معرفی رگرسیون منطقی و کاربردآن برای پیش بینی بیماریها

یدالله محرابی[۱]، پروین سربخش[۲]، علی اکبر خادم معبودی[۲]
[۱] گروه اپیدمیولوژی، دانشگاه علوم پزشکی شهید بهشتی
[۲] گروه آمار زیستی، دانشگاه علوم پزشکی شهید بهشتی

در بسیاری از مسائل آماری ، متغیرها اثرات برهمکنشی روی یکدیگر دارند. روشهای آماری موجود برای تعیین مدلهای پیش‌بینی از جمله روشهای رگرسیونی و درختهای تصمیم، قابلیت تشخیص و لحاظ کردن چنین اثراتی را ندارند و اثرات متقابل بین متغیرها در صورت شناسایی و لحاظ کردن در مدل، به دلیل پیچیده شدن آن، نهایتا از دوطرفه و سه طرفه تجاوز نمی کند. برای غلبه بر این نقص این مطالعه به معرفی رگرسیون منطقی به عنوان یک روش رگرسیونی تعمیم یافته و جدید می پردازد که در آن متغیرهای پیشگو به صورت ترکیبات بولی از متغیرهای دوحالتی ساخته می شوند. برای یافتن چنین ترکیباتی در فضای حالتهای ممکن و همچنین برآورد پارامترهای مربوط به این ترکیبات از الگوریتم جستجوی Annealing استفاده می شود.آزمونهای تصادفی سازی برای تایید وجود ارتباط بین داده ها بکار میرود. به منظور اجتناب از بیش برآورد شدن، تعداد بهینه ترکیبات منطقی و متغیرهای مدل به روش اعتبار متقاطع تعیین می گردد. به عنوان کاربردی از این روش داده های حاصل از مطالعه کوهورت قند و لیپید تهران، با استفاده از رگرسیون منطقی تحلیل شدند که در آن اثر متغیرهای تن‌سنجی، قند و لیپیدها، فشار خون و ... بر بروز دیابت بررسی شد و در نهایت مدلی برای پیش‌بینی ابتلا به دیابت ارائه گردید.

واژگان کلیدی: رگرسیون منطقی، منطق بولی، اثرات متقابل، الگوریتم Annealing.

۱ مقدمه

رگرسیون یکی از مهمترین ابزارهای آماری در زمینه آنالیز داده‌ها و بررسی ارتباط بین متغیرهای پیش بین و متغیر پاسخ است. ولی در اکثر مسائل، یک مدل رگرسیونی تنها می‌تواند ارتباط اثرات اصلی متغیرهای پیش‌بین را روی پاسخ بررسی کند و اثرات متقابل بین متغیرها در صورت لحاظ شدن در مدل، به دلیل پیچیده شدن آن، از دوطرفه و نهایتا سه طرفه تجاوز نمی کند.[1] زمانی‌که تعداد متغیرهای پیش بین زیاد باشد و به ویژه این

متغیرها دو حالتی باشند (بله و خیر ، سالم و بیمار و...) اثرات متقابل مراتب بالاتر بین این متغیرها میتواند روی برازش متغیر پاسخ تاثیر بگذارد. این موضوع بیشتر در مسائلی مثل داده‌کاوی و دادههای ریزآرایه که حجم داده ها زیاد است روی می دهد برای شناسایی و لحاظ کردن چنین تقابلهایی در مدلهای رگرسیونی، می‌توان به جای استفاده از تمام متغیرها در برازش مدل، یک متغیر ترکیبی از آنها ساخت و به عنوان متغیر مستقل جدید وارد مدل کرد. رگرسیون منطقی راه حلی برای رفع این گونه مشکلات می باشد.[1]

۱-۱ تعریف مدل رگرسیون منطقی

رگرسیون منطقی یک روش رگرسیونی تعمیم یافته و جدیدی است که در آن متغیرهای پیشگو به صورت ترکیبهای بولی از متغیرهای دو حالتی ساخته می‌شود. در رگرسیون منطقی، به دنبال یک متغیر دو حالتی هستیم که حاصل یک ترکیب منطقی بولی مطلوب از متغیرهای دو حالتی اولیه است طوری که استفاده از این متغیر جدید به عنوان متغیر پیش‌بین، در مقایسه با سایر ترکیبات بولی ممکن، بهترین برازش را برای متغیر پاسخ داشته باشد. این رگرسیون در سال ۲۰۰۲ توسط اینگو روزینسکی [1] معرفی شده است و در زمینه داده‌های SNP، توالی ژنی، غربالگری بیماریهای چند عاملی و... کاربرد دارد و به دلیل استفاده از ترکیبات بولی منطقی رگرسیون منطقی (Logic Regression) نامیده شده‌است.[1]

فرض کنید $x_1, x_2, ..., x_k$ متغیرهای پیشگوی دو حالتی و y متغیر پاسخ است.هدف برازش مدل رگرسیونی به این فرم است:

$$g(E(y)) = \beta_0 + \sum_{j=1}^{t} \beta_j L_j$$

که در آن L_j یک عبارت بولی از متغیرهای پیشگوی X_i و$g(E(y))$ یک تابع پیوند است. این مدل یک مدل منطقی نامیده می‌شود. قالب ارائه شده در بالا می تواند شامل رگرسیون خطی $g(E(y)) = E(y)$ و رگرسیون لجستیک دو حالتی $g(E(y)) = \log\left[\frac{E(y)}{1 - E(y)}\right]$ ، مدل مخاطرات متناسب کاکس و یا سایر مدلهای خطی تعمیم یافته باشد. به طور کلی برای هر مدلی یک تابع امتیاز (score)تعریف می‌شود که نشان دهنده کیفیت

[1] Ingo Ruczinski

مدل مفروض می باشد. برای مثال در رگرسیون خطی تابع امتیاز مجموع مربعات خطا $RSS(\beta) = \sum (y_i - y_i')^2 = (Y - X\beta)'(Y - X\beta)$ و در رگرسیون لجستیک آماره انحراف (Deviance) $D = \sum d(y_i, \widehat{\pi}_i)^2$ می باشد. در رگرسیون منطقی هدف یافتن عبارت بولی است که تابع امتیازتعیین شده را مینیمم کند. برآورد β_j به طور همزمان، با جستجو برای عبارت L_j با استفاده از الگوریتمAnnealing پیدا می‌شود.[1]

۱-۲ عبارات و درخت منطقی

معمواترین و آسان ترین راه نمایش عبارت بولی مثل A and B or C or D but not E استفاده از عملگرهای منطقی

$$\vee(\text{"}or\text{"}),\ \wedge(\text{"}and\text{"}),\ ^c(\text{"}not\text{"})$$

و استفاده از کروشه ها می باشد. یک مثال از این نحوه نمایش عبارت منطقی فوق است که با استفاده از عملگرها به شکل زیر است:

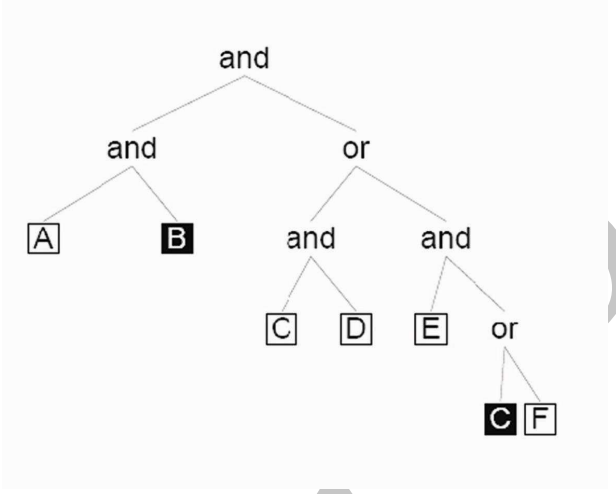
$$\{(A \wedge B^c) \wedge [(C \wedge D) \vee (E \wedge (C^c \vee F))]\}$$

با اسنفاده از کروشه ها هر عبارت بولی را میتوان مکررا با ترکیب دو متغیر، یک متغیر و یک عبارت بولی یا دو عبارت بولی تولید کرد. برای مثال عبارت منطقی فوق را می توان با ترکیبات زیر تولید کرد:

$$\underbrace{\underbrace{(A \wedge B^c)}_{1} \wedge \underbrace{[\underbrace{(C \wedge D)}_{2} \vee \underbrace{(E \wedge \underbrace{(C^c \vee F)}_{3})}_{4}]}_{5}}_{6}$$

این شکل نمایش ما را قادر می سازد که عبارت منطقی را در قالب یک درخت دوحالتی نشان دهیم. درخت منطقی برای عبارت منطقی فوق را می توان به صورت زیر رسم کرد که در آن

حروف با رنگ سفید در زمینه نشانگر نقیض آن حرف یا متغیر است.



عبارات و اصطلاحات زیر برای درخت منطقی به کار میرود: ۱. موقعیت هر عنصر(متغیر، نقیض متغیر و عملگر) در درخت یک گره است. ۲. هر گره صفر یا دو زیر گره دارد ۳. زیر گرهها همسایه های یکدیگرند ۴. گرهی که زیر گره نیست ریشه نامیده می شود ۵. گرهی که زیر گره ندارد برگ نامیده می شود ۶. برگ فقط می تواند متغیر یا متتم متغیر باشد بقیه گرهها اپراتورها هستند

۳-۱ جابجایی در فضای جستجو

یک همسایه برای درخت منطقی درختی است که می تواند از یک جابجایی تکی از درخت اولیه دست آید. هر جابجایی یک جابجایی برگشت پذیر است یعنی یک حرکت برای برگشتن از درخت جدید به درخت قدیمی وجود دارد. برگشت پذیری یک اصل مهم برای نظریه زنجیر مارکف در الگوریتم

Simulated Annealing است.[1] برای ایجاد همسایگی جدید برای درخت اولیه جابجایی های زیر وجود دارد که هر کدام با مثالی در شکل توضیح داده شده اند. در این شکل درخت اولیه در گوشه پایین سمت چپ شکل قرار دارد. سایر درختها با یک تک جابجایی از این درخت اولیه حاصل شده اند. ۱. تعویض برگها (Alternate Leaf): برداشتن یک برگ و جایگزینی آن با برگ دیگر در همان نقطه.(شکل a-1)

۲. تعویض عملگرها (Alternate Operator): هر $\wedge$ می تواند با $\vee$ عوض شود و برعکس .(

شکل b-1)

۳. رویش (Grow): در هر گرهی که برگ نیست (عملگر است) می توان یک شاخه جدید ایجاد کرد. (شکل c-1)

۴. برش (Pruning): حرکت برگشتی رویش، برش است که میتوان شاخه ای را برید.(شکل d-1)

۵. تقسیم برگ (Split Leaf): هر برگ می‌تواند شکسته شود و برگ جدیدی اضافه شود. (شکل e-1)

۶. حذف کردن برگ (Delete Leaf): هر برگ (متغیر) می‌تواند از درخت حذف شود. (شکل f-1)

با توجه به نظریه نافروکاستنی[2] بودن زنجیر مارکف با این سری جابجایی های داده شده یک درخت منطقی می تواند از هر درخت منطقی دیگر با یک تعداد جابجایی متناهی بدست آید.

۱-۴ جستجوی بهترین مدلها

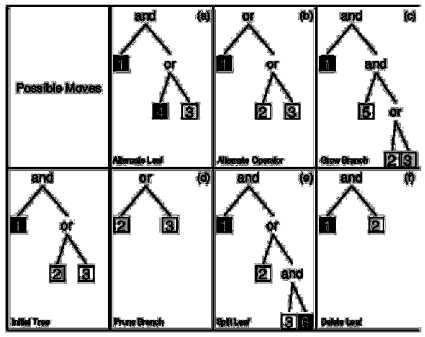
در عمل می‌توان با یک سری متغیر داده شده تعداد بسیار زیادی ترکیب منطقی ساخت و روش مستقیم نیز برای فهرست کردن همه درختهای منطقی وجود ندارد که بتوان برای گزینش بهترین مدل، همه پیش بینی های متفاوت رادر اختیار داشت پس امکان ارزیابی کامل از همه درختهای منطقی ممکن وجود ندارد. برای یافتن بهترین ترکیب منطقی ازالگوریتم‌های Annealing Simulated استفاده میشود.

معیار مطلوب بودن در این جستجو، کمتر بودن تابع امتیاز[3] متناسب با مدل رگرسیونی در نظر گرفته شده میباشد. برای یافتن چنین ترکیباتی در فضای حالتهای ممکن[4] مربوط به این ترکیبها و نیز پارامترهای مربوط به ترکیبات یافت شده از الگوریتم جستجوی Annealing استفاده می‌شود. هر ترکیب بولی را می توان به صورت یک درخت منطقی متشکل از برگهایی که متغیرها ی مطلوب هستند، نشان داد.

[2]Irreducibility in Markov chain theory

[3]Score function

[4]Score function



الگوریتم Annealing Simulated یک الگوریتم جستجوی تصادفی است و در فضای حالتهای ممکن ترکیبات منطقی، بر مبنای تابع امتیاز تعیین شده دنبال بهترین ترکیب می‌باشد

الگوریتم Annealing روی فضای حالات S (حالتهای ممکن ترکیبات منطقی)تعریف می‌شود. حالتها به خاطر سیستم همسایگی با هم مرتبط هستند و یک مجموعه از زوجهای همسایه در S با یک زیرساختار M در $S \times S$ تعریف می شود. عناصر در M، جابجایی نامیده می‌شود. اگر حالت s بتوانند با یک تک جابجایی به s^t تبدیل شود s, s^t را مجاور هم می‌نامند $(s, s^t \in M)$. به طور مشابه $s, s^t \in M^k$ عناصری هستند که با k حرکت به هم می‌رسند. این الگوریتم در بین جابجایی های ممکن، با تابع امتیاز تعریف شده دنبال جابجایی می گردد که منجر به بهتر شدن امتیاز مدل شود.[1]

۱-۵ آزمون اعتبار متقاطع

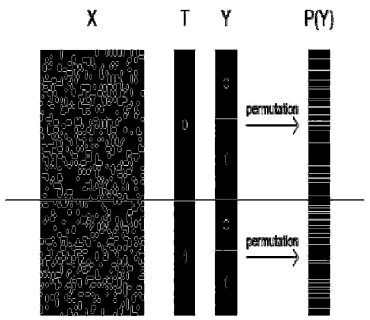
تعداد کل متغیرهای موجود در مدل منطقی به عنوان اندازه مدل در نظر گرفته می‌شود . زمانی که دنبال بهترین مدل از لحاظ امتیاز می گردیم ممکن است به مدلی برسیم که تعداد متغیرهای بیشتری از آنچه مدل بهینه دارد برسیم می توان با مقایسه عملکرد بهترین مدلها در ابعاد مختلف، مدل با بعد بهینه را انتخاب کرد. زمانی که به اندازه کافی داده موجود باشد می توان از روش مجموعه آموزش ـ آزمون[5] استفاده کرد به این صورت که به طور تصادفی داده‌ها به دوگروه با اندازه از پیش تعیین شده تقسیم میشوند با استفاده از یک قسمت از داده‌ها

[5]Training & Test set

،به عنوان مجموعه training(آموزش) و قسمت دیگر به عنوان مجموعه Test(آزمون) ،اندازه مدل مطلوب را بدست آوریم. بنابراین به جای استفاده از کل داده‌ها در برازش و ارزیابی مدل، مدلهایی با اندازه های ثابت را با استفاده از گروه آموزش برازش می دهیم و سپس همه مدلها را روی داده‌های گروه آزمون اعمال می کنیم و مدلی انتخاب می‌شود که بهترین امتیاز را داشته باشد.[1]

۱-۶ آزمون تصادفی سازی : مدل صفر

وجود ارتباط بین پاسخ و متغیر مستقل با مقایسه امتیازهای حاصل از برازش تصادفی پاسخ و بهترین مدل منطقی یافت شده بوسیله الگوریتم را می توان با این آزمون بررسی کرد. فرض صفر این است که هیچ ارتباطی بین X, Y وجود ندارد. اگر ارتباطی بین X, Y وجود نداشته باشد بهترین مدل منطقی امتیازی مشابه مدل تصادفی خواهد داشت با تکرار برازش تصادفی نسبت امتیازهای این مدل که کمتر از بهترین مدل منطقی بپد را به عنوان p مقدار دقیق برای آزمون صفر در نظر می گیریم.[1]



۱-۷ کاربرد روش رگرسیون منطقی برای پیش بینی دیابت

دیابت نوع دو یکی از بیماریهای چندعاملی است که با توجه به اهمیت و بار فردی و اجتماعی، لزوم شناسایی افراد پرخطر برای ابتلا به آن مشهود است. تاکنون مطالعات متعددی جهت پیش بینی بروز دیابت با استفاده از مدل های آماری موجود انجام شده است ولی

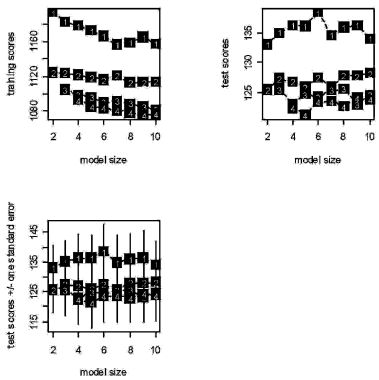
علی رغم اهمیت بالینی اثرات متقابل عوامل خطر روی بروز دیابت، امکان لحاظ کردن همه اثرهای متقابل ممکن در مدل‌های آماری فعلی وجود ندارد. در این مطالعه، به منظور یافتن ترکیبات منطقی مناسب از عوامل خطر مرتبط با دیابت نوع ۲ از روش رگرسیون لجستیک منطقی استفاده گردید.

جمعیت مورد بررسی، از افراد بخش کوهورت مطالعه قند و لیپید تهران (TLGS) [2] انتخاب شدند.۳۵۲۳ نفر (۵۷/۸% زن و ۴۲/۲% مرد) در تحلیل وارد شدند. متغیرهای مورد بررسی طبق تعریف عوامل خطر[3] ، به متغیرهای دو حالتی (عامل خطر دارد/ ندارد) تبدیل شدند

تحلیل‌های مربوط با استفاده از روش رگرسیون لجستیک منطقی(Logistic Logic Regression) با تابع امتیاز "آماره انحرف" انجام شد. پارامترهای مدل با به کارگیری الگوریتم Annealing برآورد شد. به منظور اجتناب از بیش برآورد شدن، تعداد بهینه ترکیبات منطقی و متغیرهای مدل به روش اعتبار متقاطع تعیین گردید.

چهارده عامل خطر دوحالتی در ارتباط با بروز دیابت وارد مدل رگرسیون لجستیک منطقی شدند. تاثیر تغییرات بعد مدل (تعداد متغیرهای مشمول در مدل) روی آماره انحراف مدل رگرسیون لجستیک منطقی برازش داده شده با ۱و۲و۳ و ۴ ترکیب منطقی واندازه های متفاوت از ۲ تا ۱۰ برگ، در شکل ۲ نشان داده شده است. امتیازات آزمون اعتبار متقاطع. برای تعیین بعد مناسب مدل ، پیشنهاد انتخاب مدلی با۴ ترکیب منطقی و ۵ متغیر را می دهد.

به این ترتیب بعد از تعیین اندازه مناسب مدل با استفاده از آزمون اعتبار متقاطع، درصدد یافتن بهترین مدل با ۴ ترکیب منطقی و ۵ متغیر هستیم. الگوریتم Annealing با جستجو در فضای حالات چنین مدلهایی، ترکیبی از متغیرها را می‌یابد که کمترین آماره انحراف را دارند. چون این الگوریتم یک الگوریتم تصادفی و احتمالاتی است الزاما نتیجه منحصر به فرد و یکتایی در جستجو حاصل نمی‌شود بنابراین مدلی که در ۱۰ بار اجرای الگوریتم با فراوانی نسبی بالایی مشاهده شد به عنوان مدل منتخب الگوریتم معرفی گردید. سایر مدلهای مشاهده شده نسبت به مدل منتخب در این ۱۰ بار تغییرپذیری کمی داشتند. مدل حاصل در شکل ۱ ارائه شده است.



شکل (۱): آزمون اعتبار متقاطع

برای ارزیابی و مقایسه مدل منطقی بدست آمده، آماره انحراف و میزان حساسیت و ویژگی مدل محاسبه شد و با مقادیر حاصل از رگرسیون لجستیک معمولی که در آن فقط اثرات اصلی متغیرها وارد شدند مقایسه شد. برای مقایسه صحت مدل‌ها در پیش‌بینی بروز دیابت منحنی مشخصه عملکرد[۶] ROC برای هرکدام از آن‌هارسم و سطح زیر آنها محاسبه گردید. برای ارزیابی قابلیت تعمیم دهی مدل‌ها نیز، فواصل اطمینانی برای سطح زیر منحنی‌های ROC به دو روش معمول و خودگردان (Bootstrap) بدست آمد. نقطه برش برای محاسبه حساسیت و ویژگی مدل طبق رفرنس [4] محاسبه شد.

$$\sqrt{(۱ - Sensitivity)^۲ + (۱ - Specificity)^۲}$$

از نرم افزارهای R نسخه ۲.۸.۱ برای برازش رگرسیون لجستیک منطقی استفاده شد.

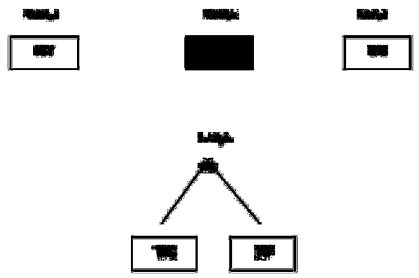
۲ یافته ها

۳۲۵۳ از مجموع نفر شامل ۸۳۰۲ زن (۸/۵۷%) و ۵۸۴۱ مرد (۲/۴۲%) مورد بررسی، تعداد ۰۸ نفر از مردان (۴/۵%) و ۳۳۱ نفر از زنان (۵/۶%) در طول مدت پیگیری به

[6] Receiver Operating Charactristic

دیابت مبتلا شدند که تفاوت معنی داری در میزان ابتلا بین این دو گروه مشاهده نشد ($p = ۰/۱$). . مقایسه عوامل خطر مرتبط با دیابت در دو گروه دیابتی و غیر دیابتی نشان داد که همه عوامل خطر غیر از فعالیت بدنی، جنسیت ، سیگار کشیدن و HDL روی بروز دیابت تاثیر معنی داری داشتند. (جدول ۲)

الگوریتم Anealing برای رگرسیون لجستیک منطقی با ۴ ترکیب بولی و ۵ متغیر مشمول در مدل، ترکیباتی به این صورت را یافت: داشتن اختلال تحمل قند ناشتا با نسبت شانس ((۷/۵۹ و ۴/۰۳):OR=۵/۵۳ CI %95)، داشتن اختلال تحمل قند دوساعته با نسبت شانس ((۷/۴۹ و ۳/۹۶):OR=۵/۴۵ CI %95) نداشتن سابقه فامیلی دیابت با نسبت شانس ((۲/۶۳ و ۱/۳۸) :OR=۱/۸۹ CI %95) ، تری‌گلیسرید بالا داشتن یا دور کمر بزرگ داشتن با نسبت شانس ((۳/۳۲ و۱/۷۳) OR=۲/۴ CI %95). همه این ترکیبات با p-value کمتر از ۰/۱۰۰ معنی‌دار بودند. ۳ تا از ترکیبات به صورت اثر اصلی و یکی از آنها به صورت اثر متقابل «تری‌گلیسرید بالا یا دور کمر بالا« ظاهر شد که در جدول ۳ نشان داده شده است. نمایش درختی این مدل نیز در شکل ۳ آمده است. در این شکل درخت اول قند خون ناشتای بالا داشتن، درخت دوم که به صورت ترکیب متضاد ظاهر شده و سابقه فامیلی دیابت نداشتن را نشان میدهد ، درخت سوم قند دوساعته بالا داشتن و درخت چهارم دور کمر بالا داشتن یا تری گلیسرید بالا داشتن (که شامل حالت هر دو غیر نرمال باشد نیز می شود) به عنوان ترکیبات بولی موثر بر دیابت را نشان می دهند. نتایج آخرین گام رگرسیون لجستیک مرسوم پیشرو نیز شامل همین متغیرها ولی فقط با اثر اصلی شان بود(نتایج نشان داده نشده است). جدول مربوط به سطح زیر نمودار مدلها و فواصل اطمینان پارامتری و ناپارامتری نیز در جدول ۴ آمده‌است. دراین فواصل به دلیل بالا بودن حجم نمونه، قضیه حد مرکزی بخوبی برقرار بود و فواصل اطمینان مشابهی در دو روش مجانبی و Bootstrap مشاهده شد. آماره انحراف مدل لجستیک منطقی براب‌ر ۱۲۰۳/۳ و برای مدل لجستیک پیشرو ۱۲۰۶/۸ محاسبه شد. برای مدل منطقی ۴ درختی، حساسیت مدل ۷۴٪ و ویژگی آن ۸۳٪ و برای مدل لجستیک پیشرو، حساسیت



شکل (۲) نمایش درختی ترکیبات بولی یافت شده با الگوریتم Annealing در مدل لجستیک منطقی با ۴درخت و ۵ برگ برای پیش‌بینی بروز دیابت. متغیرهای مشاهده شده در درختها در جدول ۱ معرفی شده اند. متغیرهایی که با زمینه سیاه در درخت ظاهر شده‌اند در مدل به صورت نقیض آن متغیر تفسیر می شوند

جدول ۱: ترکیبات بولی یافت شده با الگوریتم Annealing و ضرایب مربوط به هر کدام در مدل لجستیک منطقی با ۴ترکیب بولی و ۵ متغیر برای پیشبینی بروز دیابت

ترکیبات بولی	درخت	ضریب	خطای معیار	نسبت بخت	فاصلهٔ اطمینان ۹۵%		p-value
					حد پایین	حد بالا	
IFG	درخت ۱	۱/۷۱	۰/۱۶	۵/۵۳	۴/۰۳	۷/۵۹	<۰/۰۰۱
$(FH-DM)^c$	درخت ۲	-۰/۶۳	۰/۱۶	۰/۵۳	۰/۳۸	۰/۷۲	<۰/۰۰۱
IGT	درخت ۳	۱/۶۹	۰/۱۶	۵/۴۵	۳/۹۶	۷/۴۹	<۰/۰۰۱
$TG \vee WC$	درخت ۴	۰/۸۷	۰/۱۶	۲/۴۰	۱/۷۳	۳/۳۲	<۰/۰۰۱
ثابت مدل	ثابت مدل	-۳/۸۵	۰/۱۸	۰/۰۲			<۰/۰۰۱

جدول ۲: سطح زیر نمودار وفواصل اطمینان پارامتری و Bootstrap

مدل	سطح زیر نمودار ROC	خطای معیار	فاصلهٔ اطمینان ۹۵% با فرض توزیع نرمال		فاصلهٔ اطمینان ۹۵% با استفاده از Bootstrap	
			p-value	حد پایین	حد بالا	حد پایین	حد بالا
رگرسیون لجستیک معمولی	۰/۸۳۹	۰/۰۱۶	<۰/۰۰۱	۰/۸۰۸	۰/۸۷۱	۰/۸۰۸	۰/۸۶۹
مدل منطقی با ۴ ترکیب بولی	۰/۸۴۳	۰/۰۱۵	<۰/۰۰۱	۰/۸۱۳	۰/۸۷۴	۰/۸۱۲	۰/۸۷۲

۳ بحث و نتیجه گیری

علاوه بر رگرسیون منطقی، روشهای ارزشمندی در ساختن قواعد دو حالتی وجود دارد از جمله درخت تصمیم، قواعد تصمیم در علوم کامپیوتر، یادگیری ماشینی[7] و ... ولی رگرسیون منطقی در مقایسه با سایر روشهای تصمیم‌گیری برای متغیرهای دوحالتی، تنها روشی است که به دنبال ترکیبات بولی از متغیرهای دو حالتی در کل فضای حالت چنین ترکیباتی می باشد. قابلیت لحاظ کردن اثرات متقابل بین چندین متغیر در قالب یک عبارت بولی و تلخیص متغیرها به طوری که مدل نهایی همچنان قالب یک مدل رگرسیونی را داشته و ضرایب به سادگی تفسیر و آزمون می شونداز مزایای این روش نسبت به روشهای موجود ا ست.[1]

متغیرهای وارد شده در آخرین گام لجستیک معمولی همان متغیرهای مشمول در مدل منطقی با ۴ ترکیب منطقی و ۵ متغیر بودند با این تفاوت که در مدل لجستیک ۵ اثر اصلی

[7] Machine learnin

وجود دارد در حالی که در مدل منطقی ۳ ترکیب به صورت اثر اصلی و یک ترکیب به صورت اثر متقابل (دور کمر یا تری‌گلیسرید) ظاهر شده‌است. با توجه به این تحلیل، شاید بتوان نتیجه گرفت که در مورد بروز دیابت با استفاده از عوامل خطر ذکر شده، وجود اثرات متقابل و لحاظ نکردن آنها چندان نگران کننده نیست و آنچه مهم است اثرات اصلی متغیرهاست و ظاهرا متغیرها به‌صورت مستقل از هم در بروز دیابت نقش دارند. تنها اثر متقابل مشاهده شده در این مدل مربوط به دور کمر و تری‌گلیسرید است که در نظر گرفتن همین اثر متقابل نیز باعث کاهش آماره انحراف مدل از ۱۲۰۶/۸۸ برای لجستیک حاصل از اثرات اصلی به ۱۲۰۳/۰۳ برای مدل منطقی و افزایش قدرت پیش‌بینی مدل شده است. سطح زیر نمودار مدل منطقی با ۵ متغیر برابر۰/۸۴۳ و سطح زیر نمودار لجستیک حاصل از اثرات اصلی ۰/۸۳۹ می‌باشد.

در مورد پیشبینی دیابت با رگرسیون لجستیک منطقی، مطالعه مشابهی یافت نشد ولی در مطالعاتی که به بررسی قواعد تصمیمگیری در مورد دیابت پرداخته‌اند عوامل خطر مشابه با پژوهش حاضر بدست آمده است. از جمله ویلسون و همکارانش [5] در سال ۲۰۰۷ در مطالعه‌ای برای پیشبینی بروز دیابت در افراد بالای ۵۰ سال، عوامل خطر سن بالا، دور کمر بالا، سابقه فامیلی دیابت، اختلال تحمل قند ناشتا، تری گلیسرید بالا و HDL پایین را به عنوان متغیرهای پیشبین معرفی میکنند. در مطالعه دیگری که بارک و همکاران [6] در سال ۱۹۹۸ انجام دادند نژاد، چاقی، بیماری‌های قلبی، تری‌گلیسرید بالا و اختلال تحمل قند ناشتا و دوساعته را به عنوان عوامل موثر در بروز دیابت معرفی کردهاند.

دوحالتی کردن متغیرهای تحقیق به صورت وجود و یا عدم وجود عامل خطر، و ارائه مدلی با استفاده از این متغیرهای دوحالتی از لحاظ بالینی و سادگی استفاده از آن بدون نیاز به اطلاعات جزئی افراد، حائز اهمیت و ارزش فراوان می‌باشد. اختلاف مدل منطقی مشاهده شده با لجستیک معمولی کم است که حاکی از ناچیز بودن اثرات متقابل بین عوامل خطر دوحالتی دیابت است.

هرچند یکی از محدودیت‌های رگرسیون منطقی ممکن است مشکل داده های گمشده ولی مزیتی که نسبت به روش‌هایی مانند مدل شبکه‌های عصبی مصنوعی دارد این است که کاملا به شکل یکی از رگرسیون‌های موجود از جمله خطی، لجستیک یا کاکس، بسته به نوع مطالعه است و قابلیت تفسیر ضرایب، ارزیابی مدل با امتیازها و آماره های مربوط به نوع رگرسیون استفاده شده در آن وجود دارد. همچنین قابلیت لحاظ کردن اثرات متقابل بین چندین متغیر در قالب یک عبارت بولی و تلخیص متغیرها از مزایای این روش نسبت به روشهای قبلی و فعلی است.[1]

قدردانی:

در این تحقیق، از داده های طرح قند و لیپید تهران که توسط پژوهشکده علوم غدد درون ریز و متابولیسم دانشگاه علوم پزشکی شهید بهشتی اجرا شده است، استفاده شد. بر خود لازم می دانیم از کلیه کسانی که در طراحی و جمع آوری داده های TLGS مشارکت داشتند نهایت قدردانی را به عمل آوریم.

این مقاله از پایان نامه کارشناسی ارشد آمار زیستی استخراج شده است.[7]

مراجع

[1] Rzinski I, Koperberg C, LeBlanc M. Logic Regression. *Journal of Computational and Graphical statistics 2003* ; **12(3):475-511**

[2] Azizi F,Rahmani M, Emami H, Mirmiran P,Hajipour R, Madjid M, et al Cardiovascular risk factor in an Iranian urban population: Tehran lipid and glucose study(phase 1). *SozPraventivmed. 2002* ; **47:408-26**

[3] diabetes and their Parents: a Notion Whose Time Has Come? *Diabetes care 2008*; **S14**

[4] Perkins N, Schisterman E. The Inconsistency of "Optimal" Cut-point Using Two ROC Based Criteria. *Am J Epidemial. 2006 April* ; **1;163(7):670-675** .

[5] Wilson P, James B, Sullivan L, Fox C, Nothan D, DAgostino R. Prediction of Incident Diabetes Mellitus in Middle-aged Adult. *Arch Intern Med. 2007* ; **1068-1074**

[6] Burker J, Haffiner S, Gaskill S, Williams K, Stern M. Reversion from type2 Diabetes to nondiabetic statuse. *Diabetes Care 1998* ; **volume 21, Number 8**

[7] Sarbakhsh P. "Logic Regression and its Application in predicting diabetes anong 20year old and over population in district 13 of Tehran". *MSC thesis, Shahid Beheshti University of medical sciences,2009* .