معرفی رگرسیون منطقی و کاربرد آن برای پیش بینی بیماریها

پروین سربخش[۱]، یدالله محرابی [۲]، علی اکبر خادم معبودی[۳] و فرزاد حدائق[۴]

چکیده:

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، نهایتاً از دوطرفه و سه طرفه تجاوز نمی‌کند. برای غلبه بر این نقص این مطالعه به معرفی رگرسیون منطقی به عنوان یک روش رگرسیونی تعمیم‌یافته و جدید می پردازد که در آن متغیرهای پیشگو به صورت ترکیبات بولی از متغیرهای دوحالتی ساخته می‌شوند. برای یافتن چنین ترکیباتی در فضای حالتهای ممکن و هم‌چنین برآورد پارامترهای مربوط به این ترکیبات از الگوریتم جستجوی Annealing استفاده می‌شود. آزمونهای تصادفی‌سازی برای تأئید وجود ارتباط بین داده‌ها بکار می‌رود. به منظور اجتناب از بیش‌برآورد شدن، تعداد بهینه ترکیبات منطقی و متغیرهای مدل به روش اعتبار متقاطع تعیین می‌گردد. به عنوان کاربردی از این روش داده‌های 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.

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

۱ مقدمه

رگرسیون یکی از مهمترین ابزارهای آماری در زمینه آنالیز داده‌ها و بررسی ارتباط بین متغیرهای پیش‌بین 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 نمی‌کند [۱۵]. زمانی‌که تعداد متغیرهای پیش‌بین زیاد باشد، به ویژه وقتی این متغیرها دوحالتی باشند (بله و خیر، سالم و بیمار و...)، ممکن است اثرات متقابل مراتب بالاتر بین این متغیرها بر روی برازش متغیر پاسخ تاثیرگذار باشد. این موضوع بیشتر در مسائلی مثل داده‌کاوی و داده‌های ریزآرایه که حجم داده‌ها زیاد است روی می‌دهد. برای شناسایی و لحاظ

[۱] دانشجوی دوره دکتری آمار زیستی، دانشکده پیراپزشکی دانشگاه علوم پزشکی شهید بهشتی
[۲] استاد گروه اپیدمیولوژی، دانشکده بهداشت دانشگاه علوم پزشکی شهید بهشتی
[۳] استادیار گروه آمار زیستی، دانشکده‌پیراپزشکی دانشگاه علوم پزشکی شهید بهشتی
[۴] دانشیار پژوهشکده علوم غدد درون ریز و متابولیک،دانشگاه علوم پزشکی شهید بهشتی

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 می‌تواند راه‌حلی برای رفع این گونه مشکلات باشد [۱۵]. برای متغیرهای پیش‌بین دوحالتی روشهای متنوع رگرسیونی و کلاس‌بندی در علوم آماری و کامپیوتر و زبان ماشین وجود دارد. در منابع زبان ماشین، روش‌ها و الگوریتم‌هایی که از توابع بولی استفاده می‌کند بر مبنای درخت تصمیم یا قواعد تصمیم هستند. در میان این الگوریتم ها، الگوریتمهای SLIQ [۱۰]، CART [۲]، ID3 [۱۳]، c4.5 [۱۶] و M5 [۱۴] مشهورتر هستند. روشهای بر مبنای قواعد نیز شامل: [۷RIPPER]، SLIPPER [۸]، CN2 [۵]، SWAP1 [۱۹] و AQ [۱۲] می‌باشند. تفاوت روشهای درختی و قاعده تصمیم‌گیری در رویکرد کلاس‌بندی روش درختی در مقابل رویکرد رگرسیونی روشهای قاعده‌ای است. اکثر روش‌های دوحالتی ارائه شده، از نوع کلاس‌بندی هستند. ابتدا ویس و ایندورخیا [۱۹] روش SWAP1 را بسط دادند به طوری که قواعد رگرسیونی را به فرم ترکیبات منطقی آموزش می‌داد. تارگو و گاما [۱۸] این روش را گسترش دادند به طوری که تقریبا هر مسئله کلاس‌بندی را با الگوریتم R2 و شرطهای منطقی اگر و آن‌گاه به مدل رگرسیونی خطی تبدیل می‌کرد. الگوریتم M5 نیز برای ساختن مدل درختی خطی ابداع شد [۱۴]. روش‌های یادگیری ماشین[۵] مانند شبکه‌های عصبی[۶] [۲۱] و ماشین بردار پشتیبان[۷] [۹] نیز از روشهایی هستند که می‌تواند برای مدل‌بندی داده‌هایی که پیش‌فرضهای لازم برای رگرسیون را ندارند، هم‌چنین برای مدل‌بندی روابط پیچیده غیرخطی یا اثرات متقابل مراتب بالا به کار رود. روشهای یادگیری ماشین، با استفاده از معادلات یادگیری، داده‌ها را کلاس‌بندی می‌کنند. ایراد این روشها این است که مانند یک جعبه سیاه عمل کرده و فقط نتایج کلاس‌بندی را ارائه داده و تابع یا ضابطه‌ای برای این کلاس‌بندی در اختیار محقق قرار نمی‌دهند در نتیجه تاثیر متغیرهای پیش‌بین و شدت اثر آنها روی متغیر پاسخ ارزیابی نمی‌شود. روش رگرسیون منطقی توسط روژینسکی [۱۵]، برای مسائلی با متغیرهای مستقل دوحالتی با رویکرد رگرسیونی و یافتن متغیرهای پیش‌بین از ترکیبات بولی متغیرهای پیش‌بین دوحالتی اولیه ارائه شد. الگوریتم جستجوی استفاده شده در این روش، Simulated Annealing است. این روش به دلیل جستجوی ترکیبات بولی در کل فضای حالات ممکن چنین ترکیباتی و نیز ارائه تابع امتیاز برای مقایسه کفایت مدلها، نسبت به سایر مدلها ارجحیت دارد. برخلاف روشهای دیگر، رگرسیون منطقی به شکل یک مدل رگرسیونی بوده و برای هر ترکیب یافت شده میزان اثر و ضریب ارائه می‌هد که قابلیت تفسیر بهتر و ارزیابی مدل با استفاده از امتیازها و آماره‌های مربوط به نوع مدل رگرسیونی ایجاد می‌کند. 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 است. در این روش می‌توان به شرط

[۵] Machine learnin
[۶] Neural Networks
[۷] Suport Vector Machine

تعریف تابع امتیاز مناسب، هر نوع مدل رگرسیونی اعم از خطی،لجستیک و کاکس و... را برازش داد [۱۵].

۱.۱ تعریف مدل رگرسیون منطقی

رگرسیون منطقی یک روش رگرسیونی تعمیم‌یافته و جدید است که در آن متغیرهای پیشگو به صورت ترکیبهای بولی از متغیرهای دوحالتی ساخته می‌شود. در رگرسیون منطقی، به دنبال یک متغیر دوحالتی هستیم که حاصل یک ترکیب منطقی بولی مطلوب از متغیرهای دوحالتی اولیه باشد به‌طوری که استفاده از این متغیر جدید به عنوان متغیر پیش‌بین، در مقایسه با سایر ترکیبات بولی ممکن، بهترین برازش را برروی متغیر پاسخ داشته باشد. این روش رگرسیون که توسط روزینسکی[۸] 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[۱۵]. فرض کنید $x_1, x_2, ..., x_k$ متغیرهای پیشگوی دوحالتی و y متغیر پاسخ باشد. هدف برازش مدل رگرسیونی به فرم زیر 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 می‌تواند شامل رگرسیون خطی $g(E(y)) = E(y)$ و رگرسیون لجستیک دوحالتی $g(E(y)) = \log\left[\frac{E(y)}{1-E(y)}\right]$، مدل مخاطرات متناسب کاکس و یا سایر مدلهای خطی تعمیم یافته باشد. به طور کلی برای هر مدلی یک تابع امتیاز[۹] تعریف می‌شود که نشان‌دهنده کیفیت مدل مفروض می‌باشد. 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 $D = \sum d(y_i, \hat{\pi}_i)^2$ (Deviance) می‌باشد. در رگرسیون منطقی هدف یافتن عبارت بولی است که تابع امتیاز تعیین‌شده را مینیمم کند. برآورد β_j به طور هم‌زمان، با جستجو برای عبارت L_j با استفاده از الگوریتم Annealing پیدا می‌شود [۱۵].

۲.۱ عبارات و درخت منطقی

معمول‌ترین و آسان‌ترین راه نمایش عبارت‌های بولی استفاده از عملگرهای منطقی:

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

و استفاده از کروشه‌ها می‌باشد. یک مثال از این نحوه نمایش عبارت منطقی زیر است که با استفاده از عملگرها نشان داده شده است:

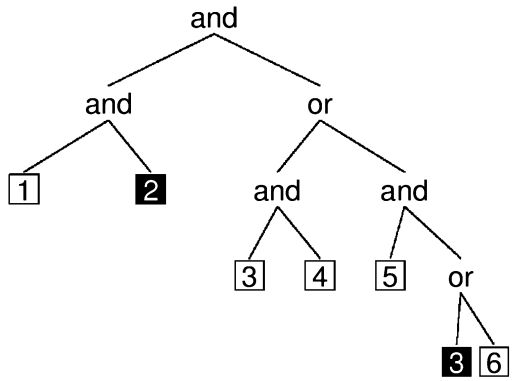
$$\{(A \wedge B^c) \wedge [(C \wedge D) \vee (E \wedge (C^c \vee F))]\}$$

با استفاده از کروشه‌ها هر عبارت بولی را می‌توان مکرراً با ترکیب دو متغیر، یک متغیر و یک عبارت بولی یا دو عبارت بولی تولید کرد. برای مثال عبارت منطقی فوق را می‌توان با ترکیبات زیر تولید کرد:

[۸] Ingo Ruczinski
[۹] Score function

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

این شکل نمایش ما را قادر می‌سازد که عبارت منطقی را در قالب یک درخت دوحالتی نشان دهیم. درخت منطقی برای عبارت منطقی فوق را می‌توان به صورت زیر رسم کرد که در آن حروف با رنگ سفید در زمینه نشان‌گر نقیض آن حرف یا متغیر است.



شکل ۱. نمایش عبارت های منطقی به صورت درخت منطقی

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 همسایه‌های یکدیگرند. ۴. گرهی که زیرگره نیست ریشه نامیده می‌شود. ۵. گرهی که زیرگره ندارد برگ نامیده می‌شود. ٦. برگ فقط می‌تواند متغیر یا متمم متغیر باشد بقیه گره‌هاعملگرها هستند.

۳.۱ جابجایی در فضای جستجو

یک همسایه برای درخت منطقی درختی است که می‌تواند از یک جابجایی تکی از درخت اولیه به دست آید. هر جابجایی یک جابجایی برگشت‌پذیر است یعنی یک حرکت برای برگشتن از درخت جدید به درخت قدیمی وجود دارد. برگشت‌پذیری یک اصل مهم برای نظریه زنجیر مارکف در الگوریتم Simulated Annealing است [۱۵]. برای ایجاد همسایگی جدید برای درخت اولیه جابجایی های زیر وجود دارد که هر کدام با مثالی در شکل ۲ توضیح داده شده‌اند. در این شکل درخت اولیه در گوشه پایین سمت چپ شکل قرار دارد. سایر درختها با یک تک جابجایی از این درخت اولیه حاصل شده‌اند. 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 $\wedge$ می‌تواند با $\vee$ عوض شود و برعکس.(شکل b-1)

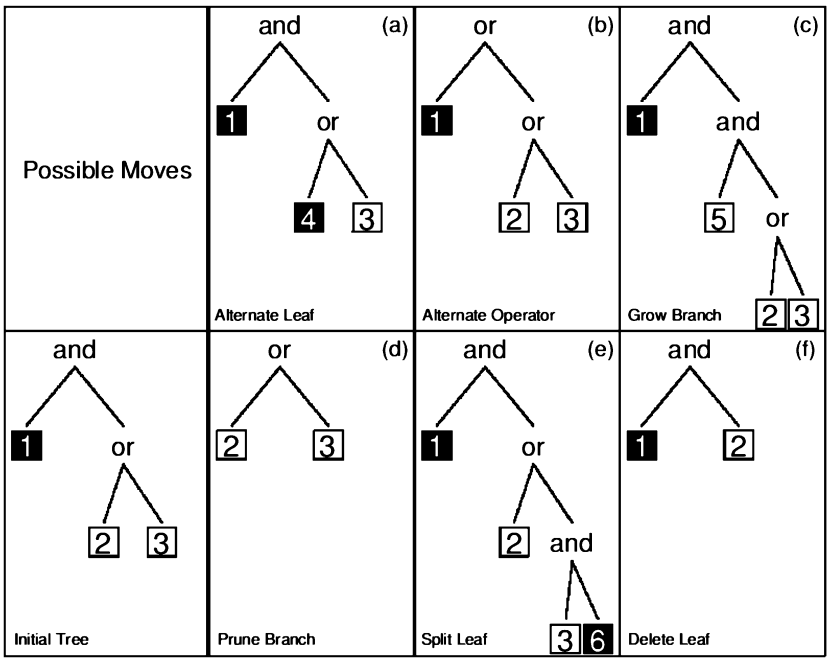
۳. رویش (Grow): در هر گرهی که برگ نیست (عملگر است) می توان یک شاخه جدید ایجاد کرد.(شکل c-1)

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

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

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

با توجه به نظریه نافروکاستنی بودن زنجیر مارکف با این سری جابجایی‌های داده شده یک درخت منطقی می‌تواند از هر درخت منطقی دیگر با یک تعداد جابجایی متناهی بدست آید.



شکل ۲جابجایی های ممکن برای درخت منطقی

۴.۱ 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 پیش‌بینی‌های متفاوت رادر اختیار داشت پس امکان ارزیابی کامل از همه درختهای منطقی ممکن وجود ندارد. برای یافتن بهترین ترکیب منطقی ازالگوریتم‌های Simulated Annealing استفاده می‌شود. معیار مطلوب بودن در این جستجو، کمتر بودن تابع امتیاز متناسب با مدل رگرسیونی در نظر گرفته شده می‌باشد. برای یافتن چنین ترکیباتی در فضای حالتهای ممکن Score function مربوط به این ترکیبها و نیز پارامترهای مربوط به ترکیبات یافت شده از الگوریتم جستجوی Annealing استفاده می‌شود. هر ترکیب بولی را می‌توان به صورت یک درخت منطقی متشکل از برگهایی که متغیرهای مطلوب هستند، نشان داد. الگوریتم Simulated Annealing 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 مرتبطبوده و یک مجموعه از زوجهای همسایه در S با یک زیرساختار M در $S \times S$ تعریف می‌شود. عناصردر M،

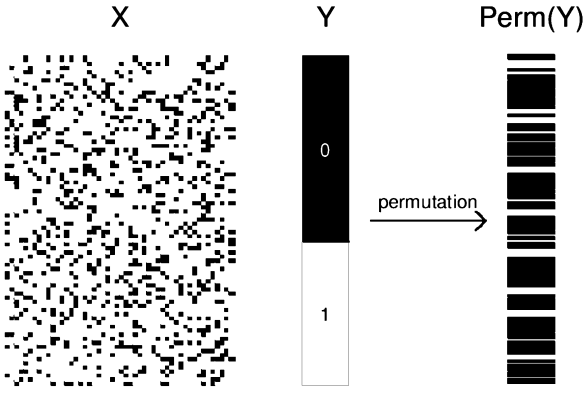
جابجایی نامیده می‌شود. اگر حالت s بتواند با یک تک جابجایی به s^t تبدیل شود s^t, s را مجاور هم می‌نامند $(s, s^t \in M)$. به طور مشابه $s, s^t \in M^k$ عناصری هستند که با k حرکت به هم می‌رسند. این الگوریتم در بین جابجایی‌های ممکن، باتوجه به تابع امتیاز به دنبال جابجایی می‌گردد که منجر به بهتر شدن امتیاز مدل شود [۱۵].

۵.۱ آزمون اعتبار متقاطع

تعداد کل متغیرهای موجود در مدل منطقی به عنوان اندازه مدل در نظر گرفته می‌شود. زمانی که در جستجوی بهترین مدل از لحاظ امتیاز هستیم ممکن است به مدلی با تعداد متغیرهای بیشتری از آنچه مدل بهینه دارد دست پیدا کنیم. می‌توان با مقایسه عملکرد بهترین مدلها در ابعاد مختلف، مدل با بعد بهینه را انتخاب کرد. این روش، آزمون اعتبار متقاطع نامیده می‌شود[۱۰]. زمانی که به اندازه کافی داده موجود باشد می‌توان از روش مجموعه آموزش و آزمون[۱۱] استفاده کرد به این صورت که به طور تصادفی داده‌ها به دو گروه با اندازه از پیش تعیین شده تقسیم می‌شوند. با استفاده از یک قسمت از داده‌ها به عنوان مجموعه training (آموزش) و قسمت دیگر به عنوان مجموعه Test (آزمون)، اندازه مدل مطلوب را بدست می آوریم. بنابراین به جای استفاده از کل داده‌ها در برازش و ارزیابی مدل، مدل‌هایی با اندازه‌های ثابت را با استفاده از گروه آموزش برازش داده و سپس همه مدل‌ها را روی داده‌های گروه آزمون اعمال می‌کنیم و مدلی انتخاب می‌شود که بهترین امتیاز را داشته باشد [۱۵].

۶.۱ آزمون تصادفی‌سازی: 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 می‌توان با این آزمون بررسی کرد. فرض صفر این است که هیچ ارتباطی بین X و Y وجود ندارد. اگر ارتباطی بین X و 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 است را به عنوان p – مقدار دقیق برای آزمون صفر در نظر می‌گیریم. این آزمون، آزمون تصادفی‌سازی: مدل صفر[۱۲] نامیده می‌شود. [۱۵]



شکل ۳.آزمون تصادفی سازی

[۱۰] Cross Validation Test
[۱۱] Training & Test set
[۱۲] Null Model Randomization Test

۷.۱ کاربرد روش رگرسیون منطقی برای پیش‌بینی دیابت [۱۱]

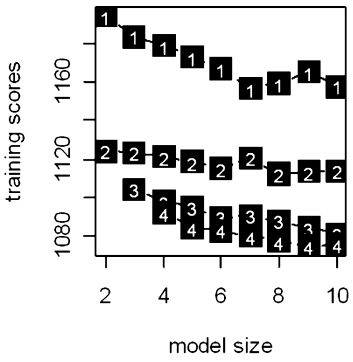
دیابت نوع دو یکی از بیماریهای چندعاملی است که با توجه به اهمیت و بار فردی و اجتماعی، لزوم شناسایی افراد پرخطر برای ابتلا به آن مشهود است. تا کنون مطالعات متعددی جهت پیش‌بینی بروز دیابت با استفاده از مدل‌های آماری موجود انجام شده است ولی علی‌رغم اهمیت بالینی اثرات متقابل عوامل خطر روی بروز دیابت، امکان 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 دو از روش رگرسیون لجستیک منطقی استفاده گردید. الگوریتم مورد استفاده در رگرسیون منطقی برای یافتن ترکیبات بولی، الگوریتم Annealing می‌باشد. رگرسیون منطقی برای توابع پیوندی مثل خطی، لجستیک، مدل کاکس و... قابل اجرا است. بنابراین در این مقاله از رگرسیون منطقی با لینک لجستیک استفاده شد که با الگوریتم Annealing بهترین ترکیبات بولی که منجر به یافتن مدل لجستیک منطقی با کمترین آماره انحراف می شد، جستجو و یافت شد.

جمعیت مورد بررسی، از افراد بخش کوهورت مطالعه قند و لیپید تهران (TLGS) [۱] انتخاب شدند.۳۵۲۳ نفر (۵۷/۸% زن و ۴۲/۲% مرد) مورد مطالعه قرار گرفتند. متغیرهای مورد بررسی طبق تعریف عوامل خطر [۴]، به متغیرهای دوحالتی (عامل خطر دارد/ ندارد) تبدیل شدند.

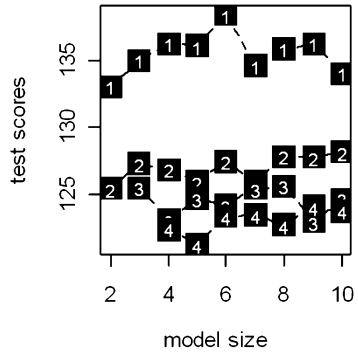
تحلیل‌های مربوط با استفاده از روش رگرسیون لجستیک منطقی(Regression Logistic Logic) با تابع امتیاز "آماره انحراف" انجام شد. پارامترهای مدل با به کارگیری الگوریتم Annealing برآورد شد. به منظور اجتناب از بیش‌برآورد شدن، تعداد بهینه ترکیبات منطقی و متغیرهای مدل به روش اعتبار متقاطع تعیین گردید.

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 واندازه‌های متفاوت از ۲ تا ۱۰ برگ، در شکل ۲ نمایش داده شده‌است. امتیازات آزمون اعتبار متقاطع برای تعیین بعد مناسب مدل، پیشنهاد انتخاب مدلی با۴ ترکیب منطقی و ۵ متغیر را می‌دهد.

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 ۴ ارائه شده است.



امتیازات مربوط به مدلهایی با تعداد درختها وبرگهای متفاوت در گروه آموزش



امتیازات مربوط به مدلهایی با تعداد درختها وبرگهای متفاوت در گروه آزمون

شکل ۴.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 ROC[۱۳] برای هرکدام از آن‌هارسم و سطح زیر آنها محاسبه گردید. برای ارزیابی قابلیت تعمیم‌دهی 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 از فرمول زیر محاسبه شد [٦]

$$\sqrt{(1-Sensitivity)^2+(1-Specificity)^2}.$$

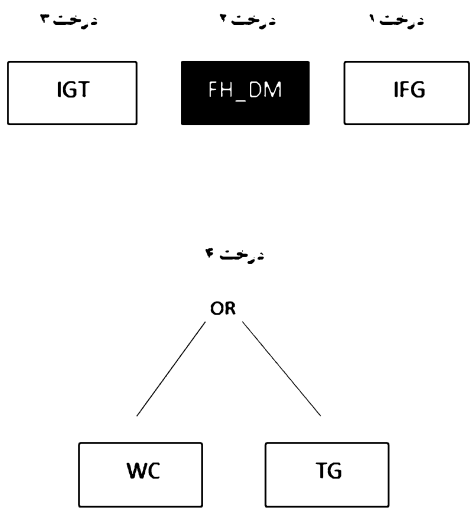
از نرم افزارهای R نسخه ۲.۸.۱ برای برازش رگرسیون لجستیک منطقی استفاده شد.

۲ یافته‌ها

از مجموع ۳۲۵۳ فرد مورد بررسی که شامل ۸۳۰۲ زن (۸/۵۷%) و ۵۸۴۱ مرد (۲/۴۲%) بودند، تعداد ۸۰ نفر از مردان (۴/۵%) و ۱۳۳ نفر از زنان (۵/٦%) در طول مدت پی‌گیری به دیابت مبتلا شدند که تفاوت معنی‌داری در میزان ابتلا بین این دو گروه مشاهده نشد ($p=0/1$). 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 بر روی بروز دیابت تاثیر معنی‌داری داشتند. 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 (OR=۵/۵۳ و CI %95: (۴/۰۳ و ۷/۵۹))، داشتن اختلال تحمل قند دوساعته با نسبت شانس (OR=۵/۴۵ و CI %95 : (۳/۹٦ و ۷/۴۹))، نداشتن سابقه فامیلی دیابت با نسبت شانس (OR=۱/۸۹ و CI %95 : (۱/۳۸ و ۲/٦۳))، تری‌گلیسرید بالا داشتن یا دور

[۱۳] Receiver Operating Charactristic

کمر بزرگ داشتن با نسبت شانس ((۳/۳۲ و۱/۷۳)CI : 95% و OR=۲/۴). همه این ترکیبات با p – مقدار کمتر از ۰/۰۰۱ معنی‌دار بودند. سه 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 ۱ نشان داده شده است. نمایش درختی این مدل نیز در شکل ۵ 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 غیرنرمال باشد نیز می‌شود) را به عنوان ترکیبات بولی موثر بر دیابت نشان می‌دهند. متغیرهایی که با زمینه سیاه در درخت ظاهر شده‌اند در مدل به صورت نقیض آن متغیر تفسیر می‌شوند. متغیرها : (tolerance: family history diabetes triglyceride , IGT: impaired glucose , FH-DM fasting glucose, WC: waist circumference, TG: impaired : IFG) نتایج آخرین گام رگرسیون لجستیک مرسوم پیشرو نیز شامل همین متغیرها ولی فقط با اثر اصلی‌شان بود(نتایج نشان داده نشده است). جدول مربوط به سطح زیر نمودار مدلها و فواصل اطمینان پارامتری و ناپارامتری نیز در جدول ۲ آمده‌است. دراین فواصل به دلیل بالا بودن حجم نمونه، 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 لجستیک پیشرو، حساسیت ۷۵% و ویژگی ۸۲% بدست آمد.



شکل ۵نمایش درختی ترکیبات بولی یافت شده در مدل منطقی

۳ بحث و نتیجه‌گیری

علاوه بر رگرسیون منطقی، روشهای ارزشمندی در ساختن قواعد دوحالتی وجود دارد از جمله درخت تصمیم، قواعد تصمیم، یادگیری ماشینی و ... ولی رگرسیون منطقی در مقایسه با سایر روشهای تصمیم‌گیری برای متغیرهای دوحالتی، تنها روشی است که به دنبال ترکیبات بولی از متغیرهای دوحالتی در کل فضای حالت چنین ترکیباتی می‌باشد. 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 می‌شوند از مزایای این روش نسبت به روشهای موجود است [۱۵].

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 ∨ 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	p-value	فاصلهٔ اطمینان ۹۵% با فرض توزیع نرمال		فاصلهٔ اطمینان ۹۵% با استفاده از Bootstrap	
				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

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 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 نگران‌کننده نیست و آنچه مهم است اثرات اصلی متغیرهاست و ظاهراً 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 [۲۰] درسال 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 [۳] 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 است. 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 داده‌های گمشده باشد ولی مزیتی که نسبت به روش‌هایی مانند مدل شبکه‌های عصبی مصنوعی دارد این است که کاملاً به شکل یکی از رگرسیون‌های موجود از جمله خطی، لجستیک یا کاکس، بسته به نوع مطالعه است و قابلیت تفیسر ضرایب، ارزیابی مدل با امتیازها و آماره‌های 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 [۱۵].

قدردانی:

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

این مقاله از پایان نامه کارشناسی ارشد آمار زیستی استخراج شده‌است [۱۷].

مراجع

[1] Azizi, F.,Rahmani, M.,et al.(2002), Cardiovascular risk factor in an Iranian urban population: Tehran lipid and glucose study(phase 1). *Social and Preventive Medicine*, 47,408-26

[2] BREIMAN, L. (1996), Bagging Predictors. *Machine Learning* , 123-140

[3] BURKE, J. P., HAFFNER, S. M. , et al.(1998), Reversion from type ۲diabetes to nondiabetic status. Influence of the 1997 American Diabetes Association criteria. *Diabetes Care*, 21, 1266-70

[4] CAMERON, F. J., NORTHAM, E. ,et al.(2007), Routine Psychological Screening in Youth With Type ۱Diabetes and Their Parents. *Diabetes Care*, 30, 2716-2724

[5] CLARK , P. , NIBLETT , T.(1989), The CN2 Induction Algorithm. *Machine Learning*, 3, 261-283

[6] COFFIN, M. , SUKHATME, S.(1997), Receiver operating characteristic studies and measurement errors. *Biometrics* , 53(3), 823-37

[7] COHEN, W. (1995),Fast Effective Rule Induction. *International Conference on Machine Learning*

[8] COHEN, W., SINGER, Y.(1999), A Simple, Fast, and Effective Rule Learner. *in Proceedings of Annual Conference of American Association for Artificial Intelligence*

[9] CORTES, C. , VAPNIK, V.(1995), Support—vector networks. *Machine Learning* , 20, 273-297

[10] MEHTA,M. , AGRAWAL,R. , et al.(1996), SLIQ: A fast scalable classifier for data mining. *In Extending Database Technology* , 18-32

[11] MEHRABI, Y., SARBAKHSH, P., et al.(2010), Prediction of Diabetes Using Logic Regression. *Iranian Journal of Endocrinology and Metabolism* , 12, 16-24

[12] MICHALSKI R.S , M. I., HONG J, et al.(1986), The Multi—Purpose Incremental Learning System AQ15 and Its Testing Application to Three Medical Domains. *In Proceedings of AAAI* , 1041-1047

[13] QUINLAN, J. R.(1986),Induction of Decision Trees. *Machine Learning*, 81-106

[14] QUINLAN, J. R.(1992), Learning with Continuous Classes. *5 th Australian Joint Conference on Artificial Intelligence*

[15] RUCZINSKI, I., KOOPERBERG,C. , et al.(2003),Logic Regression. *Journal of Computational and Graphical statistics*, 12(3):475-511

[16] SALZBERG, S. L.(1994), C 4.5: Programs for Machine Learning by J Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Machine Learning* , 16, 235-240

[17] SARBAKHSH , P. (2009), Logic regression and its application in predicting diabetes among 20year old and over population in district 13 of Tehran. *MSc thesis in Biostatistic. Tehran, Shahid Beheshti University of Medical Sciences*;

[18] TORGO, L. , GAMA, J.(1996), Regression by Classification. *Brazilian Symposium on Artificial Intelligence*

[19] WEISS , M. , INDURKHYA , N. (1993), Optimized Rule Induction. *IEEE Expert*: *Intelligent Systems and Their Applications* , 8, 61-69

[20] WILSON, P. W. F., MEIGS, J. B., , et al.(2007), Prediction of Incident Diabetes Mellitus in Middle—aged Adults: The Framingham Offspring Study. *Arch Intern Med* , 167, 1068-1074.

[21] WU, B.(1992), An introduction to neural networks and their applications in manufacturing. *Journal of Intelligent Manufacturing* 3,391-403.