کاربرد روش کاهش بعد چندعاملی (MDR) در شناسایی مدل‌های چند لوکوسی دخیل در رخداد بیماری بهجت

انوشیروان کاظم‌نژاد [۱]، پریسا ریاحی [۲]، شایان مصطفایی [۳]

تاریخ دریافت: ۱۳۹۹/۱۰/۱۸
تاریخ پذیرش: ۱۴۰۰/۰۸/۰۸

چکیده:
رگرسیون لجستیک به دلیل تنکی و وجود متغیرهای رگرسیونی جداکننده‌های کامل و حجم بالایی محاسبات در اثرات متقابل مرتبه بالا، دقت کافی برای تشخیص اثرات اصلی و متقابل بین جهش‌های ژنتیکی را در مراتب خیلی بالا ندارد. الگوریتم کاهش بعد چندعاملی به‌عنوان یک الگوریتم توانمند برای شناسایی اثرات متقابل مراتب بالا در ساختارهای ابربعد محسوب می‌شود. در این تحقیق با استفاده از اطلاعات ۷۴۸ مورد بیمار مبتلا به بیماری بهجت که به مرکز تحقیقات روماتولوژی، بیمارستان شریعتی تهران مراجعه کرده بودند و ۷۷۶ شاهد سالم، برای شناسایی اثرات متقابل بین پلی‌مورفیسم‌های ژن ERAP1 دخیل در رخداد بیماری بهجت از الگوریتم کاهش بعد چندعاملی استفاده‌شده است. محاسبات با استفاده از نرم‌افزار mdr 3.0.2 انجام‌گرفته است. مدل‌های حاصل از الگوریتم کاهش بعد چندعاملی با دقت متعادل بالای ۰٫۶ دخیل در افزایش ریسک بیماری بهجت تعیین‌شده‌اند. الگوریتم کاهش بعد چندعاملی توان و سرعت بالایی در محاسبه اثرات متقابل پلی‌مورفیسم‌ها یا جهش‌های ژنتیکی و شناسایی اثرات متقابل مهم و معنی‌دار دارد.

واژه‌های کلیدی: الگوریتم کاهش بعد چندعاملی، بیماری بهجت، اثرات متقابل ژن-ژن

۱ مقدمه

محاسبه اثرات متقابل ژن-ژن و ژن با محیط در بیماری‌های پیچیده چندعاملی و شناسایی اثرات متقابل مهم و معنی‌دار، یکی از مهم‌ترین و چالش‌برانگیزترین زمینه‌های مطالعاتی دانشمندان ژنتیک و آمار است. در محاسبه اثرات متقابل در مراتب بالاتر به‌ویژه در ساختارهای ابربعد، استفاده از روش مرسوم کلاسیک رگرسیون به دلیل یک سری محدودیت‌ها و بار محاسباتی نسبتاً بالا، امکان‌پذیر نیست. رگرسیون لجستیک به دلیل تنکی [۴] و وجود متغیرهای رگرسیونی جداکننده‌های [۵] کامل، دقت کافی برای تشخیص اثرات اصلی و متقابل بین ژن‌ها را در مراتب بالاتر ندارد، چراکه برآورد پارامترها نامعتبر است. در حضور متغیرهای جداکننده نیز درصورتی‌که متغیر توضیحی بزرگ‌تر از یک مقدار T باشد، یک رابطه بین یک متغیر پیشگو و یکی از متغیرهای توضیحی می‌شود. کاربرد الگوریتم کاهش بعد چندعاملی برای شناسایی اثرات متقابل بین ژن‌ها یا اپیستاسیس [۶] در مورد هر فنوتیپی از پیشامد موردبررسی است. به‌طورکلی تحلیل‌های محاسباتی برای اپیستاسیس مشکل بوده و دارای محدودیت‌های است زیرا این ساختار بسیار پیچیده است. روش کاهش بعد چندعاملی اولین بار توسط ریچی و همکارانش برای به دست آوردن اثرات متقابل بین ژن‌های دخیل بیماری سرطان پستان غیرارثی معرفی‌شده است [۱]. اولین مزیت کاهش بعد چندعاملی سرعت عمل آن در تشخیص و شناسایی هم‌زمان چند لوکوس است. مزیت دیگر آن ناپارامتری بودن آن است که این یک تفاوت اصلی در مقایسه با روش‌های پارامتری آماری قدیمی مثل مدل‌های خطی تعمیم‌یافته است. مثلاً در رگرسیون لجستیک، با وارد شدن هر اثر اصلی به مدل، تعداد اثرات متقابل نیز به‌صورت نمایی افزایش می‌یابد که این یک مشکل اساسی است. مزیت سوم آن، بدون پذیره بودن این روش است که این در بیمارهایی مثل سرطان سینه که در آن مد وراثت، نامعلوم و تقریباً پیچیده است، مهم است. مزیت چهارم، حداقل شدن مقدار مثبت کاذب در آزمون‌های چندگانه است با استفاده از اعتبارسنجی متقاطع [۷] در انتخاب بهترین مدل با اثرات متقابل است.

[۱] استاد آمار زیستی، گروه آمار زیستی، دانشکده علوم پزشکی، دانشگاه تربیت مدرس، تهران، ایران، (نویسنده مسئول: kazem_an@modares.ac.ir)
[۲] دانش‌آموخته رشته آمار زیستی، گروه آمار زیستی، دانشکده علوم پزشکی، دانشگاه تربیت مدرس، تهران، ایران
[۳] استادیار آمار زیستی، گروه آمار زیستی، دانشکده بهداشت، دانشگاه علوم پزشکی کرمانشاه، کرمانشاه، ایران

[4] Sparsity
[5] Separation
[6] Epistasis
[7] Cross-validation, CV

بیماری بهجت یک بیماری التهابی مزمن سیستمیک است در این بیماری درگیری‌های چشمی که مکرراً اتفاق می‌افتد و کاهش بینایی را در پی دارد، زخم‌های دهانی، زخم‌های تناسلی و اختلالات پوستی شایع‌ترین تظاهرات بالینی هستند. این بیماری در کل دنیا وجود دارد، ولی عمدتاً در کشورهایی که در امتداد جاده ابریشم، از ژاپن تا خاورمیانه و حوضه مدیترانه شایع است. اگرچه علت بیماری هنوز شناخته‌شده نیست اما همانند بسیاری از بیماری‌های سندروم خودایمنی و خودالتهابی، عوامل ژنتیکی در این بیماری دخیل هستند. بیماری بهجت برای نخستین بار توسط هولوسی بهجت [8] توصیف‌شده است [۲]. علاوه بر اثرات چشم و پوست، افت دهان و ناحیه تناسلی معیار اصلی تشخیص بیماری بهجت هستند. اگرچه بهجت به‌طور خاص در آمریکا (۵/۲ در ۱۰۰۰۰۰) و در اروپا (۲/۴ در ۱۰۰۰۰۰) شایع نیست. با توجه به میزان شیوع بهجت در کشورهای مجاور جاده ابریشم از جمله ایران، که شیوع آن ۸۰ نفر در هر ۱۰۰۰۰۰ است باعث شده است که مدیریت آن، یک مشکل برای متخصصان روماتیسم ایران و آسیا محسوب شود [۳، ؟]. لذا این تحقیق با هدف شناسایی اثرات متقابل پلی‌مورفیسم‌های ژن ERAP1 دخیل در رخداد بیماری بهجت با کاربرد روش کاهش بعد چندعاملی است.

۲ روش کار

این پژوهش یک مطالعه درون رایانه‌ای است در این مطالعه از اطلاعات ۷۴۸ مورد بیمار مبتلا به بیماری بهجت و ۷۷۶ شاهد سالم استفاده‌شده است. اطلاعات افراد مورد از بیماران بهجتی مراجعه‌کننده به مرکز تحقیقات روماتولوژی، بیمارستان شریعتی تهران با روش نمونه‌گیری تصادفی ساده انتخاب‌شده و موردمطالعه قرارگرفته‌اند. تمام تشخیص‌ها با معیارهای بین‌المللی بیماری بهجت (ICBD) تائید شده است. گروه شاهد شامل ۷۷۶ فرد سالم، بدون سابقه خانوادگی و یا علائم بالینی هر نوع اختلالات روماتیسمی یا سایر اختلالات خودایمنی و همسان شده به لحاظ سن و جنسیت و نژاد با گروه مورد بیمار هستند [۵]. همه محاسبات با استفاده از نرم‌افزار mdr 3.0.2 انجام‌گرفته است. میزان خطای نوع اول به‌منظور سطح معنی‌داری آماری ۰/۰۵ در نظر گرفته‌شده است.

۱.۲ آماده‌سازی DNA و ژنوتیپ کردن پلی‌مورفیسم‌ها

پروتکل مطالعه توسط کمیته اخلاقی دانشگاه علوم پزشکی تهران (کمیته اخلاقی ۹۱-۰۴-۴۱-۱۹۳۸۰-۲۹۶۳۷۱) تصویب‌شده است. نمونه‌های خون استخراج‌شده از آزمودنی‌ها با استفاده از پروتکل‌ها و تکنیک‌های استاندارد، ژنومیک DNA از آن‌ها استخراج‌شده است و سپس با استفاده از MGB-TaqMan Allelic Discrimination method [9] و با استفاده از StepOnePlus Real-Time PCR System ژنوتایپ‌ها و آلل‌های پلی‌مورفیسم‌های این ژن به دست می‌آیند. همه پلی‌مورفیسم‌ها از کنترل کیفیت آماری برخوردار هستند [۵].

۲.۲ روش کاهش بعد چندعاملی

روش کاهش بعد چندعاملی ابتدا همه ترکیبات دوعاملی را در نظر گرفته و سپس بهترین مدل دوعاملی (مدل با اثرات متقابل مرتبه دو) را با کمترین خطای پیش‌بینی در بین همه مدل‌های دوعاملی انتخاب می‌کند. سپس این فرایند در بین همه مرتبه‌های بالاتر ترکیبات عاملی یا اثرات متقابل تکرار می‌شود و در هر مرحله "بهترین" مدل انتخاب می‌شود. خطای طبقه‌بندی بر روی مجموعه آموزش و خطای پیش‌بینی بر روی مجموعه آزمون محاسبه می‌شود. در چنین مدل‌های به دلیل پیچیدگی و وجود اثرات مقابل زیاد، پدیده بیش برازشی بسیار رایج است. اعتبارسنجی متقاطع در مدل بندی تاحدودی قادر است مانع از رخداد این پدیده جلوگیری نماید. در اعتبارسنجی متقاطع مدل‌ها به دنبال مدلی هستیم که در مقایسه با دیگر مدل‌ها، سازگاری [10] اعتبارسنجی متقاطع آن بیشتر و خطای پیش‌بینی برای آن مدل کمتر باشد درصورتی‌که چند مدل با چنین شرایطی وجود داشته باشد آنگاه برای انتخاب بهترین مدل، از مدل صرفه‌جو [11] استفاده می‌شود. بااین‌وجود هرچه تعداد لوسی‌ها افزایش پیدا می‌کند، خطای طبقه‌بندی همواره کاهش می‌یابد و این به دلیل بیش برازش مشاهده‌شده در خطای طبقه‌بندی است. در انتخاب مدل ضروری است تا خطای پیش‌بینی در نظر گرفته شود گرفته شود.

اعتبارسنجی متقاطع، مهم‌ترین مرحله در الگوریتم‌های یادگیری ماشین نظیر کاهش بعد چندعاملی است. چراکه توان پیش‌بینی یک مدل را ارزیابی کرده و از بیش برازشی تاحدودی جلوگیری می‌کند. بررسی‌های صورت گرفته نشان داده است که کاهش بازه‌های تعداد اعتبارسنجی متقاطع از ۱۰ به ۵، توان آزمون را کاهش نمی‌دهد و

[8] Hulusi Behcet
[9] Applied Biosystems, Foster City, CA, USA
[10] Consistency
[11] Statistical parsimony

زمان محاسباتی را نصف کرده و برای تعداد مشاهدات کم پیشنهاد می‌شود [۶].

روش اجرای الگوریتم کاهش بعد چندعاملی در شکل ۱ نشان داده‌شده است که در آن هر سلول چندعاملی در فضای n بعدی به‌عنوان «ریسک بالا» یا «ریسک پایین» نام‌گذاری می‌شود و خطای پیش‌بینی هر مدل برآورد شده است.

مراحل اجرای کاهش بعد چندعاملی:

۱. در گام اول، یک مجموعه n تایی ورینت‌های ژنتیکی گسسته از ترکیب تمام عوامل انتخاب می‌کنیم.

۲. در گام دوم، تمام n عامل و کلاس‌های چندعاملی یا سلول‌ها در فضای n بعدی نمایش داده می‌شوند. سپس نسبت تعداد موارد به تعداد شاهدها در درون هر کلاس چندعاملی برآورد می‌شود.

۳. هر سلول در فضای n بعدی به دو دسته "ریسک بالا"، اگر نسبت موردها به شاهدها برابر باشد یا بیشتر از مقدار آستانه تعیین‌شده (به‌عنوان‌مثال ۱/۰) باشد، یا "ریسک پایین" اگر نسبت محاسبه‌شده از مقدار آستانه بزرگ‌تر نباشد منتسب می‌شود. بدین طریق، یک مدل برای موردها و شاهدها با ادغام خانه‌های با ریسک بالا به یک گروه و خانه‌های با ریسک پایین به یک گروه دیگر ایجاد می‌شود. بدین طریق ما مدل n بعدی را به مدل تک‌بعدی تبدیل می‌کنیم (یعنی اینکه یک متغیر با دو کلاس چندعاملی داریم ریسک بالا، ریسک پایین).

۴. در مرحله چهارم، خطای پیش‌بینی‌شده هر مدل به‌وسیله اعتبارسنجی متقابل ۱۰-تایی برآورد می‌شود. به این صورت که داده‌ها به‌صورت تصادفی به ۱۰ مجموعه مساوی تقسیم می‌کنیم. مدل کاهش بعد چندعاملی را برای هر ۹/۱۰ داده‌ها اجرا می‌کنیم و سپس به‌منظور پیش‌بینی وضعیت بیماری (متغیر پاسخ) ۱/۱۰ داده‌ها از مطالعه کنار گذاشته می‌شوند. نسبت داده‌هایی که موجب پیش‌بینی نادرست شده‌اند، برآورد خطای پیش‌بینی است. برای کاهش احتمال برآوردهای ضعیف خطای پیش‌بینی که به علت تقسیم تصادفی مجموعه داده‌ها است، اعتبارسنجی متقابل ۱۰- برابری، را ۱۰ مرتبه تکرار می‌کنیم و سپس متوسط مقدار خطاهای پیش‌بینی را محاسبه می‌کنیم.

روش کاهش بعد چندعاملی در ابتدا همه ترکیبات چندعاملی را در نظر گرفته و «بهترین» مدل چندعاملی را با کمترین خطای پیش‌بینی در بین همه مدل‌های هر مرتبه انتخاب می‌کند. سپس این فرایند در بین همه مرتبه‌های بالاتر ترکیبات عاملی تکرار می‌شود و در هر مرحله «بهترین» مدل انتخاب می‌شود. در بین همه «مدل‌های نهایی» مدلی که خطای پیش‌بینی را حداقل و سازگاری اعتبارسنجی متقابل را حداکثر می‌کند، انتخاب می‌شود. اگر چندین مدل، خطای پیش‌بینی و سازگاری اعتبارسنجی متقابل (تعداد دفعاتی که مدل به‌دست‌آمده از بین ۹/۱۰ داده‌ها یکسان باشند) یکسانی داشته باشد، مدل با کوچک‌ترین تعداد پلی‌مورفیسم‌ها انتخاب می‌شود [۷].

۳ نتایج

اطلاعات دموگرافیک شرکت‌کنندگان میانگین سنی بیماران بهجتی ۱۰/۸۸ ± ۴۰/۲۶ و در گروه کنترل سالم ۱۱/۵۴ ± ۳۸/۸۸ سال بود (با p-مقدار برابر با ۰/۰۷۶). از بین ۷۴۸ بیمار بهجتی، ۴۴۸ نفر (۵۹/۸٪) و از ۷۷۶ نفر فرد کنترل سالم، ۴۷۶ نفر (۶۱/۳٪) مرد بودند. همچنین جدول ۱ فراوانی الل-های ژن ERAP1 را نشان می‌دهد.

با توجه به جدول ۲، پلی‌مورفیسم rs27044 به‌عنوان اثر اصلی با بیشترین مقدار دقت متعادل [12] ۰/۵۴۳۲ در مقایسه با ۱۰ پلی‌مورفیسم دیگر شناسایی شد. این پلی‌مورفیسم دارای بیشترین سازگاری اعتبارسنجی را در بین مدل‌های تک لوکوسی داشته است (۷/۱۰). مدل دولوکسی به‌دست‌آمده از این روش با دقت متعادل ۰/۵۶۴۱ و همچنین سازگاری اعتبارسنجی متقاطع ۶/۱۰ به‌عنوان بهترین مدل دولوکوسی گزارش می‌شود. مدل سه لوکوسی گزارش‌شده شامل پلی‌مورفیسم‌های rs30187، rs469876 و rs13167972 می‌باشد. که دقت متعادل برای این مدل ۰/۵۸۴۴ است. مدل چهار لوکوسی rs13167972×rs469876×rs28096×rs27434 با دقت متعادل ۰/۶۰۷۸ به‌عنوان بهترین مدل در مرتبه چهار گزارش می‌شود. در اثرات متقابل مرتبه پنج پلی‌مورفیسم‌های rs27044، rs26653، rs469876، rs28096 و rs13167972 با دقت متعادل ۰/۶۲۳ و میزان از سازگاری اعتبارسنجی متقاطع ۳/۱۰ به‌عنوان بهترین مدل پنج لوکوسی از بین مدل‌های دیگر انتخاب‌شده‌اند.

در مرتبه شش، پلی‌مورفیسم‌های rs106547، rs30187، rs26653، rs469876، rs28096 و rs13167972 با دقت متعادل ۰/۶۴۱۴ و میزان سازگاری اعتبارسنجی متقاطع ۲/۱۰ بهترین مدل شش لوکوسی بین پلی‌مورفیسم‌های ژن ERAP1 را به دست دادند. پلی‌مورفیسم‌های rs1065407، rs30187، rs2287987، rs26653، rs469876، rs28096 و rs13167972 با دقت متعادل 0.6631 و سازگاری اعتبارسنجی متقاطع ۶/۱۰ به‌عنوان بهترین مدل هفت لوکوسی معرفی‌شده است. مدل هشت لوکوسی × rs1065407 × rs2287987 × rs30187 × rs27044 × rs26653 × rs469876 ×

[12] Balanced Accuracy

rs13167972 × rs28096 با دقت متعادل ۰٫۶۷۳۱ و سازگاری اعتبارسنجی متقاطع ۵٫۱۰ و مدل نه لوکوسی × rs1065407 × rs28096 × rs469876 × rs26653 × rs27044 × rs2287987 × rs30187 × rs1748078 × rs13167972 با دقت متعادل ۰٫۶۷۸۹ و دقت متعادل ۴٫۱۰ به‌عنوان بهترین مدل‌ها به‌دست‌آمده‌اند. مدل‌های ده، یازده و دوازده لوکوسی آمده در جدول به ترتیب با دقت متعادل ۰٫۶۸۲۱، ۰٫۶۸۴۸ و ۰٫۶۸۴۸ میزان سازگاری اعتبارسنجی متقاطع ۵٫۱۰، ۱۰٫۱۰ و ۱۰٫۱۰ به‌عنوان بهترین مدل‌ها انتخاب می‌شوند (جدول ۲). ضرایب جینی اندازه و جهت ارتباط بین پلی‌مورفیسم‌ها در قالب یک نمودار دایره‌ای در شکل ۲ نمایش می‌دهد. پلی‌مورفیسم‌های rs17481856، rs27044 و rs2287987 بزرگ‌ترین ضرایب را در بین پلی‌مورفیسم‌های دیگر داشته‌اند.

جدول ۱. فراوانی الل‌های ژن ERAP1

SNP	مکان پلی‌مورفیسم در کروموزوم پنج	الل‌ها	فراوانی الل کوچک(%)		pـ مقدار	نسبت بخت (فاصله اطمینان)
			مورد	شاهد		
rs1065407	۳۷۹،۷۷۶،۹۶	G < T	۳۶٫۶	۳۲٫۵	۰٫۰۱۸	۱٫۲۰(۱٫۰۳, ۱٫۳۹)
rs27044	۱۴۸،۷۸۳،۹۶	G < C	۲۸٫۵	۲۹٫۱	۰٫۷۴	۰٫۹۷(۰٫۸۳, ۱٫۱۴)
rs17482078	۱۶۲،۷۸۳،۹۶	T < C	۱۲٫۶	۱۰٫۳	۰٫۰۵۲	۱٫۲۵(۱٫۰۰, ۱٫۵۶)
rs10050860	۵۰۶،۷۸۶،۹۶	T < C	۱۲٫۵	۱۰٫۱	۰٫۰۳۹	۱٫۲۷(۱٫۰۱, ۱٫۵۹)
rs30187	۶۲۷،۷۸۸،۹۶	T < C	۴۰٫۱	۳۹٫۷	۰٫۸۲	۱٫۰۲(۰٫۸۸, ۱٫۱۸)
rs2287987	۸۳۲،۷۹۳،۹۶	C < T	۱۲٫۵	۱۰٫۲	۰٫۰۴۰	۱٫۲۷(۱٫۰۱, ۱٫۵۹)
rs27895	۸۴۰،۷۹۳،۹۶	T < C	۹٫۸	۹٫۹	۰٫۹۸	۱٫۰۰(۰٫۷۹, ۱٫۲۶)
rs26618	۱۳۳،۷۹۵،۹۶	C < T	۲۰٫۱	۲۲٫۹	۰٫۰۵۹	۰٫۸۵(۰٫۷۱, ۱٫۰۱)
rs26653	۵۴۷،۸۰۳،۹۶	C < G	۴۰٫۲	۳۹٫۷	۰٫۷۵	۱٫۰۲(۰٫۸۹, ۱٫۱۸)
rs3734016	۷۶۱،۸۰۳،۹۶	T < C	۱٫۹	۲٫۴	۰٫۴۰	۰٫۸۱(۰٫۵۰, ۱٫۳۲)
rs72773968	۸۹۲،۸۰۳،۹۶	A < G	۹٫۸	۹٫۹	۰٫۸۸	۰٫۹۸(۰٫۷۷, ۱٫۲۵)

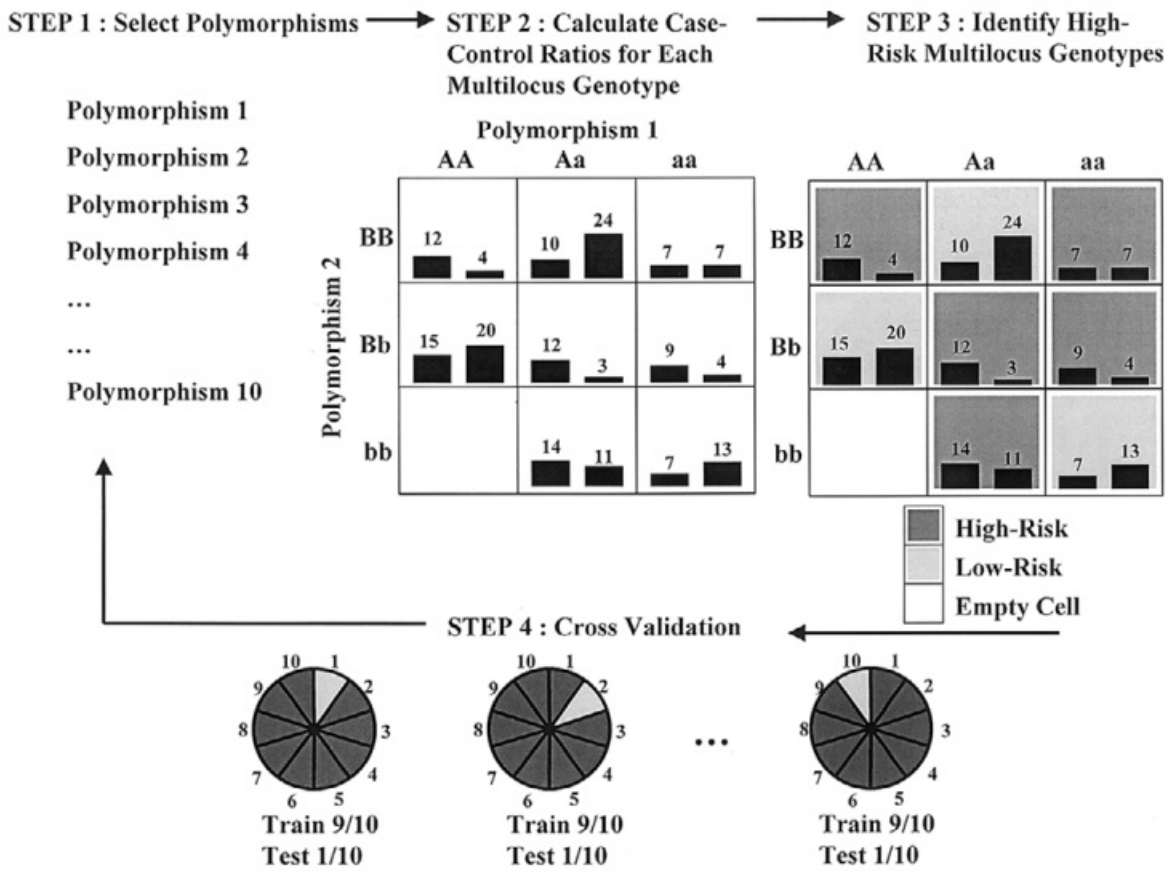
در این نمودار ضرایب اثرات متقابل بین هر جفت پلی‌مورفیسم بر روی خطی که این دو را به هم وصل می‌کند نوشته‌شده است. همچنین علامت ضریب نشان از جهت اثر دارد، به‌طوری‌که ضرایب با علامت منفی دارای اثر کاهنده و مثبت دارای اثر فزاینده‌اند. اندازه اثر اصلی هر پلی‌مورفیسم در کنار اسم آن گزارش‌شده است.رنگ نارنجی نشانگر رابطه فزاینده با درجه کم، رنگ طلایی نشانگر نبود و یا رابطه فزاینده بسیار کم، رنگ سبز نشانگر رابطه کاهنده با درجه کم و رنگ آبی نشانگر رابطه کاهنده با درجه بیشتر از گروه قبلی است.

۴ بحث و نتیجه‌گیری

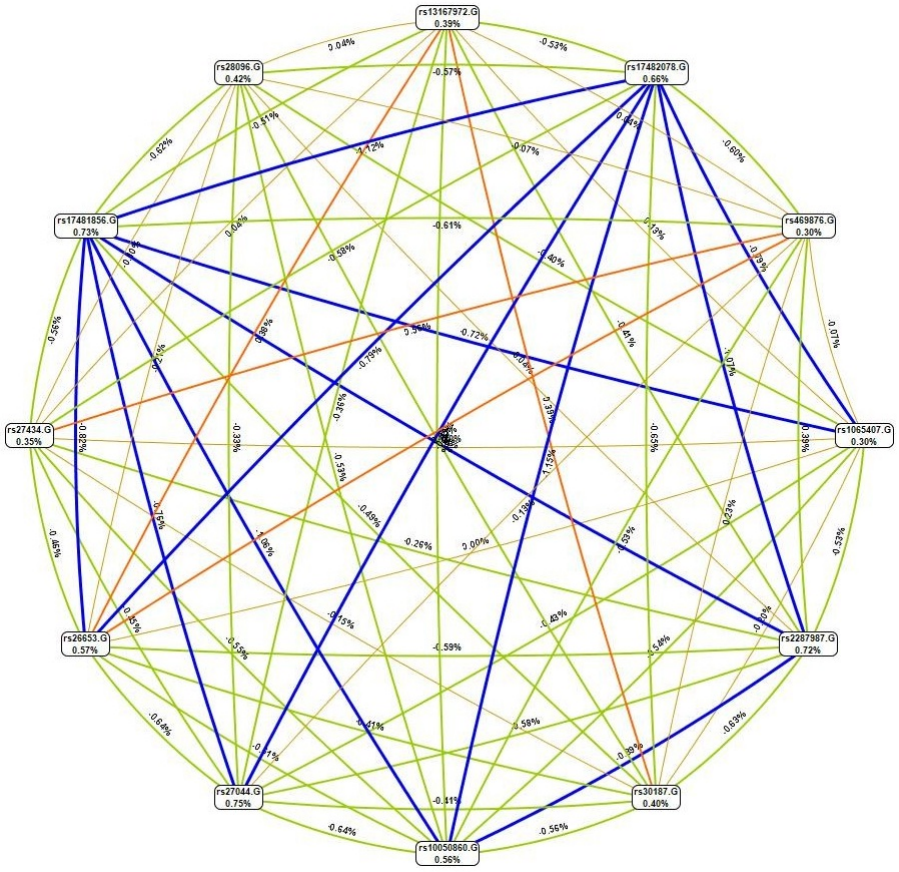
روش کاهش بعد چندعاملی یکی از ابزارهای قوی معرفی‌شده برای محاسبه اپیستاسیس (اثر متقابل ژن‌ـژن) در حوزه علم آمار ژنتیک است. از مزایای روش کاهش بعد چندعاملی می‌توان به‌سرعت عمل آن در تشخیص و شناسایی هم‌زمان چند لوکوس نام برد. محاسبه اثرات متقابل در حضور تعداد پلی‌مورفیسم‌های بالا و همچنین حجم نمونه بالا نیازمند یک الگوریتم توانمند باقدرت تشخیص بالاست که درعین‌حال سرعت پردازش بالایی داشته باشد. مزیت دیگر آن ناپارامتری بودن و بدون مدل بودن آن است که این یک تفاوت اصلی در مقایسه با روش‌های پارامتری آماری قدیمی مثل مدل‌های خطی تعمیم‌یافته است. این مدل علاوه بر مزیت‌های ذکرشده یک مزیت دیگر نیز دارد که این مدل را به‌طور برجسته‌ای از دیگر مدل‌ها متمایز می‌کند، حداقل شدن مقدار مثبت کاذب در آزمون‌های چندگانه است که این به دلیل استراتژی اعتبارسنجی متقابل در انتخاب بهترین مدل است [۱].

جدول۲. نتایج الگوریتم کاهش بعد چندعاملی برای پلی‌مورفیسم‌های ژن ERAP1

مدل	نتایج اعتبار سنجی متقاطع						
	دقت متعادل	حساسیت	ویژگی	CV	دقت متعادل	حساسیت	ویژگی
rs27044	۰/۴۹۴۸	۰/۵۴۸۵	۰/۴۴۱۱	۷/۱	۰/۵۴۳۲	۰/۶۶۲۲	۰/۴۲۴۲
rs27434 × rs469876	۰/۴۸۳۵	۰/۴۰۸۰	۰/۵۵۸۹	۶/۱	۰/۵۶۴۱	۰/۴۷۴۹	۰/۶۵۳۲
rs30187 × rs469876 × rs13167972	۰/۵۰۲۱	۰/۳۸۸۰	۰/۶۱۶۲	۶/۱	۰/۵۸۴۴	۰/۴۴۱۵	۰/۷۲۷۳
rs27434 × rs28096 × rs469876 × rs13167972	۰/۵۲۵۵	۰/۴۳۱۴	۰/۶۱۹۵	۵/۱	۰/۶۰۷۸	۰/۴۷۴۹	۰/۷۴۰۷
rs27044 × rs26653 × rs469876 × rs28096 × rs13167972	۰/۴۷۸۶	۰/۳۶۷۹	۰/۵۸۹۲	۳/۱	۰/۶۲۳۰	۰/۴۷۸۲	۰/۷۶۷۷
rs1065407 × rs30187 × rs26653 × rs469876 × rs28096 × rs13167972	۰/۴۶۳۳	۰/۴۰۱۳	۰/۵۲۵۳	۲/۱	۰/۶۴۱۴	۰/۵۴۱۸	۰/۷۴۰۷
rs1065407 × rs30187 × rs2287987 × rs26653 × rs469876 × rs28096 × rs13167972	۰/۴۸۵۱	۰/۴۲۱۴	۰/۵۴۸۸	۶/۱	۰/۶۶۳۱	۰/۵۶۱۹	۰/۷۶۴۳
rs1065407 × rs2287987 × rs30187 × rs27044 × rs26653 × rs469876 × rs28096 × rs13167972	۰/۴۹۶۸	۰/۴۳۸۱	۰/۵۵۵۶	۵/۱	۰/۶۷۳۱	۰/۵۸۸۶	۰/۷۵۷۶
rs1065407 × rs2287987 × rs27044 × rs26653 × rs469876 × rs28096 × rs13167972 × rs1748078 × rs30187	۰/۵۰۵۲	۰/۴۴۴۸	۰/۵۶۵۷	۴/۱	۰/۶۷۹۸	۰/۶۰۵۴	۰/۷۵۴۲
rs1065407 × rs2287987 × rs30187 × rs27044 × rs26653 × rs27434 × rs469876 × rs28096 × rs13167972 × rs17482087	۰/۵۰۳۵	۰/۴۵۱۵	۰/۵۵۵۶	۵/۱	۰/۶۸۲۱	۰/۶۱۲۰	۰/۷۵۴۲
rs1065407 × rs2287987 × rs30187 × rs10050860 × rs27044 × rs26653 × rs27434 × rs469876 × rs28096 × rs13167972 × rs17482078	۰/۵۰۵۲	۰/۴۶۴۹	۰/۵۴۵۵	۱۰/۱	۰/۶۸۴۸	۰/۶۱۲۰	۰/۷۵۷۶
rs1065407 × rs2287987 × rs30187 × rs10050860 × rs27044 × rs26653 × rs27434 × rs469876 × rs28096 × rs13167972 × rs17482078 × rs17481856	۰/۵۰۵۲	۰/۴۶۴۹	۰/۵۴۵۵	۱۰/۱	۰/۶۸۴۸	۰/۶۱۲۰	۰/۷۵۷۶



شکل ۱. مراحل انجام الگوریتم کاهش بعد چندعاملی



شکل ۲. نمودار دایره‌ای نتایج MDR

در کنار مزیت‌ها و توانمندی‌های مدل کاهش بعد چندعاملی، این روش مشابه سایر روش‌های دیگر از معایب و کاستی‌هایی همراه است. یکی از عمده مشکلات روش کاهش بعد چندعاملی خالی بودن خانه‌های جدول توافقی ساخته‌شده در مدل‌هایی با ابعاد بالاست که در این صورت قادر نیست که این خانه‌ها را به‌عنوان خانه با ریسک بالا (فزاینده) و یا با ریسک پایین (کاهنده) نام‌گذاری کرد که این ممکن است در محاسبه درست و دقیق اثرات متقابل ابعاد بالا دچار مشکل شود. مشکل دیگر آن است که نسبت موردها به شاهدها نزدیک به یک باشد که در این صورت نام‌گذاری خانه‌های جدول کار مشکلی خواهد بود [۸]. از عمده مشکلات این مدل دودویی بودن متغیر پاسخ است و نمی‌توان مقیاس‌های دیگری برای متغیر پاسخ در نظر گرفت. همچنین این مدل قادر نیست تا اثر متغیرهای مخدوش گر و یا هیچ کوریتی را تعدیل کند [۹]. در هنگام استفاده از روش کاهش بعد چندعاملی باید در نظر داشت که اگر تعداد عامل‌ها، بیش از ۱۵ باشد، محاسبات پیچیده‌تر شده و زمان محاسباتی بیشتر می‌شود و توان مدل در حضور داده‌های گمشده، خطای ژنوتیپی، ناهمگونی ژنتیکی و فنوکپی کم است [۱۰].

اگر کاهش بعد چندعاملی در حضور اثرات اصلی یا کوریت‌های مهم اجرا شود، تعیین مدل نهایی سخت می‌شود. به‌عنوان‌مثال، اگر در تحلیل کاهش بعد چندعاملی معلوم شود که مدل نهایی، شامل ۴ پلی‌مورفیسم است، کاملاً مشخص نمی‌شود که آیا مدل نهایی یک مدل با اثر متقابل ۴ طرفه است، یا دارای دو اثر متقابل دوطرفه جداگانه، و یا دو اثر اصلی و یک اثر متقابل دوطرفه است [۷]. به عبارتی در مدل کاهش بعد چندعاملی تعیین اینکه مدل نهایی به‌دست‌آمده فقط دارای یک اثر متقابل است یا این پلی‌مورفیسم‌ها اثرات متقابل با مرتبه کوچک‌تری نیز دارند یا خیر. همچنین تفسیر نتایج در مواقعی که تعداد متغیرها تقریباً زیاد و حجم نمونه کم است سخت بوده و در این صورت توان پیش‌بینی مدل کم است(در حجم نمونه کوچک، نتایج مثبت کاذب و منفی کاذب به دست می‌آید) [۱]. در این مطالعه، مدل کاهش بعد چندعاملی در محاسبه و تشخیص مهم‌ترین اثرات متقابل بین پلی‌مورفیسم‌های ژن ERAP1 مرتبط با رخداد بیماری بهجت از دقت و سرعت خوبی برخوردار بود. این مدل علیرغم کاستی‌ها و معایبش در حضور داده‌های گمشده و ناهمگونی‌های ژنتیکی، از توان بالایی در محاسبه و به دست آوردن مدل‌های چند لوکوسی ژنتیکی دارد.

مراجع

[1] Ritchie, M.D., Hahn, L.W., Roodi, N., Bailey, L.R., Dupont, W.D., Parl, F.F. and Moore, J.H. (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *The American Journal of Human Genetics*, **69(1)**, 138–147.

[2] Mahr, A., Belarbi, L., Wechsler, B., Jeanneret, D., Dhote, R., Fain, O., Lhote, F., Ramanoelina, J., Coste, J., and Guillevin, L. (2008). Population-based prevalence study of Behçet's disease: differences by ethnic origin and low variation by age at immigration. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, **58(12)**, 3951–3959.

[3] Davatchi, F., Shahram, F., Chams-Davatchi, C., Shams, H., Nadji, A., Akhlaghi, M., Faezi, T., Ghodsi, Z., Larimi, R., Ashofteh, F., and Sadeghi Abdollahi, B. (2010). Behcet's disease in Iran: analysis of 6500 cases. *International journal of rheumatic diseases*, **13(4)**, 367–373.

[4] International Team for the Revision of the International Criteria for Behçet's Disease (ITR-ICBD). (2014). The International Criteria for Behçet's Disease (ICBD): a collaborative study of 27 countries on the sensitivity and specificity of the new criteria, *Journal of the european Academy of Dermatology and Venereology*, **28(3)**, 338–347.

[5] Mahmoudi, M., Jamshidi, A.R., Amirzargar, A.A., Farhadi, E., Nourijelyani, K., Fallahi, S., Oraei, M., Noori, S., and Nicknam, M.H. (2012). Association between endoplasmic reticulum aminopeptidase-1 (ERAP-1) and susceptibility to ankylosing spondylitis in Iran, *Iranian Journal of Allergy, Asthma and Immunology*, **11**, 294–300.

[6] Mostafaei, S. and Riahi, P. (2020). Interaction Effects of Plasma Vitamins A, E, D, B9, and B12 and Tobacco Exposure in Urothelial Bladder Cancer: A Multifactor Dimensionality Reduction Analysis. *Nutrition and Cancer*, 1-2.

[7] Coffey, C.S., Hebert, P. R., Ritchie, M. D., Krumholz, H. M., Gaziano, J. M., Ridker, P. M., Brown, N. J., Vaughan, D. E., and Moore, J. H. (2004). An application of conditional logistic regression and multifactor dimensionality reduction for detecting gene-gene interactions on risk of myocardial infarction: the importance of model validation. *BMC bioinformatics*, **5(1)**, 49.

[8] Park, M.Y. and Hastie, T. (2008). Penalized logistic regression for detecting gene interactions. *Biostatistics*, **9(1)**, 30–50.

[9] Riahi, P., Kazemnejad, A., Mostafaei, S., Meguro, A., Mizuki, N., Ashraf-Ganjouei, A., Javinani, A., Faezi, S.T., Shahram, F., and Mahmoudi, M. (2020). ERAP1 polymorphisms interactions and their association with Behçet's disease susceptibly: Application of Model-Based Multifactor Dimension Reduction Algorithm (MB-MDR), *PloS one*, **15(2)**, e0227997.

[10] Calle, M.L., Urrea Gales, V., Malats i Riera, N., Van Steen, K. (2008) MB-MDR: model-based multifactor dimensionality reduction for detecting interactions in high-dimensional genomic data. *Technical Report 24*, Department of Systems Biology, Universitat de Vic, Vic,: Spain; available at http://repositori.uvic.cat/xmlui/handle/10854/408.

Application of multifactor dimensionality reduction (MDR) algorithm in detecting the n-locus models related to Behcet's disease

A. Kazemnejad [1] , P. Riyahi [2] , S. Mostafaei [3]

Abstract:

Due to the sparsity and separation and a large amount of calculations in high-order interactions, Logistic regression is not accurate enough to detect the main and interaction effects between genetic markers at very high orders. The multifactorial dimension reduction algorithm is considered a powerful algorithm for identifying high-order interactions in high dimensional structures. In this study, information of 748 patients with Behcet's disease who were referred to the Rheumatology Research Center, Shariati Hospital, Tehran, and 776 healthy controls were used to identify the interaction effects between ERAP1 gene polymorphisms involved in the occurrence of Behcet's disease using the multifactor dimensionality reduction algorithm. Data analysis was performed using MDR 3.0.2 software. The models obtained from the multifactorial dimensional reduction algorithm with balanced accuracy above 0.6 have been determined to increase the risk of Behcet's disease. The multi-factor reduction algorithm has high power and speed in calculating the interaction effects of polymorphisms or genetic mutations and identifying important interactions.

Keywords: Multifactor dimensionality reduction algorithm, Behcet's disease, Gene-gen interaction.

[1]Professor of Biostatistics, department of Biostatistics, faculty of Medical Sciences, Tarbiat modares University, Tehran, Iran. Corresponding author, kazem_an@modares.ac.ir

[2]Graduated student of Biostatistics, department of Biostatistics, faculty of Medical Sciences, Tarbiat modares University, Tehran, Iran.

[3]Assistant professor of Biostatistics, department of Biostatistics, faculty of Health, Kermanshah University of Medical Sciences, Krmanshah, Iran.