مجله علوم آماری، بهار و تابستان ۱۳۹۳
جلد ۸، شماره ۱، ص ۱-۱۸

مدل‌بندی داده‌های فازی با رگرسیون اسپلاین تطبیقی چندگانه

جلال چاچی [۱]، غلامرضا حسامیان [۲]

[۱] گروه آمار، دانشگاه سمنان
[۲] گروه آمار، دانشگاه پیام‌نور

تاریخ دریافت: ۱۳۹۲/۷/۳۰ تاریخ آخرین بازنگری: ۱۳۹۳/۳/۱۸

چکیده : در این مقاله به مدل‌بندی داده‌های ورودی دقیق-خروجی فازی پرداخته می‌شود و رویکرد رگرسیون مارس فازی با پارامترهای دقیق و جملات خطای فازی معرفی می‌گردد. روش پیشنهادی شامل دو مرحله است: در مرحله اول با استفاده از رگرسیون اسپلاین تطبیقی چندگانه (مارس) مراکز متغیر وابسته برآورد می‌شوند، و در مرحله دوم کمترین مقادیر خطاهای فازی بر اساس یک مسأله بهینه‌سازی غیر خطی به‌دست می‌آیند. در انتها کاربرد مدل پیشنهاد شده در مدل‌بندی داده‌های واقعی در مهندسی آب نشان داده می‌شود. نتایج تجربی این مثال برتری روش پیشنهادی را در مقایسه با برخی از روش‌های متداولِ رگرسیون فازی کمترین توان‌های دوم خطا نشان می‌دهد.

واژه‌های کلیدی : رگرسیون اسپلاین تطبیقی چندگانه (مارس)، داده‌های فازی، سامانهٔ استنتاج فازی، دبی رودخانه.

---

* آدرس الکترونیک مسئول مقاله: جلال چاچی، jchachi@profs.semnan.ac.ir
* کد موضوع‌بندی ریاضی (۲۰۱۰): ۶۲J۸۶، ۶۲G۰۸، ۹۴D۰۵

۱ مقدمه

مدل‌های رگرسیونی (آماری و فازی) الگوهایی فراهم می‌آورند که می‌توان بر پایه آنها ارتباط بین مجموعه‌ای از متغیرهای تبیینی و پاسخ را بررسی کرد. بر خلاف رگرسیون آماری، رگرسیون فازی را می‌توان در موارد زیر به‌کار برد:

- داده‌ها نادقیق باشند؛
- رابطه بین متغیرها از نوع امکانی (و نه احتمالی) باشد؛
- فرضیات زیربنایی در مدل‌های رگرسیون آماری مورد تردید باشند.

در رگرسیون فازی، همانند رگرسیون آماری، یک فرم تابعی بین مشاهدات متغیرهای تبیینی و پاسخ و به منظور کنترل و پیش‌بینی مقادیر متغیر پاسخ در نظر گرفته می‌شود. بر این اساس و به منظور کنترل ابهام مقدار پیش‌بینی یا فاصله‌آن با مقدار واقعی متغیر پاسخ، شیوه‌های رگرسیون فازی را می‌توان از لحاظ رویکرد به دو نوع کلی طبقه‌بندی کرد:

۱) **رویکردهای موسوم به رگرسیون امکانی:** در رگرسیون امکانی خطای مدل در قالب توزیع‌های امکانی ضرایب مدل منظور می‌شود. در این مدل‌ها خطای پیش‌بینی متغیر وابسته تعبیر امکانی دارد. به سخن دیگر هنگامی که بر پایه مدل رگرسیونی و به‌ازای مقادیر متغیرهای تبیینی مقدار متغیر پاسخ را پیش‌بینی می‌کنیم، مقادیر پیش‌بینی تعبیر امکانی دارد و نه تعبیر احتمالی. یافتن ضرایب این گونه مدل‌ها اغلب مستلزم استفاده از روش‌های برنامه‌ریزی خطی یا غیرخطی است. در این رویکردها ابهام کل مدل که برابر با مجموع پهناهای مقادیر برآورد شدهٔ متغیر پاسخ است، تحت شرایط و قیودی کمینه می‌شود.

۲) **شیوه‌های مبتنی بر روش کمترین توان‌های دوم:** در این رویکردها بر پایه تعریف‌هایی برای فاصله بین اعداد فازی روش کمترین توان‌های دوم کلاسیک تعمیم می‌یابد و بر اساس آن ضرایب مدل رگرسیون فازی برآورد می‌شوند. به‌عبارتی در این شیوه، مجموع فاصله‌ها بین مقادیر متغیر پاسخ

مشاهده شده و مقادیر متغیر پاسخ برآورد شده، بر اساس تعریف یک فاصله بین اعداد فازی، کمینه می‌شود. این رویکردها نیز تنوع بسیاری دارند. این تنوع عوامل مختلفی دارد، از جمله: دقیق یا نادقیق بودن مشاهدات متغیرها، تنوع در تعریفی که برای فاصله بین مجموعه‌های فازی در نظر گرفته می‌شود، دقیق یا نادقیق منظور نمودن عرض از مبدا و برخی عوامل دیگر است.

از آنجا که رگرسیون فازی از جملهٔ موضوعاتی است که بسیار مورد توجه و مطالعه محققان قرار گرفته است و تنوع گسترده‌ای در روش‌ها و رویکردهای معرفی شده وجود دارد، می‌توان برای مطالعه و مرور روش‌های معرفی شده در این زمینه می‌توان به ارقامی (۱۳۸۱)، میرزایی و ارقامی (۱۳۸۶)، چاچی (۱۳۹۱) و دورسو و همکاران (۲۰۱۰، ۲۰۱۱) مراجعه نمود. برخی از این رویکردها از جهات مختلف مورد بررسی و نقد قرار گرفته‌اند. در جهت رفع مشکلات بالا تحقیقات بسیاری صورت گرفته که می‌توان به چن و دنگ (۲۰۰۸) و لو و ونگ (۲۰۰۹) رجوع کرد.

در ادامه برای رفع برخی از انتقادات مطرح شده، از روش رگرسیون اسپلاین تطبیقی چندگانه (مارس[۱]) به عنوان جایگزینی برای روش کمترین توان‌های دوم در محیط فازی استفاده می‌شود. روش مارس یک شیوه مدل‌بندی رگرسیون قطعه‌ای ناپارامتری است که هیچ فرضیهٔ زیربنایی درباره رابطه تابعی بین متغیرهای تبیینی و پاسخ درنظر نمی‌گیرد. این روش برای مدل کردن داده‌هایی که دامنه وسیعی دارند، یا رابطه بین متغیرها خطی نمی‌باشد، بسیار مناسب است (فریدمن، ۱۹۹۱؛ هستی و همکاران، ۲۰۰۹). در طول چند دههٔ اخیر روش مارس با بسیاری از روش‌های پارامتری و ناپارامتری متداول (از جمله روش کمترین توان‌های دوم خطا) از لحاظ دقت برازش، کارایی، استواری، و سادگی محاسبات مورد مقایسه قرار گرفته است (دی-آندرس و همکاران، ۲۰۱۱؛ کرینر، ۲۰۰۷؛ لی و چن، ۲۰۰۵؛ لی و همکاران، ۲۰۰۶).

در این مقاله با روش مارس داده‌های فازی مدل‌بندی می‌شوند و روش پیشنهادی با روش‌های مشابه در رگرسیون فازی با رویکرد کمترین توان‌های دوم مورد مقایسه

---

۱ Multivariate Adaptive Regression Splines (MARS)

قرار می‌گیرد. از آنجا که روش مارس در مقایسه با روش کمترین توان‌های دوم نسبت به داده‌های پرت استوار است، به نظر می‌رسد که روش پیشنهادی برتری قابل توجهی به روش کمترین توان‌های دوم در مدل‌بندی داده‌هایی که در آنها مشاهدات پرت وجود داشته باشد، دارد. روش پیشنهادی یک روش دو مرحله‌ای است. در مرحله اول، با استفاده از مارس مراکز متغیر پاسخ برآورد می‌شوند. سپس در مرحله دوم و با استفاده از روش‌های بهینه‌سازی، جملات خطای فازی متناظر با داده‌ها به‌دست می‌آیند. در این روش ضرایب مدل دقیق هستند و جملات خطای فازی برای در نظر گرفتن ابهام به مدل افزوده می‌شوند.

در بخش ۲، انگیزه‌ها و اهداف ارائهٔ مدل جدید بیان می‌شود. در بخش ۳، مدل رگرسیون مارس فازی با پهناهای متغیر بیان و تشریح می‌شود. نحوه پیش‌بینی یک مقدار جدید با استفاده از سامانهٔ استنتاج فازی در بخش ۴ بیان و تشریح می‌شود. دو ملاک نیکویی برازش برای مدل‌های رگرسیون فازی در بخش ۵ ذکر می‌شوند. در بخش ۶، با استفاده از داده‌های واقعی در مهندسی آب یک مطالعه مقایسه‌ای بین روش پیشنهادی و چند روش مطرح در رگرسیون فازی انجام می‌شود. در انتها به بحث و نتیجه‌گیری پرداخته می‌شود.

۲ اهداف

تا کنون رویکردهای مختلف و متنوعی در زمینه رگرسیون فازی ارایه شده است. بسیاری از این رویکردها از جهات مختلف مورد بررسی و نقد قرار گرفته‌اند (چاچی، ۱۳۹۱). مشکل اساسی بسیاری از روش‌های موجود عبارتند از:

- مدل‌های رگرسیون امکانی و مدل‌های مبتنی بر کمترین توان‌های دوم فازی نسبت به داده(های) پرت حساس هستند؛
- در برخی از روش‌های موجود با افزایش مقدار متغیر مستقل پهنای برآورد شدهٔ متغیر پاسخ افزایش می‌یابد، که به آن مشکل افزایش پهناها گفته می‌شود؛
- در برخی از روش‌های مدل‌بندی مشاهدات فازی متغیر پاسخی که پهناهای مشاهده شده آن روند صعودی، نزولی یا متغیر دارند، ممکن است برآوردهایی

با خطای زیاد حاصل شوند.

اگرچه برای حل انتقادات فوق برخی روش‌ها در رگرسیون فازی ارائه شده است، اما با وجود این هنوز روشی که به‌طور کامل این مشکلات را برطرف کند، ارائه نشده است. برای مثال، روش‌های زیر در برطرف نمودن حساسیت مدل نسبت به داده پرت معرفی شدند:

- رویکردهای رگرسیون فازی مبتنی بر روش کمترین قدرمطلق خطا (چاچی، ۱۳۹۱؛ کلکین‌نما و طاهری، ۲۰۱۲)؛
- رویکردهایی که در آنها از روش‌های رگرسیون آماری استوار، از جمله M-برآوردگرها، S-برآوردگرها و غیره، استفاده می‌شود (دورسو و همکاران، ۲۰۱۱).

اما در رویکردهای بالا مسألهٔ افزایش پهناها همچنان حل نشده است. از سوی دیگر، در زمینه برطرف کردن مسألهٔ افزایش پهناها نیز می‌توان به روش‌های زیر اشاره کرد:

- روش‌هایی که در آنها اغلب ضرایب مدل دقیق هستند، یا
- یک جمله فازی به‌منظور در نظر گرفتن ابهام فازی به مدل افزوده می‌شود (چن و دنگ، ۲۰۰۸؛ لو و ونگ، ۲۰۰۹).

این روش‌ها نیز در چارچوب رگرسیون کمترین توان‌های دوم فازی قرار دارند و از جهت حساسیت نسبت به داده پرت مورد انتقاد هستند. در میان روش‌های فوق باید به روش چن و دنگ (۲۰۰۸) اشاره نمود که یک روش رگرسیون فازی با پهناهای متغیر است که مسألهٔ افزایش پهناها را برطرف می‌کند. این روش همچنین، برآوردهای خوبی را برای مشاهدات فازی متغیر پاسخی که پهناهای آن روند صعودی، نزولی، ثابت و حتی متغیر دارند ارائه می‌دهد. اما همچنان که گفته شد، این روش نیز با مشکل تأثیرپذیری زیاد نسبت به دادهٔ پرت مواجه است.

در ادامه یک مدل رگرسیون فازی ارائه می‌شود که نسبت به داده(های) پرت استوار است و دیگر این که مسألهٔ افزایش پهناها را برحسب افزایش مقادیر متغیر تبیینی و پهناهای متغیر پاسخ همزمان برطرف می‌کند.

۳ بیان و تشریح مدل

فرض کنید مجموعه‌ای از مشاهداتِ مربوط به یک یا چند متغیر تبیینی و یک متغیر پاسخ در اختیار داریم که مشاهدات متغیر(های) تبیینی دقیق هستند ولی مشاهدات متغیر پاسخ نادقیق (فازی) هستند. برای مدل‌بندی چنین داده‌هایی یک مدل رگرسیون فازی با پهناهای متغیر ارائه می‌شود. مدل مربوطه به‌صورت

$$\tilde{y}_i = \beta_0 + \sum_{m=1}^{M} \beta_m B_m(x_i) \oplus \tilde{\varepsilon}_i, \qquad i = 1, \ldots, n$$

درنظر گرفته می‌شود، که در آن $\tilde{y}_i = (y_i, y_i^\ell, y_i^r)_{LR}$، iامین مشاهده فازی متغیر پاسخ، $\mathbf{x}_i = [x_{0i}, x_{1i}, \ldots, x_{ki}] \in \mathbb{R}^{k+1}$ $(i = 1, \ldots, n; k < n; x_{0i} = 1)$، iامین مشاهده حقیقی مقدار متغیرهای تبیینی، β_m، $m = 0, 1, \ldots, M$، ضرایب دقیق مدل مارس برای توابع پایهٔ[۲] $B_m(x_i)$ و $\tilde{\varepsilon}_i = (0, \ell_i, r_i)_{LR}$ خطای فازی مدل برای مشاهده iام است.

برای برآورد پارامترها $(\beta_m,\ m = 0, 1, \ldots, M)$ و جملات خطای فازی مدل $(\tilde{\varepsilon}_i = (0, \ell_i, r_i)_{LR}\ , i = 1, \ldots, n)$ یک روش دومرحله‌ای معرفی می‌شود. در مرحله اول ضرایب دقیق $\beta = [\beta_0, \beta_1, \ldots, \beta_M]^t$ برآورد می‌شوند، و در مرحله دوم جملات خطای فازی بر اساس یک مسألهٔ بهینه‌سازی به‌دست می‌آیند.

مرحله ۱. ضرایب رگرسیون $\beta = [\beta_0, \beta_1, \ldots, \beta_M]^t$ با مدل‌بندی مراکز متغیر پاسخ فازی بر روی مقادیر متغیر تبیینی، از طریق مدل مارس

$$y_i = \beta_0 + \sum_{m=1}^{M} \beta_m B_m(x_i) + \varepsilon_i, \qquad i = 1, \ldots, n$$

---

[۲] Basis functions

به‌دست آورده می‌شود. ضرایب مدل در روش مارس با اجرای برنامه‌های earth یا mars در نرم‌افزار R برآورد می‌شوند (فوکس و ویزبرگ، ۲۰۱۱). سپس مراکز متغیر پاسخ فازی به‌صورت زیر به‌دست می‌آیند.

$$\widehat{y}_i = \widehat{\beta}_0 + \sum_{m=1}^{M} \widehat{\beta}_m B_m(x_i), \qquad i = 1, \ldots, n$$

مرحله ۲. عبارت خطای فازی $\tilde{\varepsilon}_i = (0, \ell_i, r_i)_{LR}$ به‌ازای $i = 1, \ldots, n$، به‌گونه‌ای برآورد می‌شود که:

الف- پهنای برآورد شده هر $\tilde{\varepsilon}_i$، یعنی $\ell_i + r_i$، مساوی پهنای مشاهده شده $\tilde{y}_i$، یعنی $y_i^\ell + y_i^r$، باشد،

ب- برآورد مشاهده $\tilde{y}_i$، یعنی $\widehat{\tilde{y}}_i$، کمترین خطای برآورد را داشته باشد.

برای این منظور خطاهای فازی را از طریق حل یک مسألهٔ برنامه‌ریزی غیرخطی و با در نظر گرفتن شرط اول به عنوان قیود آن و شرط دوم به عنوان تابع هدف آن به‌دست آورده می‌شود که در آن تابع هدف مجموع اندازه‌های تشابه بین توابع عضویت مشاهده شده و برآورد شدهٔ متغیر پاسخ به‌صورت (لو و ونگ، ۲۰۰۹)

$$\sum_{i=1}^{n} \frac{\int \min\{\tilde{y}_i(x), \widehat{\tilde{y}}_i(x)\}\, dx}{\int \max\{\tilde{y}_i(x), \widehat{\tilde{y}}_i(x)\}\, dx}$$

است و قیود به‌گونه‌ای اختیار می‌شوند که iامین پهنای برآورد شده متغیر پاسخ مساوی iامین پهنای مشاهده شده آن باشد. یعنی شرط $\ell_i + r_i = y_i^\ell + y_i^r$ برای هر i، $i = 1, \ldots, n$، برقرار باشد. این قید به این دلیل در نظر گرفته می‌شود که پهناهای برآورد شده کنترل شود، یا به‌عبارتی مقادیر پهناهای خیلی کم یا خیلی زیاد برای متغیر پاسخ برآورد نشود (چن و دنگ، ۲۰۰۸). برای در نظر گرفتن این قیود فرض می‌شود y_m^ℓ و y_m^r به‌ترتیب کمترین مقدار پهناهای چپ و راست مشاهدات متغیر وابسته باشند، یعنی

$$y_m^\ell = \min\{y_1^\ell, \ldots, y_n^\ell\}, \qquad y_m^r = \min\{y_1^r, \ldots, y_n^r\}.$$

قسمتی از پهنای برآورد شدهٔ متغیر پاسخ مقدار ثابت $y_m^\ell + y_m^r$ اختیار می‌شود. بدیهی است که

$$D_i = (y_i^\ell + y_i^r) - (y_m^\ell + y_m^r) \geq ۰, \qquad i = ۱, \ldots, n.$$

اما این مقدار ثابت باید از دو طرف گسترش یابد تا در نهایت مساوی با iامین پهنای مشاهده شدهٔ متغیر وابسته شود. برای این منظور، مقدار D_i به دو قسمت d_i و $D_i - d_i$ تقسیم می‌شود، که در آن $۰ \leq d_i \leq D_i$. بنابراین، قیود مسألهٔ بهینه‌سازی غیرخطی به‌صورت

$$\tilde{\varepsilon}_i = (۰, y_m^\ell + d_i, y_m^r + D_i - d_i)_{LR}, \quad ۰ \leq d_i \leq D_i, \ i = ۱, \ldots, n,$$

در نظر گرفته می‌شود. سرانجام، با توجه به تابع هدف و قیود بیان شده مسألهٔ بهینه‌سازی غیرخطی که از طریق آن مقادیر $d_۱, \ldots, d_n$ برآورد می‌شوند به‌صورت

$$\max_{d_۱, \ldots, d_n} \sum_{i=۱}^{n} \frac{\int \min\{\tilde{y}_i(x), \widehat{\tilde{y}}_i(x)\}dx}{\int \max\{\tilde{y}_i(x), \widehat{\tilde{y}}_i(x)\}dx}, \quad \widehat{\tilde{y}}_i = \widehat{\beta}_۰ + \sum_{m=۱}^{M} \widehat{\beta}_m B_m(x_i) \oplus \tilde{\varepsilon}_i,$$

$$\tilde{\varepsilon}_i = (۰, y_m^\ell + d_i, y_m^r + D_i - d_i)_{LR}, \ ۰ \leq d_i \leq D_i, \ i = ۱, \ldots, n.$$

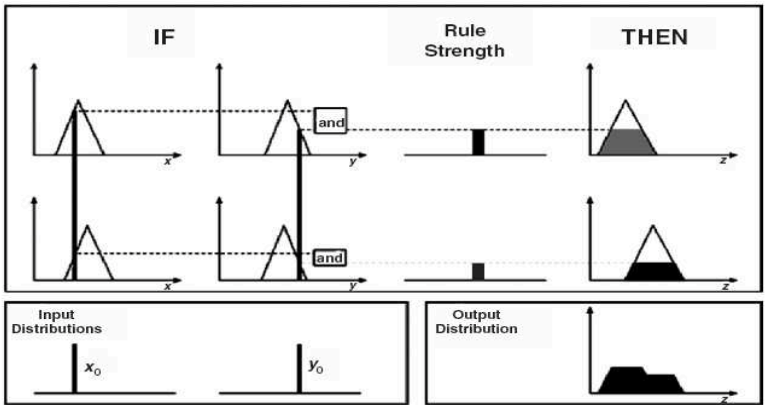
در نظر گرفته می‌شود. پس از تعیین مقادیر $\widehat{d}_۱, \ldots, \widehat{d}_n$، مدل رگرسیون مارس فازی به‌صورت زیر حاصل می‌شود

$$\begin{aligned} \widehat{\tilde{y}}_i &= \widehat{\beta}_۰ + \sum_{m=۱}^{M} \widehat{\beta}_m B_m(x_i) \oplus \widehat{\tilde{\varepsilon}}_i \\ &= \widehat{\beta}_۰ + \sum_{m=۱}^{M} \widehat{\beta}_m B_m(x_i) \oplus (۰, y_m^\ell + \widehat{d}_i, y_m^r + D_i - \widehat{d}_i)_{LR}, \ i = ۱, \ldots, n. \end{aligned}$$

۴ پیش‌بینی متغیر پاسخ با سامانهٔ استنتاج فازی

پس از تعیین مقادیر $\widehat{\beta}_۰, \widehat{\beta}_۱, \ldots, \widehat{\beta}_M$ و $\widehat{d}_۱, \ldots, \widehat{d}_n$، مدل رگرسیون مارس فازی با پهناهای متغیر به‌صورت

$$\widehat{\tilde{y}}_i = \widehat{\beta}_۰ + \sum_{m=۱}^{M} \widehat{\beta}_m B_m(x_i) \oplus (۰, y_m^\ell + \widehat{d}_i, y_m^r + D_i - \widehat{d}_i)_{LR}, \qquad i = ۱, \ldots, n$$



شکل ۱: سامانهٔ استنتاج فازی با دو متغیر ورودی دقیق و دو قانون اگر-آنگاه فازی

نتیجه می‌شود. اکنون با استفاده از این مدل می‌توان به پیش‌بینی مقادیر متغیر پاسخ بر اساس مشاهدات جدیدی از متغیر تبیینی پرداخت. برای پیش‌بینی، چن و دنگ (۲۰۰۸) از سامانهٔ استنتاج فازی استفاده کردند که در ادامه توضیح داده می‌شود و برای پیش‌بینی متغیر پاسخ بر اساس مدل پیشنهاد شده در این مقاله از این شیوه استفاده می‌شود. ساختار اصلی یک سامانهٔ استنتاج فازی از سه بخش اصلی تشکیل می‌شود (زیمرمن، ۲۰۰۱):

الف- **بخش قوانین:** شامل مجموعه‌ای از قوانین اگر-آنگاه فازی است؛

ب- **بخش مجموعه داده‌ها یا اطلاعات فازی داده‌ها:** شامل توابع عضویت داده‌های فازی به‌کار رفته در قوانین اگر-آنگاه فازی است؛

ج- **فرآیند استنتاج:** شامل قواعد تصمیم‌گیری بر اساس اطلاعات به‌دست آمده و روش الحاق نتایج حاصله از قوانین اگر-آنگاه فازی است.

در شکل ۱ سامانهٔ استنتاج فازی ممدانی با دو متغیر ورودی دقیق و دو قانون اگر-آنگاه فازی نشان داده شده است.

فرض کنید $\mathbf{x}^* = [1, x_1^*, \ldots, x_k^*]$ مقدار جدید متغیر تبیینی و $Y = \widehat{\beta}_0 + \sum_{m=1}^{M} \widehat{\beta}_m B_m(x_i^*)$ پیش‌بینی مرکز متغیر پاسخ بر اساس مدل مارس

باشد. اکنون $\widehat{\tilde{\epsilon}}$ جمله خطای فازی متناظر با این مشاهده برای مدل

$$\tilde{Y} = \widehat{\beta}_{\circ} + \sum_{m=1}^{M} \widehat{\beta}_m B_m(x_i^*) \oplus \widehat{\tilde{\epsilon}},$$

از طریق یک سامانه استنتاج فازی ممدانی با یک متغیر ورودی–خروجی به‌دست آورده می‌شود. فرض کنید

$$(\tilde{y}_a, \widehat{\tilde{\varepsilon}}_a) \qquad a \in A = \{i : \tilde{y}_i(Y) > \circ,\ i = 1, \ldots, n\},$$

مشاهداتی از متغیر پاسخ به همراه جملات خطای فازی متناظر با آنها باشند که توسط مقدار جدید $\mathbf{x}^*$ فعال می‌شوند. حال aامین قانون اگر–آنگاه فازی به‌صورت

R^a: اگر Y به‌صورت $\tilde{y}_a$ باشد، آنگاه $\widehat{\tilde{\epsilon}}$ به‌صورت $\widehat{\tilde{\varepsilon}}_a$ است،

نوشته می‌شود، که در آن $a \in A$ و R^a قانون aام است. در سامانه استنتاج فازی مقدم و تالی هر قانون اگر–آنگاه فازی به‌صورت گزاره‌های فازی هستند. به‌علاوه، بر اساس قوانین در نظر گرفته شده خروجی این سامانه نیز به‌صورت یک مجموعه فازی است. اکنون برای مقدار Y، هر تابع عضویت $\tilde{y}_a$ در خروجی به میزان درجه عضویت $\tilde{y}_a(Y)$ فعال می‌شود، که از آن در به‌دست آوردن جمله خطای فازی استفاده می‌شود. در نهایت، با الحاق خروجی‌های تمام قوانینی که فعال شده‌اند، جمله خطای فازی به‌دست می‌آید.

توجه شود که حتی اگر تمام مشاهدات متغیر پاسخ اعداد فازی LR باشند، تابع عضویت به‌دست آمده برای خطای فازی هر شکل نامنظمی می‌تواند داشته باشد (شکل ۱). اما مناسب است که خطای برآورد شده نیز به‌صورت یک عدد فازی LR به‌دست آید، زیرا معمولا ترجیح داده می‌شود که مقدار پیش‌بینی نیز به‌صورت یک عدد فازی LR باشد. برای این منظور، خطای فازی تبدیل به یک عدد فازی LR می‌شود تا پیش‌بینی متغیر پاسخ نیز یک عدد فازی LR شود. بنابراین، خطای فازی به‌صورت عدد فازی $\widehat{\tilde{\epsilon}} = (\widehat{\epsilon}, \widehat{\ell}, \widehat{r})_{LR}$ در نظر گرفته می‌شود که در آن $\widehat{\epsilon}$ مقدار غیرفازی شده عبارت خطای فازی بر اساس روش مرکز ثقل است و $\widehat{\ell}$ و $\widehat{r}$ کمترین و بیشترین مقادیر ممکن برای پهناهای چپ و راست عبارت خطای فازی هستند.

سرانجام، برای مقدار جدید $\mathbf{x}^*$ مقدار پیش‌بینی متغیر پاسخ به‌صورت

$$\widehat{\widetilde{Y}}_{x^*} = Y \oplus \widehat{\widetilde{\epsilon}} = (Y + \widehat{\epsilon}, \widehat{\ell}, \widehat{r})_{LR},$$

به‌دست می‌آید.

۵ ملاک‌های ارزیابی مدل

برای ارزیابی مدل‌های رگرسیون فازی چندین ملاک توسط مولفان مورد استفاده قرار گرفته است (چن و دنگ، ۲۰۰۸؛ لو و ونگ، ۲۰۰۹؛ کلکین‌نما و طاهری، ۲۰۱۲). در ادامه دو ملاک متداول برای بررسی نیکویی برازش مدل‌های رگرسیون فازی یادآوری می‌شوند. از این دو ملاک برای مقایسه مدل رگرسیونی پیشنهاد شده در این مقاله و چند رویکرد دیگر در زمینه رگرسیون فازی استفاده می‌شود. این دو ملاک میانگین اندازه‌های تشابه[۳] و میانگین قدرمطلق خطاهای[۴] برآورد هستند که به ترتیب به‌صورت زیر تعریف می‌شوند

$$\begin{aligned} MSM &= \frac{1}{n}\sum_{i=1}^{n} \frac{\int \min\{\tilde{y}_i(x), \widehat{\tilde{y}}_i(x)\}\, dx}{\int \max\{\tilde{y}_i(x), \widehat{\tilde{y}}_i(x)\}\, dx}, \\ MAE &= \frac{1}{n}\sum_{i=1}^{n} \int |\tilde{y}_i(x) - \widehat{\tilde{y}}_i(x)|\, dx. \end{aligned}$$

شایان ذکر است که در یک مدل رگرسیون فازی MAE می‌تواند هر مقدار بزرگتر از صفر باشد، ولی MSM همواره عددی بین صفر و یک است. اگر $MAE = 0$ یا $MSM = 1$ باشد، در این‌صورت مدل رگرسیون فازی مربوطه برازش کامل به داده‌ها دارد. بنابراین در مقایسه بین دو مدل رگرسیون فازی هر مدلی که مقدار MAE کوچکتری (نزدیک‌تر به صفر) داشته باشد، یا مقدار MSM بزرگتری (نزدیک‌تر به یک) داشته باشد، آن مدل برازش بهتری به داده‌ها دارد.

---

۳ Mean of similarity measures

۴ Mean of absolute errors

۶ مثال کاربردی

یکی از مسایل مهم در مهندسی آب اندازه‌گیری دبی یا بار معلق[۵] و سرعت آب [۶] در حوزه‌های آبریز است. در مطالعات کاربردی برآورد دبی بر اساس سرعت آب اهمیت زیادی دارد. بر اساس مطالعه‌ای در دربند واقع در خراسان شمالی با استفاده از ابزارهای استاندارد، برخی خصوصیات آب‌شناسی حوزه‌های آبریز منطقه ثبت گردید. در این خصوص از ایستگاهای مختلف تعداد ۱۸۴ جفت مشاهده مربوط به دبی و سرعت آب حوزه‌های آبریز منطقه به‌دست آمد. بنا به محدودیت‌های آزمایشگاهی و ابزارهای اندازه‌گیری و ماهیت متغیر پاسخ، این مشاهدات به‌صورت اعداد فازی مثلثی در جدول ۱ گزارش شده‌اند. نمودار پراکنش مراکز متغیر پاسخ بر حسب متغیر مستقل به همراه برآورد مارس آنها در شکل ۲ نشان داده شده است. با توجه به این نمودار واضح است که یک خط راست دقت کافی برای مدل‌بندی این داده‌ها را ندارد. در ادامه با رگرسیون مارس فازی به برآورد متغیر پاسخ فازی $(\tilde{y} = (y, s)_T)$ بر اساس مقادیر متغیر تبیینی (x) پرداخته می‌شود. اما بر خلاف شیوه‌های مدل‌سازی پارامتری متداول، در روش مارس ضرایب مدل برای تمام دامنه مشاهدات متغیرها یکسان در نظر گرفته نمی‌شود و به‌جای آن از خطوط شکسته‌ای[۷] به منظور برازش یک تابع پیوسته چند ضابطه‌ای بهینه به داده‌ها استفاده می‌شود. این رویکرد برای مدل‌بندی مسائلی که در آنها تعداد متغیرها بسیار زیاد است یا با حجم وسیعی از داده‌ها روبرو هستیم مفید است.

با به‌کار بردن روش پیشنهاد شده در این مقاله، مدل

$$\begin{aligned}\widehat{\tilde{y}}_i \;=\; & ۲/۸۲ + ۲/۲۷ \max\{۰, x_i + ۰/۲۴\} - ۰/۷۶ \max\{۰, -۰/۲۴ - x_i\} \\ & -۱۱/۹۸ \max\{۰, x_i - ۲/۱۷\} + ۱۵/۹۸ \max\{۰, x_i - ۲/۳۲\} \\ & \oplus (۰, ۰/۰۳ + \widehat{d}_i, ۰/۰۳ + D_i - \widehat{d}_i)_T, \qquad i = ۱, \ldots, ۱۸۴,\end{aligned}$$

---

[۵] Suspended load

[۶] Discharge

[۷] Splines

جدول ۱: داده‌های مهندسی آب

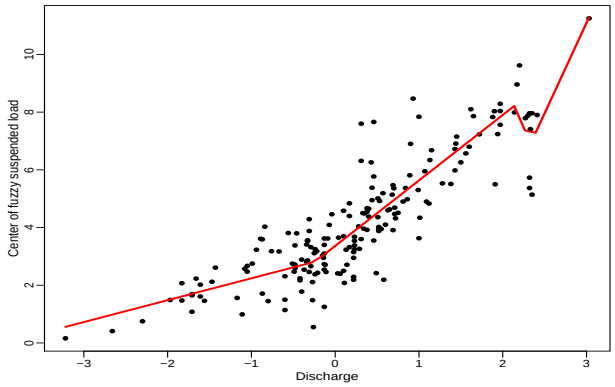
i	x_i	$\tilde{y}_i = (y_i, s_i)_T$	$\widehat{d}_i$
۱	−۰/۵۱	(۲/۷۵, ۰/۵۵)$_T$	۰/۳۳
۲	−۰/۲۱	(۲/۴۳, ۰/۲۳)$_T$	۰/۴۰
۳	−۰/۲۴	(۳/۲۵, ۱/۱۳)$_T$	۰/۵۰
۴	−۰/۳۳	(۳/۵۳, ۱/۴۹)$_T$	۰/۳۸
۵	۰/۰۴	(۳/۶۵, ۱/۶۹)$_T$	۱/۳۹
⋮	⋮	⋮	⋮
۱۸۰	۰/۹۰	(۶/۹۰, ۴/۳۸)$_T$	۲/۲۶
۱۸۱	۰/۴۶	(۷/۶۶, ۵/۰۰)$_T$	۰/۴۸
۱۸۲	۰/۸۱	(۴/۹۰, ۲/۷۸)$_T$	۳/۱۸
۱۸۳	۱/۴۴	(۶/۹۱, ۴/۲۵)$_T$	۳/۸۳
۱۸۴	۲/۳۰	(۷/۸۸, ۵/۷۶)$_T$	۴/۵۳

برای این داده‌ها به‌دست می‌آید. برآوردهای $\widehat{d}_i$، $i = ۱, \ldots, ۱۸۴$ در جدول ۱ آورده شده‌اند.

در ادامه روش پیشنهاد شده در این مقاله با رویکردهای ژو و لی (۲۰۰۱) (XL)، چن و دنگ (۲۰۰۸) (CD)، و فرارو و همکاران (۲۰۱۰) (F) در رگرسیون فازی مقایسه می‌شود. توجه شود که در روش ژو و لی (۲۰۰۱) مقادیر متغیر تبیینی نامنفی در نظر گرفته می‌شوند و در اینجا بدون اینکه از کلیت مساله کاسته شود، داده‌های متغیر تبیینی به $x'_i = x_i - \min_i\{x_i\} = x_i - (-۳/۲۲) \geq ۰$ تبدیل می‌شود. این سه مدل به‌صورت

$$\begin{aligned}
\widehat{\tilde{y}}_i^{XL} &= (-۱/۵۳ + ۱/۶۴\, x'_i, \max\{۰, -۲/۲۰ + ۱/۲۲\, x'_i\})_T, \\
\widehat{\tilde{y}}_i^{CD} &= ۳/۷۶ + ۱/۶۴\, x_i \oplus \widehat{\tilde{\varepsilon}}_i, \\
\widehat{\tilde{y}}_i^{F} &= (۳/۷۶ + ۱/۶۴\, x_i, \exp\{۰/۱۰ + ۰/۷۴\, x_i\})_T,
\end{aligned}$$

به‌دست می‌آیند. نتایج نیکویی برازش سه مدل بالا به همراه مدل رگرسیون مارس فازی در جدول ۲ بیانگر برتری مدل رگرسیون مارس فازی بر سه روش دیگر است.



شکل ۲: نمودار پراکنش مراکز متغیر پاسخ بر حسب مقادیر متغیر تبیینی به همراه برآورد مارس آنها

جدول ۲: مقادیر نیکویی برازش مدل‌های رگرسیون فازی

مدل‌های پیشنهاد شده توسط	MSM	MAE
ژو و لی (۲۰۰۱)	۰/۴۳۸۲	۱/۳۰۵۵
چن و دنگ (۲۰۰۸)	۰/۶۲۲۹	۰/۵۴۳۱
فرارو و همکاران (۲۰۱۰)	۰/۳۹۲۷	۱/۳۳۴۵
رگرسیون مارس فازی	۰/۶۷۴۶	۰/۴۶۸۳

برای پیش‌بینی مقدار فازی متغیر پاسخ بر اساس مدل رگرسیون مارس فازی به‌ازای $x = ۳/۴۰$ ابتدا مرکز متغیر پاسخ به‌صورت

$$\begin{aligned} Y &= ۲/۸۲ + ۲/۲۷ \max\{۰, ۳/۴۰ + ۰/۲۴\} \\ &\quad - ۰/۷۶ \max\{۰, -۰/۲۴ - ۳/۴۰\} - ۱۱/۹۸ \max\{۰, ۳/۴۰ - ۲/۱۷\} \\ &\quad + ۱۵/۹۸ \max\{۰, ۳/۴۰ - ۲/۳۲\} \\ &= ۱۳/۶۰, \end{aligned}$$

برآورد می‌شود. سپس طبق روش معرفی شده در بخش ۴، به منظور فراهم آوردن یک سامانه استنتاج فازی، داریم

$$A = \{i \mid \tilde{y}_i(۱۳/۶۰) \geq ۰,\ i = ۱, \ldots, ۱۸۴\}$$

$$= \{۸۲, ۹۷, ۹۸, ۱۴۱, ۱۴۶, ۱۸۴\}.$$

لذا خطاهای فازی

$$\widehat{\tilde{\varepsilon}}_i = (۰, ۰/۰۳ + \widehat{d}_i, ۰/۰۳ + D_i - \widehat{d}_i)_T, \qquad i \in A,$$

در این سامانه فعال هستند. حال فرض کنید Y مرکز برآورد شده و $\widehat{\tilde{\epsilon}}$ خطای فازی برآورد شدهٔ متناظر با آن باشد. بنابراین در سامانه استنتاج فازی قوانین اگر-آنگاه زیر را داریم:

۱) اگر Y به‌صورت $(۸/۴۷, ۵/۴۷)_T$ باشد، آنگاه $\widehat{\tilde{\epsilon}}$ به‌صورت $(۰, ۱/۳۱, ۹/۶۳)_T$ است؛

۲) اگر Y به‌صورت $(۸/۲۹, ۵/۶۳)_T$ باشد، آنگاه $\widehat{\tilde{\epsilon}}$ به‌صورت $(۰, ۴/۹۹, ۶/۲۷)_T$ است؛

۳) اگر Y به‌صورت $(۸/۹۶, ۵/۴۵)_T$ باشد، آنگاه $\widehat{\tilde{\epsilon}}$ به‌صورت $(۰, ۴/۵۰, ۶/۴۰)_T$ است؛

۴) اگر Y به‌صورت $(۹/۶۲, ۵/۰۲)_T$ باشد، آنگاه $\widehat{\tilde{\epsilon}}$ به‌صورت $(۰, ۲/۷۴, ۷/۳۰)_T$ است؛

۵) اگر Y به‌صورت $(۱۱/۲۵, ۶/۶۵)_T$ باشد، آنگاه $\widehat{\tilde{\epsilon}}$ به‌صورت $(۰, ۶/۷۰, ۶/۶۰)_T$ است؛

۶) اگر Y به‌صورت $(۷/۸۸, ۵/۷۶)_T$ باشد، آنگاه $\widehat{\tilde{\epsilon}}$ به‌صورت $(۰, ۴/۵۶, ۶/۹۶)_T$ است.

چون $Y = ۱۳/۶۰$، با اجرای قوانین بالا در قسمت سامانهٔ استنتاج فازی در نرم‌افزار MATLAB (مطلب، ۲۰۰۷)، خطای فازی $\widehat{\tilde{\epsilon}} = (۰/۲۹, ۶/۹۹, ۹/۳۴)_T$ به‌دست می‌آید. اکنون با جایگزین کردن خطای فازی به‌دست آمده در مدل رگرسیون مارس فازی داریم

$$\widehat{\tilde{Y}} = ۱۳/۶۰ \oplus (۰/۲۹, ۶/۹۹, ۹/۳۴)_T = (۱۳/۸۹, ۶/۹۹, ۹/۳۴)_T.$$

بحث و نتیجه‌گیری

در این مقاله روشی دو مرحله‌ای برای معرفی مدل رگرسیون فازی با پهناهای متغیر بیان شد. در این رویکرد افزایش مقادیر متغیر تبیینی تاثیری بر افزایش یا کاهش پهناهای برآورد شده برای متغیر پاسخ ندارد. همچنین این روش می‌تواند مشاهدات فازی را که در آنها پهناهای متغیر پاسخ، روند نزولی، صعودی، ثابت یا متغیر دارند، به خوبی مدل‌بندی کند.

چون در روش پیشنهادی، مراکز متغیر پاسخ فازی بر اساس روش مارس که یک روش استوار نسبت به داده‌های پرت است، برآورد شد، اثرات منفی و نامناسب داده‌های پرت در برآورد مراکز متغیر پاسخ وارد نمی‌شوند. از طرف دیگر، در برآورد جملات خطای فازی از مراکز برآورد شده استفاده می‌شود، بنابراین این برآوردها نیز تحت تاثیر اثرات منفی داده(های) پرت واقع نخواهند شد.

اگرچه رویکرد پیشنهادی با الهام از رویکرد چن و دنگ (۲۰۰۸) تدوین شده است، اما این دو رویکرد در شیوهٔ برآورد ضرایب رگرسیونی، قیود به‌کار رفته در مسأله بهینه‌سازی برای به‌دست آوردن جملات خطا و تعداد پارامترهای برآورد شده، متفاوت هستند. از طرفی می‌توان گفت که مدل پیشنهادی در مقایسه با روش چن و دنگ کمتر تحت تاثیر داده (یا داده‌های) پرت قرار می‌گیرد، که این موضوع دلیل برتری روش پیشنهاد شده در این مقاله بر روش آنها است. روش چن و دنگ یک روش رگرسیون کمترین توان‌های دوم فازی با پهناهای متغیر است که در برازش، برتری قابل توجهی بر بسیاری از روش‌های رگرسیون فازی دارد.

تقدیر و تشکر

نویسندگان مقاله از جناب آقای مهندس رضایی پژند کمال تشکر و قدردانی را بابت انتشار و در اختیار گذاردن داده‌های واقعی به‌کار رفته در این مقاله را دارند. همچنین از داوران محترم که نظرات ارزشمند ایشان باعث بهبود مطالب ارائه شده در این مقاله گردید، کمال تشکر و قدردانی را داریم.

مراجع

ارقامی، ن. ر. (۱۳۸۱)، مروری بر رگرسیون فازی، گزارش نخستین سمینار مجموعه‌های مشکک و کاربردهای آن، دانشگاه شهید باهنر کرمان، ۱–۱۸.

چاچی، ج. (۱۳۹۱)، روش‌های آماری بر اساس اطلاعات نادقیق، رساله دکترای آمار، دانشگاه صنعتی اصفهان، دانشکده علوم ریاضی.

میرزایی یگانه، ش.، ارقامی، ن. ر. (۱۳۸۶)، رگرسیون فازی: مروری بر چند رویکرد، اندیشهٔ آماری، **۱۲**، ۳۵–۴۷.

De Andrés, J., Lorca, P., De Cos Juez, F. J. and Sánchez-Lasheras, F. (2011), Bankruptcy Forecasting: A Hybrid Approach Using Fuzzy C-Means Clustering and Multivariate Adaptive Regression Splines (Mars), *Expert Systems with Applications*, **38**, 1866-1875.

Chen, S. P. and Dang, J. F. (2008), A Variable Spread Fuzzy Linear Regression Model with Higher Explanatory Power and Forecasting Accuracy, *Information Sciences*, **178**, 3973-3988.

D'Urso, P., Massari, R. and Santoro, A. (2010), A Class of Fuzzy Clusterwise Regression Models, *Information Sciences*, **180**, 4737-4762.

D'Urso, P., Massari, R. and Santoro, A. (2011), Robust Fuzzy Regression Analysis, *Information Sciences*, **181**, 4154-4174.

Ferraro, M. B., Coppi, R., González-Rodríguez, G. and Colubi, A. (2010), A Linear Regression Model for Imprecise Response, *International Journal of Approximate Reasoning*, **51**, 759-770.

Fox, J. and Weisberg, S. (2011), *An R Companion to Applied Regression*, 2nd Edt., Sage Publications, Thousand Oaks, CA.

Friedman, J. (1991), Multivariate Adaptive Regression Splines, *The Annals of Statistics*, **19**, 1-141.

Hastie, T., Tibshirani, R. and Friedman, J. H. (2009), *The Elements of Statistical Learning*, 2nd Edt., Springer.

Kelkinnama, M. and Taheri, S. M. (2012), Fuzzy Least-Absolutes Regression Using Shape Preserving Operations, *Information Sciences*, **214**, 105-120.

Kriner, M. (2007), *Survival Analysis with Multivariate Adaptive Regression Splines*, Ph.D. Dissertation, Fakultät für Mathematik, Informatik und Statistik, Ludwig-Maximilians-Universität München.

Lee, T. S. and Chen, I. F. (2005), A Two Stage Hybrid Credit Scoring Model Using Artificial Neural Networks and Multivariate Adaptive Regression Splines, *Expert Systems with Applications*, **28**, 743-752.

Lee, T. S., Chiu, C. C., Chou, Y. C. and Lu, C. J. (2006), Mining The Customer Credit Using Classification and Regression Tree and Multivariate Adaptive Regression Splines, *Computational Statistics and Data Analysis*, **50**, 1113-1130.

Lu, J. and Wang, R. (2009), An Enhanced Fuzzy Linear Regression Model with More Flexible Spreads, *Fuzzy Sets and Systems*, **160**, 2505-2523.

Xu, R. and Li, C. (2001), Multidimensional Least-Squares Fitting with a Fuzzy Model, *Fuzzy Sets and Systems*, **119**, 215-223.

Zimmermann, H. J. (2001), *Fuzzy Set Theory and Its Applications*, 4th Edt., Kluwer Nihoff, Boston.