مجله دانشگاه علوم پزشکی خراسان شمالی
۱۳۹۰؛ ۳ (ویژه نامه آمار زیستی و اپیدمیولوژی):۱۰۱-۹۳

Journal of North Khorasan University of Medical Sciences
2011;3(Biostatistics and Epidemiology Supplement):93-101

آنالیز جنگل های تصادفی: یک روش آماری مدرن برای غربالگری در مطالعات با بعد بالا و کاربرد آن در یک مطالعه همبستگی ژنتیکی جمعیت–پایه

سحر نوری[۱]، کرامت نوری جلیانی[*۲]*، کاظم محمد[۳]، محمد حسین نیکنام[۴]، مهدی محمودی[۵]، لاریس آندونیان[۶]، آرش اکابری[۷]*

[۱] دانشجوی دوره دکتری تخصصی آمار زیستی، گروه اپیدمیولوژی و آمار زیستی، دانشکده بهداشت و انستیتو تحقیقات بهداشتی، دانشگاه علوم پزشکی تهران، تهران، ایران
[۲] دانشیار آمار زیستی، دانشکده بهداشت و انستیتو تحقیقات بهداشتی، دانشگاه علوم پزشکی تهران ، تهران، ایران
[۳] استاد آمار زیستی، دانشکده بهداشت و انستیتو تحقیقات بهداشتی، دانشگاه علوم پزشکی تهران، تهران، ایران
[۴] دانشیار ایمونولوژی، مرکز تحقیقات ایمونولوژی ملکولی، دانشگاه علوم پزشکی تهران، تهران، ایران
[۵] مرکز تحقیقات روماتولوژی و مرکز تحقیقات ایمونولوژی ملکولی، دانشگاه علوم پزشکی تهران، تهران، ایران
[۶] مربی ژنتیک انسانی، دانشکده بهداشت و انستیتو تحقیقات بهداشتی، دانشگاه علوم پزشکی تهران، تهران، ایران
[۷] مربی آمارزیستی دانشگاه علوم پزشکی خراسان شمالی، بجنورد، ایران
[*] **نویسنده مسئول:** تهران، دانشگاه علوم پزشکی تهران، دانشکده بهداشت، گروه اپیدمیولوژی و آمار زیستی
پست الکترونیک: nourik@tums.ac.ir

چکیده

زمینه و هدف: *پیشرفت های سریع تکنولوژی قرن اخیر در زمینه مطالعات ژنتیکی ما را با حجم زیاد اطلاعات مواجه کرده و چالشی را در تحلیل این قبیل داده های با تعداد بسیار زیاد متغیر پیشگو بوجود آورده است. مطالعه حاضر با در نظر گرفتن داده ها با تعداد متغیرهای بسیار زیاد همراه با اثرات متقابل آنها که ممکن است در تحلیل آماری داده های ژنتیکی با آن مواجه شویم و با هدف بررسی روش های نوین برای تحلیل اینگونه داده های با بعد زیاد انجام پذیرفت.*

مواد و روش کار: *در این مطالعه روش آماری ناپارامتری و نوین جنگل های تصادفی برای تعیین فاکتورهای مهم و اثرگذار ژنتیکی بر روی بیماری آنکیلوزان اسپوندیلیت بکار برده شد. داده ها حاوی اطلاعات مربوط به ژن HLA-B27 و ۱۲ پلی مرفیسم تک نوکلئوتیدی ژنی موسوم به ERAP-1 از ۴۰۱ بیمار مبتلا به آنکیلوزان اسپوندیلیت و ۳۱۶ کنترل سالم بود. تحلیل های فوق متعاقباً به کمک رگرسیون لجستیک نیز اجرا شد و نتایج آن با جنگل های تصادفی مقایسه گردید.*

یافته ها: *بر اساس نتایج مدل رگرسیون لجستیک گام به گام متغیرهای HLA-B27 و پلی مرفیسم rs28096 به طور معنی دار در ارتباط با بیماری مذکور بودند در حالیکه روش جنگل های تصادفی متغیرهای HLA-B27 و rs1065407 را متغیرهای اصلی اثرگذار روی این بیماری تشخیص داد و rs28096 در رتبه سوم اهمیت قرار داشت.*

نتیجه گیری: *نتایج حاصل از این مطالعه حاکی از ارتباط زیاد HLA-B27 با بیماری آنکیلوزان اسپوندیلیت بود. روش کلاسیک و متداول رگرسیون لجستیک پلی مرفیسم rs28096 را مهم ترین فاکتور خطر در رابطه با بیماری معرفی کرد در حالیکه روش جنگل های تصادفی rs1065407 را نیز مهمترین پلی مرفیسم تشخیص داد. لذا محققین بایستی نتایج آماری حاصل از روش های متداول کلاسیک را با روش های جامع و کامل تر نوین از قبیل جنگل های تصادفی در مطالعات غربالگری مدنظر داشته باشند.*

واژه های کلیدی: *جنگل های تصادفی (RF)، داده های بعد بالا، اثرمتقابل، رگرسیون لجستیک، درخت*

مقدمه

پیشرفت های تکنولوژی در قرن اخیر حجم زیادی اطلاعات از سکانس ژنتیکی انسان را فراهم می کند. این امکان دستیابی به حجم زیاد اطلاعات انواع متفاوتی از مطالعات ژنتیکی را ممکن کرده است که در تمام آنها ارتباط بین تغییرات DNA و بیماری هدف مطالعه می‌باشد [۱]. ۹۹/۹٪ از سکانس DNA انسان ها با یکدیگر یکسان است. بیش از ۸۰٪ از ۰/۱٪ اختلاف بین سکانس DNA مربوط به SNP (پلی مرفیسم تک نوکلئوتیدی[1]) هاست. بطور خلاصه یک SNP جانشینی یک تک باز از یک نوکلئوتید با تک باز دیگری است. بر خلاف بیماری های مندلین که در آن ها یک SNP یا یک ژن عامل بیماری می باشد، در بیماری های پیچیده چندین عامل ژنتیکی و محیطی در یک شبکه پیچیده، بطور جمعی و در بسیاری از موارد بطور ضربی، با هم تعامل دارند تا فنوتیپ نهایی بیماری را تعیین کنند [۲]. این تعداد زیاد SNP ها و امکان وجود اثر های متقابل بین عوامل ژنتیکی با هم و با عوامل محیطی یک چالش را در تحلیل اینکه کدامیک از این SNP ها در ارتباط با بیماری هستند پدید آورده است [۳]. هدف اصلی در تحلیل این اطلاعات با بعد بالا[2] ($p >> n$) تعیین اینکه کدام SNP یا مجموعه SNP ها روی بیماری مؤثرند نیست. بسیاری از SNP ها اثرات حاشیه ای ناچیزی دارند در حالیکه اثرات متقابل آنها بسیار قوی است. بلکه هدف اصلی اولویت بندی SNP ها برحسب اهمیت و نقش آنها در ارتباط با بیماری به منظور مطالعه و بررسی بیشتر آنهاست.

نتیجه استفاده از آزمون های تک متغیره در تحلیل این داده های با بعد بالا کنار گذاشتن SNP های با اثرات اصلی کوچک است در حالیکه آن SNP اثر متقابل بزرگی دارد. ممکن است پا فراتر نهاده و علاوه بر آزمون های تک متغیره برای تک تک SNP ها، تمام جفت های ممکن را نیز آزمون کرد اما وقتی که با تعداد زیاد SNP ها مواجه هستیم این روش دشواری است و این سؤال را ایجاد می کند که چرا نباید مجموعه های ۳ تایی، ۴ تایی و حتی بالاتر را آزمون کرد؟

بسیاری از روش های مدل پایه برای بکارگیری بعد بالای داده ها وجود دارد که در آنها باید اثرات متقابل از قبل مشخص شوند[۴,۵,۶] . این روش ها ی مدل سازی محدودیت هایی در تعداد متغیر های پیشگویی که میتوانند یکباره وارد مدل شوند دارند. لذا محقق مجبور است که مدل سازی را در دو مرحله انجام دهد. در مرحله اول تنها اثرات اصلی در نظر گرفته میشود و در مرحله دوم اثرات متقابل بین متغیرهای با اثرات اصلی بزرگ در نظر گرفته میشود. این روش ها نیز منجر به از دست دادن اثرات متقابل مهم با اثرات اصلی کوچک می شوند.

اگرچه مدل رگرسیون لجستیک یکی از روشهای کلاسیک رایج در مطالعات همبستگی است و تا حدودی نیز در مطالعات همبستگی ژنتیکی کارآمد بوده، اما این روش محدودیت هایی دارد.مدل رگرسیون لجستیک یک روش پارامتریک است که فرض میکند متغیرهای پیشگو با لگاریتم شانس ابتلا به بیماری رابطه خطی دارند. در مدل رگرسیون لجستیک اثرات متقابل باید از قبل مشخص شود. از طرفی به ازاء هر متغیر اضافی در مدل تعداد اثر متقابل های ممکن به طور نمایی زیاد می شود. بنابراین برای بعد زیاد داده ها تعداد سلولهای جدول توافقی با تعداد داده های خیلی کم یا صفر افزایش میابد و این منجر به براوردهای با واریانس بالا از پارامترها میشود. از طرفی توان مدل رگرسیون لجستیک در کشف اثرات متقابل مراتب بالا کم خواهد بود. اگر چه اثرات متقابل مراتب بالا اغلب به عنوان توصیفی برای نیکویی برازش ضعیف یا واریانس بیش از اندازه باقیمانده ها استفاده می شوند اما در مطالعات ژنتیکی اندازه گیری مستقیم آنها یا آزمون کردن آنها دیگر چیز بی اهمیتی نیست. اغلب مطالعات طوری طراحی نشده که حجم نمونه آن به اندازه کافی باشد تا بتوان با استفاده از روش رگرسیون لجستیک اثر متقابل های مراتب بالا را کشف کرد. حتی اگر حجم نمونه به اندازه ای باشد که بتوان این اثرات متقابل را کشف کرد تفسیر آنها کاری دشوار است.برای اجتناب از برآورد پارامترها با واریانس زیاد و توان کم در تعیین اثرات متقابل در حجم های نمونه نسبتاً کم روش های آماری ناپارامتری که در آنها نیازی به برآورد پارامترها نیست بسیار مورد توجه می باشند.

1 -Single-nucleotide polymorphism
2 -High-dimensional data

در تحلیل داده های ژنتیکی مسئله دیگری که باید مدنظر قرار گیرد ناهمگنی ژنتیکی[1] است. ناهمگنی ژنتیکی به این معناست که راه های ممکن متعددی برای ایجاد یک بیماری وجود دارد که هر کدام شامل زیر مجموعه های مختلفی از ژنهاست. مدل های رگرسیونی توانایی محدودی در برخورد با ناهمگنی ژنتیکی دارند [۷]. وقتی نتوان نمونه را به زیر گروه هایی که همگنی ژنتیکی دارند تقسیم کرد بعید است روش رگرسیون لجستیک، که تمام افراد را یک گروه در نظر می گیرد و میانگین اثرات را در کل نمونه برآورد می کند، در تعیین SNP های مؤثر در یک بیماری پیچیده موفق باشد.

روشهای درخت–پایه[2] روشهای آماری ناپارامتری (مدل آزاد) برای اجرای انالیز کلاس بندی و آنالیز رگرسیونی با استفاده از الگوریتم افراز های بازگشتی [۹،۸] می باشند. این روش ها در انتخاب یک مجموعه از متغیرهای پیشگو، که به بهترین نحو فنوتیپ نهایی بیماری را بیان کنند، بسیار کارا هستند. روش های درخت-پایه وقتی که متغیرهای پیشگو بطور غیر خطی با در ارتباط با بیماری هستند نیز مفیدند چون هیچ قیدی را در مورد فرم رابطه بین متغیرهای پیشگو و پاسخ فرض نمی کنند. این روش ها با اغلب ناهمگنی های ژنتیکی سازگارند بدین صورت که بطور اتوماتیک مدل های جداگانه ای به زیر مجموعه هایی از داده ها، که با افراز زود هنگام در درخت مشخص میشوند، برازانده میشوند. سادگی مدل و قابل تفسیر بودن روش های درخت-پایه، انعطاف پذیری در بکارگیری تعداد زیاد متغیرهای پیشگو و حجم نمونه محدود و تواناییشان در مدنظر قرار دادن ناهمگنی ژنتیکی منجر به افزایش کاربرد آنها در مطالعات همبستگی ژنتیکی شده است.

جنگل های تصادفی[3] (RF) یک نوع مدرن از روش های درخت-پایه هستند که شامل انبوهی از درختهای کلاس بندی و رگرسیونی اند [۱۰]. مهمترین ویژگی جنگل های تصادفی عملکرد بالای آنها در اندازه گیری اهمیت متغیرها برای مشخص کردن اینکه هر متغیر چه نقشی در پیش بینی پاسخ دارد.

1-genetic heterogeneity
2- Tree-based
3- Random Forests

در این مقاله ما کاربرد جنگل های تصادفی از درخت های کلاس بندی را در غالب یک مطالعه همبستگی ژنتیکی جمعیت-پایه[4] [۱]. مورد-شاهد بین سکانس ژنتیکی و فنوتیپ بیماری آنکیلوزان اسپوندیلیت[5] نشان داده ایم. جنگل های تصادفی برای گروه بندی افراد با ناهمگنی ژنتیکی نیز به کار میروند. اما انگیزه اصلی ما در این مطالعه بکار بردن جنگل های تصادفی برای کشف اینکه کدام SNP ها پیشگوی مهمتری برای فنوتیپ بیماری آنکیلوزان اسپوندیلیت هستند و لذا می توانند در شانس ابتلا به بیماری مؤثر باشند، بود.

روش کار

جمعیت مورد مطالعه و داده ها:

در این مقاله از داده های یک مطالعه مورد-شاهدی جمعیت-پایه استفاده شد.این داده ها شامل ۴۰۱ بیمار مبتلا به آنکیلوزان اسپوندیلیت که از انجمن رماتیسم ستون فقرات ایران جمع آوری شده بودند، و ۳۱۶ کنترل سالم از افراد با نژاد ایرانی و قومیت های مختلف از کارمندان و دانشجویان دانشگاه علوم پزشکی تهران بود که سابقه هیچگونه بیماری های خودایمن[6] در آنها یا در اقوام درجه اولشان وجود نداشت. از هر دو گروه بیمار و کنترل سالم رضایت نامه مورد تأیید کمیته اخلاق پزشکی دانشگاه علوم پزشکی تهران تهیه شده بود. در این مطالعه ۱۳ متغیر پیشگو وجود داشت که شامل اطلاعات ژن HLA-B27 و ۱۲ پلی‌مرفیسم تک نوکلئوتیدی روی ژن ERAP-1 بود. متغیر وابسته نیز وضعیت فرد از لحاظ بیمار یا سالم بودن بود. به بیان آماری هر پلی‌مرفیسم تک نوکلئوتیدی سه حالت ممکن: (i) دو آلل آن هموزیگوت و از نوع وحشی[7] اند. (ii) دو آلل آن هتروزیگوت هستند. (iii) دو آلل آن هموزیگوت و از نوع تغییر یافته[8] اند(۱)؛ و و HLA-B27 دو حالت مثبت و منفی می تواند داشته باشد.

تحلیل آماری؛ رگرسیون لجستیک: در مدل رگرسیون لجستیک ترکیب خطی از متغیرهای پیشگو با میانگین

4- Population-based Genetic Association Study
5 -Ankylosing Spondylitis
6- Auto-immune
7- Wild type
8- variant

یک متغیر پاسخ با توزیع دو جمله ای تحت تابع اتصال لوجیت[1] ارتباط خطی دارد.

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \sum \beta_i X_i$$

که در ان p احتمال وقوع پاسخ مورد نظر، β_0 جمله ثابت، β_i ها ضرایب رگرسیونی و X_i ها متغیرهای مستقل هستند. یکی از رایجترین روش ها برای غربالگری متغیرها در رگرسیون لجستیک روش گام به گام[2] است (۱۵). در این روش ابتدا اثر اصلی هر متغیر آزمون می شود و متغیرهایی که اثر معنی داری روی پاسخ دارند وارد مدل چند متغیره شده و به روش حذف پس رو[3] از مدل کنار گذاشته می شوند. سپس جملات اثر متقابل برای متغیرهای با اثرات اصلی معنادار وارد می شوند. در این مقاله ما روش گام به گام انتخاب متغیر را بر اساس شاخص اطلاعات آکائیک[4] (AIC) به کار بردیم. سطح معنی داری برای اثرهای اصلی در آزمون های تک متغیره ۰/۲ و در سایر مراحل ۰/۰۵ در نظر گرفته شد.

جنگل های تصادفی (RF)

یک RF مجموعه ای از درخت های هرس نشده است که هر درخت با الگوریتم افرازهای بازگشتی[5] [۹,۸] بدست می آید. الگوریتم ساخت یک RF با T درخت از یک مجموعه داده با n مشاهده و P متغیر بدین صورت است:

(i) با روش بوت استرپ یک نمونه تصادفی با جایگذاری به حجم n از مشاهدات انتخاب می شود. (ii) برای نمونه بوت استرپ انتخاب شده یک درخت کلاس بندی با استفاده از الگوریتم افرازهای بازگشتی، رشد می کند. در هر گره افراز بر اساس یک نمونه تصادفی m تایی از P متغیر پیشگو انجام می شود. (iii) الگوریتم افرازهای بازگشتی آنقدر ادامه میابد تا درخت به بزرگترین اندازه خود (یعنی برای هر مشاهده یک گره نهایی)، برسد بدون آنکه درخت هرس شود. (iv) مراحل (i) تا (iii) T بار تکرار می شوند تا یک RF ساخته شود[۱۰]. انتخاب های رایج برای T، ۱۰۰۰ درخت و برای m، $\sqrt{P}$ و $\log(P)$ هستند [۱۲].

یک RF آنقدر بزرگ است که تفسیر آن کار بسیار دشواری است، لذا نیازمند خلاصه کردن اطلاعات آن با استفاده از شاخص های کمی هستیم. یکی از این شاخص ها اهمیت متغیر[6] (VI) است. VI شاخصی برای رتبه بندی متغیرها بر حسب اهمیت آنها در اثر گذاری روی پاسخ است. معروفترین شاخص های VI، شاخص اهمیت جینی[7] و شاخص اهمیت جایگشتی[8] می باشد.

شاخص اهمیت جینی: در طی ساخت درخت های RF برای تعیین اینکه گره بر اساس کدام متغیر افراز شود، از شاخص ناخالصی جینی [۹] استفاده می شود. اهمیت متغیر X_i در یک درخت مجموع کاهش در شاخص ناخالصی جینی روی تمام گره هایی است که بر اساس X_i افراز شده اند. میانگین اندازه اهمیت متغیر X_i روی تمام درخت های جنگل، اندازه شاخص اهمیت جینی است.

شاخص اهمیت جایگشتی: برای محاسبه این شاخص، الگوریتم RF از تمام مشاهدات نمونه برای ساخت درخت استفاده نمی کند بلکه یک نمونه تصادفی با جایگذاری به حجم n_1 (معمولاً برابر $2n/3$)از مشاهدات انتخاب می شود. به مشاهدات انتخاب شده نمونه آموزشی[9] (LS) و به بقیه آنها نمونه خارج کیسه[10] (OOB) گفته میشود. درخت ها با مشاهدات LS ساخته می شوند و از OOB برای اندازه گیری ناخالصی درخت استفاده می شود. در هر درخت ابتدا اندازه ناخالصی روی مشاهدات OOB محاسبه می شود. سپس مقادیر متغیر X_i مشاهدات OOB به طور تصادفی جابجا[11] می شوند و اندازه ناخالصی درخت روی مقادیر جابجا شده محاسبه می شود. اندازه اهمیت متغیر

1- Logit link function
2- Stepwise
3- backward elimination
4 -Akaike information criterion
5- Recursive partitioning

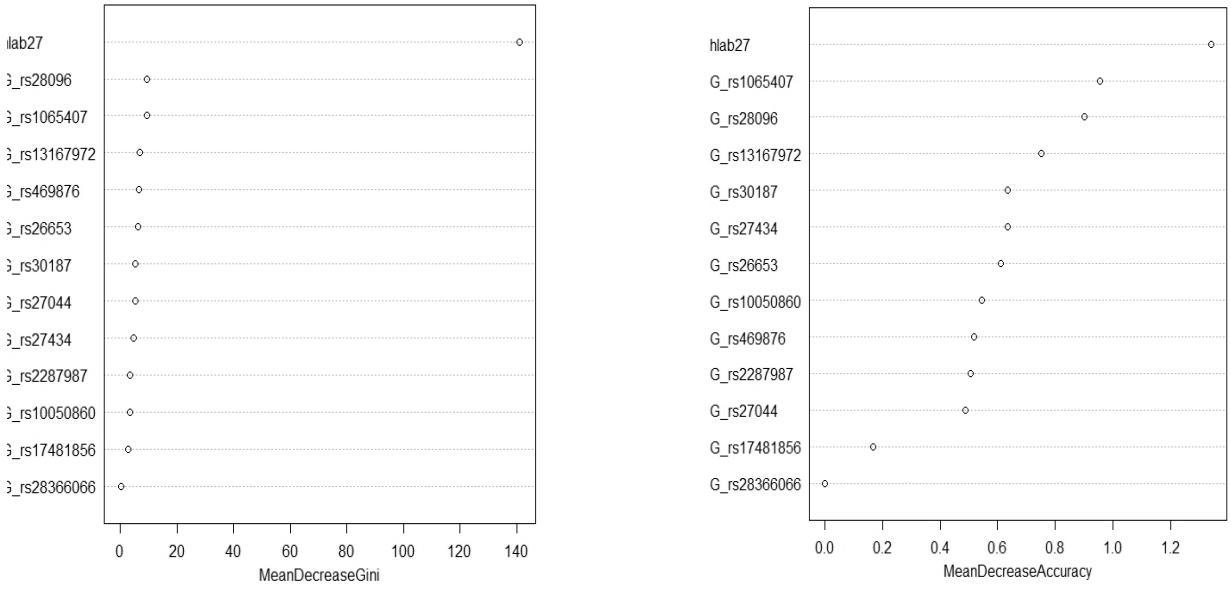
6 -Variable importance
7- Gini importance index
8 -Permutation importance index
9- Learning sample
10- Out-of-bag
11 -permute

X_i در هر درخت، اختلاف بین این دو اندازه ناخالصی است و میانگین این مقادیر شاخص اهمیت جایگشتی است. انگیزه این روش اینست که اگر X_i متغیر مهمی باشد جابجا شدن مقادیر آن بطور تصادفی منجر به افزایش ناخالصی درخت می شود در حالیکه اگر متغیر تأثیر گذاری نباشد، تغییری در ناخالصی ایجاد نمی شود. کلیه تحلیل ها با استفاده از نرم افزار R [۱۳] انجام شد. آنالیز RF با استفاده از تابع random Forest در پکیج random Forest و رگرسیون لجستیک با استفاده از تابع stepAIC در پکیج MASS اجرا شدند.

یافته ها

جدول ۱ نتایج حاصل از رگرسیون تک متغیره را نشان می دهد. از آنجا که روش رگرسیون لجستیک گام به گام از مقادیر احتمال و شاخص AIC برای غربالگری متغیرها استفاده می کند از ذکر جزئیات نتایج رگرسیون لجستیک امتناع کرده و تنها به ذکر مقادیر احتمال اکتفا کرده ایم. ضمناً چون هر SNP یک متغیر سه حالتی است که ژنوتیپ هموزیگوت نوع وحشی آن به عنوان گروه مرجع در نظر گرفته شده مقدار احتمال گزارش شده برای آن حاصل از آزمون حداکثر درستنمایی[1] است و معنی داری کلی متغیر را نشان می دهد. متغیرهای HLA-B27، rs30187، rs1065407، rs28096، rs27434، rs2287987، rs10050860، rs26653، rs27044 به ترتیب معنی دار ترین متغیرها در سطح ۰/۲ در ارتباط با AS بودند.

مدل نهایی رگرسیون لجستیک گام به گام بر اساس شاخص AIC در جدول ۳ نشان داده شده است. این مدل تنها شامل اثرات اصلی متغیرهای HLA-B27 و rs 28096 بود و هیچ اثر متقابلی معنی دار نبود.



شکل ۱ : تعیین اهمیت متغیرها بر اساس دو شاخص اهمیت جینی و بازگشتی حاصل از اجرای RF

1- Likelihood Ratio Test

جدول ۱: نتایج رگرسیون تک متغیره برای ژن HLA-B27 و پلی مرفیسم های ژن ERAP-1.

متغیر	مقدار احتمال
HLA-B27	$۲/۴ \times ۱۰^{-۱۶}$
rs1065407	۰/۰۰۰۸۶
rs2287987	۰/۰۱۵
rs30187	۰/۰۰۰۱۷
rs10050860	۰/۰۱۹
rs27044	۰/۰۲۶
rs26653	۰/۰۲۰
rs27434	۰/۰۱۱
rs469876	۰/۲۶
rs17481856	۰/۴۹
rs28366066	۰/۲۷
rs28096	۰/۰۰۷۲
rs13167972	۰/۵۵

جدول ۲: مدل نهایی رگرسیون لجستیک حاصل از روش گام به گام برای غربالگری متغیرها

متغیر	نسبت شانس	مقدار احتمال
HLA-B27		
منفی	گروه مرجع	-
مثبت	۷۶/۷۰	۲×۱۰^{-۱۶}
rs28096		
G/G	گروه مرجع	-
A/G	۰/۶۸	۰/۲۷
A/A	۱/۳۴	۰/۳۸

شکل ۱ نتایج الگوریتم RF را نشان می دهد. اهمیت متغیرها براساس شاخص اهمیت جینی (Mean Decrease Gini)و جایگشتی (Mean Decrease Accuracy) بدست آمد. در هر دو شاخص HLA-B27 مهمترین متغیر تعیین شد. شاخص جینی دومین مهمترین متغیر را rs 28096 و شاخص جایگشتی دومین مهمترین متغیر را rs 1065407 معرفی کرد.

بحث

یافته های این مطالعه نشان می دهد که HLA-B27 در هر دو روش رگرسیون لجستیک و RF مهمترین متغیر در ارتباط با آنکیلوزان اسپوندیلیت بود. براساس الگوریتم RF متغیر rs 1065407 دومین متغیر مهم و تأثیرگذار است در حالیکه در مدل رگرسیون لجستیک نهایی اثر آن معنی دار نبود. نکته قابل توجه اینست که اثر اصلی rs13167972 در مدل رگرسیون تک متغیره معنی دار نبود و مقدار احتمال آن از تمام متغیرها بزرگتر بود (P.value = ۰/۵۵) در حالیکه بر اساس روش RF این متغیر رتبه اهمیت بالاتر از rs30187 (P.value ۰/۰۰۰۱۷ =) را داشت. همانطور که گفته شد اندازه اهمیت متغیر که در RF بدست می آید هم بر اساس اثر تکی آن متغیر است و هم بر اساس اثر آن در تعامل با متغیرهای دیگر می باشد.

امروزه به لحاظ پیشرفت های زیاد تکنولوژی در حصول داده ها، داده های با بعد بالا محصول بسیاری از تحقیقات پزشکی مخصوصاً تحقیقات ژنتیک ملکولی است. داده های مربوط به میکرواری[1] و پلی مرفیسم های ژن ها از این قبیل داده ها محسوب می شوند مخصوصاً با پیشرفت های اخیر در پروژه های ژنوم انسانی مسأله ارتباط کل ژنی[2] [۱]بسیار مطرح است. تجزیه و تحلیل آماری اینگونه داده ها مشکلی قابل توجه پیش روی متخصصین آمار زیستی است. در واقع تحلیل این گونه داده ها به کمک روش های متداول و سنتی آماری چون رگرسیون خطی یا لجستیک امکان پذیر نیست و روش های نوین را میطلبد. عدم توانایی روش های کلاسیک به خاطر محدودیت آن ها در حل مسائلی از قبیل آزمون های چندگانه[3] و توان کم آنها در بررسی اثرات متقابل و همبستگی های درونی[4] این متغیر های پیشگویی کننده است. در حالیکه روش های جدید مثل جنگل های تصادفی با بکارگیری الگوی متفاوت سعی در حل داده ها با بعد بسیار بالا دارد.

تجزیه و تحلیل داده های ژنتیکی در مراحل ابتدایی است بدین معنی که علی رغم پیچیدگی های اینگونه داده ها هنوز آماشناسان بدلیل ناآگاهی نسبت به روش های نوین از روش های کلاسیک استفاده می کنند [۱۴].

جنگل های تصادفی شامل تعداد زیادی درخت است که در کاربرد آن به جای اندازه گیری مقادیراحتمال[5] ساختار اهمیت متغیرها مشخص می شود. روش های نوین متعدد دیگری نیز وجود دارند که می توانند در کاربردهای مشابه با جنگل های تصادفی بکار روند. بطور مثال رگرسیون منطقی[6] که روشی مشابه درخت های کلاس بندی و رگرسیونی دارد با این تفاوت که تفسیر آن پیچیده تر است [۱۵]. روش دیگری بنام MARS[7] اخیراً مورد توجه قرار گرفته که از الگوریتم افرازهای بازگشتی استفاده می‌کند. این روش علاوه بر تعیین مجموعه متغیرهای مهم، ساختار ارتباط آنها را با پاسخ نشان می دهد [۸]. انتخاب هر یک از این روش ها برای یک تجزیه و تحلیل خاص به درک شهودی و علمی ما در مورد ارتباط متغیرهای پیشگو و پاسخ نیاز دارد. در ساخت یک جنگل در عمل چندین سؤال پیش می آید. از جمله اینکه لازم است که جنگل شامل چند درخت باشد؟ دقت یک جنگل به دو عامل بستگی دارد: (۱) قدرت پیشگویی هر درخت به تنهایی و (۲) همبستگی میان درخت ها [۱۰]. لذا سایز درخت را تا حداقل ممکن کم می کنیم تا جایی که بیشترین قدرت پیشگویی و کمترین همبستگی را داشته باشد. سؤال دیگری که مطرح است اینست که با توجه به اینکه درخت های جنگل هرس نمی شوند آیا یک جنگل تصادفی به داده ها بیش برازش[8] نمیشود؟ بریمن با

2- micro array
3- Genome-Wide association

4- multiple testing
4- multicollinearity
5- P.values
6 -logic regression
7 -Mutivariable Adaptive Regression Splines
8 -over fitting

استفاده از قانون قوی اعداد بزرگ نشان داد که هر چه تعداد درخت های یک جنگل افزایش میابد خطای پیشگویی جنگل همگرا به یک مقدار می شود [۱۰].

نتیجه گیری

زمانی که با داده هایی مواجه می شویم که تعداد متغیرهای زیادی دارند بهترین روش تحلیل، تعیین متغیرهای مهمی است که روی پاسخ تأثیرگذارند تا بتوان مطالعات بیشتری روی آنها انجام داد. RF یک روش مدرن ناپارامتری است و برخلاف مدل های کلاسیک همچون رگرسیون لجستیک که تنها بر یک مدل تکیه دارند، با استفاده از صدها و هزاران درخت از اطلاعات بیشتری در داده ها استفاده می کند تا بتوان استنباط بهتری از متغیرها داشت.

تشکر و قدردانی

بدینوسیله از دانشگاه علوم پزشکی تهران و مرکز تحقیقات روماتولوژی که ما را در انجام این طرح پشتیبانی کرده اند سپاس گذاری می نماییم.

References

1.Foulkes AS, Applied Statistical Genetics With R For Population-based Association Studies USA, University of massachusetts School of Public Health Sciences 2009.

2.Cardon LRaBJI, Association Study Designs for Complex Diseases, Nature 2001; 2: p. 91-98.

3.Glazier AM NJAT, Finding genes that underlie complex traits, Science 2002; 298: p. 2345-2349.

4.George EI MR, Variable Selection via Gibbs Sampling, Journal of the American Statistical Association 1993; 88(423): p. 881-889.

5.Oh C YKHQMN, Locating disease genes using Bayesian variable selection with the Haseman-Elston. BMC Genet 2003; 4(Suppl 1): p. S69.

6.Yi N GVAD, Stochastic search variable selection for identifying multiple quantitative trait loci, Genetics 2003; 164: p. 1129-1138.

7.Province MA SWRD, Classification methods for confronting heterogeneity, Adv Genet 2001; 42: p. 273-286.

8.Hastie T TRFJ, The elements of statistical learning : data mining, inference and prediction, In Springer series in statistics New York , Springer 2001; xvi: p. 533.

9.Breiman L , Classification and regression trees CA, Wadsworth International Groups 1984.

10.Braiman L, Random forests, Machine Learn 2001; 45: p. 5-32.

11.Hosmer Dw LS, Applied Logistic Regression NewYork, John Wiley & Sons Inc 2000.

12.Genuer R PJMTC, Random Forest: some methodological insights 2008.

13.R Development Core Team R, A language and environment for statistical computing, R Foundation for Statistical Computing, Austria 2010.

14. David H," et al", Investigating the genetic association between ERAP1 and ankylosing Spondylitis 2009; 18(21): 4204-4212.

15. Schwender, HaIK, Identification of SNP interactions using logic regression 2008; 9(1): p. 187-198.

Random Forests Analysis: A modern statistical method for screening in high-dimensional studies and its application in a population-based genetic association study

Noori S [1], Nourijelyani K [2], Mohammad K[3], Niknam MH[4], Mahmoudi M[5], Andonian L[6], Akaberi A [7]*

[1] PhD student of Biostatistics, Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences
[2]Associate professor of Biostatistics, Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences
[3]Professor of Biostatistics, Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences
[4]Associate professor of Immunology, Molecular Immunology center, Tehran University of Medical Sciences
[5]Rheumatology center and Molecular Immunology center, Tehran University of Medical Sciences
[6]Instructor of Human Genetics, Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences
[7]Instructor of Biostatisticsn, North Khorasan University of Medical Sciences

***Corresponding Author:** Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran
Email: nourik@tums.ac.ir*

Abstract

Background & Objectives: *Technology advances in this century, especially, in molecular generics yields high volume, high dimensional data. This creates many unprecedented challenges for statisticians who are responsible for analysis of such data. Although logistic regression method is quite popular in association analysis in medical researches but it has some serious limitations in handling high dimensional data. In present study, our goal is introduce a modern model-free statistical method called random forest that we believe is able to overcome difficulties of the classical statistical methods in finding association between predictors and a trait.*

Material & Methods: *In this study, the nonparametric random forest technique was employed to determine the important factors associated with ankylosing spondylitis (AS) disease. Genetic materials including information on HLA-B27 status (positive/negative) and 12 polymorphisms of the ERAP-1 gene were collected on 401 patients and 316 healthy controls. The data were analyzed both with the logistic regression method and random forests technique and the results were compared.*

Results: *Based on a stepwise logistic regression, HLA-B27 and rs28096 polymorphism were significantly associated with the disease. However, using the random forests technique, we found that HLA-B27 and rs1065407 were the main factors associated with diseases and in fact rs28096 polymorphism becomes the third in importance ranking.*

Conclusion: *The results from our study indicate some discrepancies between logistic regression and random forest analyses of high-dimensional data such as the genetic data that we are dealing here. Although logistic regression is quite popular, easy to employ, and is a predominant statistical method among researchers, but it has some serious limitations. On the other hand, more modern statistical such random forest enjoy a more methodological sophistication and yield more accurate and reliable results. Therefore, researchers should be aware of such alternatives and should use these alternatives accordingly and as situation arise in screening tests especially in genetic data analyses.*

Key words*: random forests, High-dimensional data, interaction, logistic regression, CART*