بر خلاف تصور، روش‌های تحلیل ممیزی خطی به طور کلی از روش‌های غیرخطی همواره ضعیف‌تر نیستند و در مواردی می‌توانند دقیقتر از روش‌های دیگر باشند. اما روش‌های غیرخطی زمانی بهتر از روش‌های

خطی عمل می‌کنند که داده‌ها در گروه‌هایی کاملاً جدا از هم قرار داشته باشند. هر چه داده‌ها از گروه‌های متمایزتری تشکیل شده باشند، روش تحلیل ممیزی AAM نسبت به روش MARS عملکرد بهتری در تحلیل ممیزی داده‌ها از خود نشان می‌دهد.

بعلاوه شرایط محدود کننده روش‌های خطی و درجه دوم که عموماً در عمل محقق نمی‌باشند، بکارگیری روش‌های پیشنهاد شده در این مقاله را در اغلب مسائل کاربردی مرجح می‌سازد.

تمام محاسبات و شبیه سازی‌های این مقاله توسط برنامه‌نویسی در محیط نرم‌افزار Splus انجام شده که از طریق پست الکترونیکی هر یک از نویسندگان قابل دریافت است.

قدردانی و تشکر

نویسندگان از داوران محترم مجله به خاطر نظرات و پیشنهادات سازنده‌شان که موجب بهبود این مقاله گردید، و از حمایت قطب علمی داده های ترتیبی و فضایی دانشگاه فردوسی مشهد نهایت تشکر و قدردانی را دارند.

منابع

1. L., Breiman, and R., Ihaka, Nonlinear Discriminant Analysis Via Scaling and ACE. Technical Report, Univ. of California, Berkeley; (1984).

2. L., Breiman, J. J., Friedman, R. A., Olshen, and C. J., Classification and Regression Trees. Words Worth, Monterey, CA; (Stone, 1984).

3. A., Buja, T., Hastie, and R., Tibshirani, Linear Smoothers and Additive Models(with Discussion). Annals of Statistics, 17, 453-555; (1989).

4. R. A., Fisher, Use of Multiple Measurements in Taxonomic Problems, Ann. Eug. 7 , 179-184; (1936).

5. J. H., Friedman, , Multivariate Adaptive Regression Splines(with Discussion). Annals of Statistics, Vol. 19, N0. 1, 1-67; (1991).

6. A., Gifi, Nonlinear Multivariate Analysis. Wiley, Chichester ; (1981).

7. C., Gu., and G., Wahba,. Minimizing GCV/GML Scores with Multiple Smoothing Parameters Via Newton's Method., Technical Report 847, University of Wisconsin – Madison; (1984).

8. T., Hastie, Discussion of "Flexible Parsimonious Smoothing and Additive Modelling" by Friedman and Silverman. Technometrics, 31 , 3-39; (1989).

9. T., Hastie, R., Tibshirani, and A., Buja, Generalized Additive Models. Chapman and Hall; (1993).

10. T., Hastie, R., Tibshirani, and A., Buja, Flexible Discriminant Analysis by Optimal Scoring. Journal of American Statistical Association; (1994).

11. T., Hastie, A., Buja, and R., Tibshirani, Penalized Discriminant Analysis, Annals of Statistics, Vol. 23, No. 1, 73-102; (1995)..

12. J. T., Kent, and M., Mohammadzadeh, Global Optimization of the Generalized Cross Validation Criterion, Statistics and Computing, Vol. 10, 231-236; (2000).

13. K. V., Mardia, J. T., Kent , and J. M., Bibby, Multivariate Analysis, Academic Press; (1979).

14. M., Mohammadzadeh, Estimating the Smoothing Parameter in Smoothing Splines, COMPSTAT'98, Proceedings of the Thirteen symposium on Computational Statistics, Bristol, England; (1998).

15. J., Neyman, and E. S., Pearson, The Problem of Most Efficient Test of Statistical Hypotheses. Phil. Trans. R. Soc. 231;(1933).

16. G., Wahba, Spline Models for Observational Data. CBSM-NSF Regional Conf. Ser. Appl. Math. 59, SIAM, Philadelphia; (1990).

17. A., Wald,. Statistical Decision Function. Wiley, New York; (1944).

18. B. L., Welch, Note on Discriminant Function. Biometrika 31, 218-220; (1939).