دانشگاه علوم کشاورزی و منابع طبیعی ساری
پژوهشهای تولیدات دامی

تنظیم و کاربرد الگوریتم جنگل تصادفی در ارزیابی ژنومی

فرهاد غفوری کسبی[۱]، قدرت رحیمی میانجی[۲]، محمود هنرور[۳] و اردشیر نجاتی جوارمی[۴]

۱- دانشجوی دکتری، دانشگاه علوم کشاورزی و منابع طبیعی ساری، (نویسنده مسوول: farhad_ghy@yahoo.com)
۲- استاد، دانشگاه علوم کشاورزی و منابع طبیعی ساری
۳- استادیار گروه علوم دامی، دانشگاه آزاد اسلامی واحد شهر قدس، کرج
۴- دانشیار، پردیس کشاورزی دانشگاه تهران
تاریخ دریافت: ۹۳/۴/۸ تاریخ پذیرش: ۹۳/۶/۳۰

چکیده

یکی از مباحث مهم در انتخاب ژنومی، استفاده از روشی مناسب برای برآورد اثر نشانگرها و ارزیابی ژنومی است. اخیراً روش‌های یادگیری ماشین[۱] که جزو روش‌های ناپارامتری غیرخطی هستند وارد ارزیابی ژنومی شده‌اند. یکی از این روش‌ها الگوریتم جنگل تصادفی[۲] است که این تحقیق روی نحوه تنظیم این روش متمرکز شده است. پارامترهای مهم در الگوریتم جنگل تصادفی به ترتیب اهمیت، تعداد متغیر انتخاب شده در هر گره درخت[۳]، تعداد درخت[۴] و حداقل اندازه گره‌های پایانی[۵] می‌باشند که بهتر است برای آنها مقدار مناسبی تعیین شود و در اصطلاح مدل برای این پارامترها تنظیم[۶] شود. ژنومی ۵ کروموزومی متشکل از ۱۰۰۰۰ نشانگر تک نوکلئوتیدی دوآللی[۷] هریک به طول یک مورگان شبیه‌سازی شد و در ادامه، کارایی ترکیبات مختلف از تعداد متغیر انتخاب شده در هر گره درخت، تعداد درخت و حداقل اندازه گره‌های پایانی در قالب جمعیت شبیه‌سازی شده مورد آزمون قرار گرفته و بهترین ترکیب بر اساس پارامتر خطای خارج از کیسه[۸] انتخاب و برای تجزیه و تحلیل اطلاعات مورد استفاده قرار گرفت. برای داده‌های شبیه‌سازی شده در این مطالعه، کمترین مقدار خطای خارج از کیسه و هم‌چنین حداکثر صحت پیش‌بینی ارزش‌های اصلاحی ژنومی مربوط به مدلی با تعداد متغیر انتخاب در هر گره درخت برابر ۶۰۰۰، تعداد درخت برابر ۱۰۰۰ و حداقل اندازه گره‌های پایانی برابر ۵ بود. بقیه ترکیبات از این سه پارامتر نه تنها منجر به افزایش صحت پیش‌بینی نشدند بلکه در آن‌هایی که از تعداد بیشتری درخت استفاده شده بود، مدت زمان لازم برای انجام محاسبات نیز افزایش یافت. با توجه به این‌که صحت پیش‌بینی الگوریتم جنگل تصادفی تابعی از تعداد متغیر انتخاب شده در هر گره درخت، تعداد درخت و حداقل اندازه گره‌های پایانی است، لازم است ترکیبات مختلفی از این پارامترها مورد استفاده قرار گیرد و ترکیب بهینه با حداکثر عملکرد پیش‌بینی انتخاب شده و برای ارزیابی ژنومی استفاده شود.

واژه‌های کلیدی: ارزیابی ژنومی، جنگل تصادفی، درخت، نشانگر تک نوکلئوتیدی، ارزش‌های اصلاحی

مقدمه

در روش‌های رایج اصلاح نژاد دام، معمولاً اصلاحگران یا انتخاب را بر مبنای مقادیر فنوتیپی (انعکاسی نه چندان دقیق از ژنوتیپ) انجام می‌دهند و یا این‌که از طریق ثبت شجره و فنوتیپ حیوانات و تجزیه و تحلیل این اطلاعات با استفاده از ویژگی بهترین پیش‌بینی ناریب خطی[۹] در قالب مدل دام، ارزش‌های اصلاحی را پیش بینی کرده و بر اساس آن حیوانات برتر را انتخاب نموده و به آن‌ها اجازه تولید مثل می‌دهند و با انجام این عمل بهبود ژنتیکی را در صفت مورد نظر ایجاد می‌کنند. اگر چه اساس این روش بر پایه استفاده از اطلاعات فنوتیپی بوده و در آن معماری ژنتیکی صفات، تعداد ژن‌ها، نوع اثرات ژن‌ها و جایگاه ژن‌ها در نظر گرفته نمی‌شد اما همین روش در نیمه دوم قرن بیستم بهبود چشم‌گیری در صفات اقتصادی دام‌های اهلی ایجاد کرد (۱۳). به هر حال کاربرد این روش برای صفات با وراثت‌پذیری پایین، صفات محدود به جنس، صفات مربوط به ماندگاری و صفاتی که اندازه‌گیری آنها مشکل می‌باشد مثل صفات مربوط به لاشه با محدودیت‌هایی روبه‌رو بود. با انفجار در تکنولوژی DNA طی دهه‌های اخیر این امکان برای پژوهشگران عرصه اصلاح نژاد فراهم شد تا از فنوتیپ حیوان فاصله گرفته و فرآیند انتخاب را با کمک گرفتن از اطلاعات ژنتیکی تولید شده هرچه بیشتر به سمت ژنوتیپ حیوان سوق دهند. پیشرفت اساسی که زمینه را برای ارائه ایده انتخاب ژنومی فراهم کرد، توالی یابی ژنوم گاو بود که امکان شناسایی هزاران نشانگر DNA به شکل چندشکلی تک نوکلئوتیدی (SNP) یا تنوع تک نوکلئوتیدی[۱۰] (SNV) را فراهم نمود. اگرچه انتخاب ژنومی با نام میوسن و همکاران (۱۶) گره خورده است و این محققین چهارچوب انتخاب ژنومی و اصول ریاضی پیش‌بینی ارزش‌های اصلاحی ژنومی را توسعه دادند، مفهوم پیش‌بینی ژنومی چند سال قبل‌تر از سوی نجاتی جوارمی و همکاران (۱۹) و ویشر و هالی (۲۴) به دنیای

1- Machine Learning
2- Random Forest, RF
3- Mtry
4- Ntree
5- Nodesize
6- Tune
7- Single Nucleotide Polymorphism (SNP)
8- Out Of Bag Error (OOB Error)
9- Best Linear Unbiased Prediction
10- Single Nucleotide Variation

اصلاح نژاد دام وارد شده بود. در دسترس بودن انبوه SNPها که کل ژنوم را پوشش می‌دهند و فن‌آوری‌های سریع تعیین ژنوتیپ به توسعه انتخاب ژنومی کمک شایانی کرده است. تا سال ۲۰۱۲ پنل‌های ۸۰۰۰۰۰ SNP برای گاوهای شیری توسعه داده شد (۹) که انتظار می‌رود این تعداد در آینده نزدیک به سه میلیون SNP برسد که کل SNPهای ژنوم گاو را در بر خواهد داشت (۱۴). یکی از مسائلی که در انتخاب ژنومی مطرح است بحث برآورد اثر نشانگرها است. بر این اساس روش‌های مختلفی برای برآورد اثر نشانگرها توسعه یافته‌اند. مشابه با روش‌های مبتنی بر اطلاعات فنوتیپی مانند مدل انفرادی، مدل پدری، مدل دام تک متغیره و چند متغیره و مدل رگرسیون تصادفی که صحت پیش‌بینی ارزش اصلاحی در آنها متفاوت به نظر می‌رسد، صحت پیش‌بینی روش‌های ابداع شده برای پیش‌بینی ارزش‌های اصلاحی ژنومی نیز یکسان نبوده و این مسأله چالش اصلی پیش روی مطالعات انتخاب ژنومی است. نوِس و همکاران (۲۰) بیان نمودند که یک روش خوب باید عملکرد تقریباً یکسانی برای صفات مختلف داشته باشد، از نظر محاسباتی زیاد پیچیده و زمان بر نباشد و در ضمن اریبی کم و صحت پیش‌بینی بالا داشته باشد.

اخیراً روش‌های یادگیری ماشین به مباحث انتخاب ژنومی وارد شده‌اند. مقوله یادگیری ماشین، شاخه‌ای از هوش مصنوعی است که هدف آن دستیابی به ماشین‌هایی است که قادر به استخراج دانش (یادگیری) از محیط می‌باشند. بنابر تعریف یادگیری ماشین عبارت است از این‌که چگونه می‌توان برنامه‌ای نوشت که از طریق تجربه یادگیری کرده و عملکرد خود را در هر مرحله تصحیح و بهتر کند. ماشین زمانی یاد می‌گیرد که بتواند تغییراتی در ساختارش، برنامه‌اش یا اطلاعاتش ایجاد کند و بنابراین، انتظار می‌رود تا تغییراتی مثبت در عملکرد آینده‌اش ایجاد شود (۲۱). می‌گوییم یک برنامه کامپیوتری از تجربه E در مورد کار T یادگیری انجام داده است اگر عملکرد آن در صورت اندازه‌گیری با معیار P پس از این تجربه بهبود پیدا کند. اگرچه استفاده از الگوریتم‌های یادگیری ماشین در ارزیابی ژنومی به چند سال اخیر محدود می‌شود (۱۷،۱۱، ۲۲)، این الگوریتم‌ها سابقه طولانی‌تری در مباحث بیوانفورماتیک و پزشکی دارند (۷،۶). یکی از مزیت‌های کلیدی روش‌های یادگیری ماشین توانایی آن‌ها در تجزیه و تحلیل داده‌های با ابعاد بسیار بالا می‌باشد. در آینده نزدیک و با در دسترس بودن اطلاعات ژنوتیپی با ابعاد بالا و یا اطلاعات توالی‌یابی ژنومی با حجم بسیار بالا، جایی که قابلیت‌های روش‌های رایج به چالش کشیده خواهد شد، این روش‌ها به خوبی از عهده تجزیه و تحلیل چنین داده‌هایی بر خواهند آمد. در ضمن استخراج روابط پیچیده بین متغیرها مانند اثرات متقابل بین نشانگرها نیز از دیگر مزیت‌های مطلوب این روش‌ها است. این اثرات متقابل از طریق روش‌های رایج در ارزیابی‌های ژنومی قابل استخراج نیستند (۲۲). به دلیل این ویژگی‌های مطلوب، استفاده از روش‌های یادگیری ماشین در مبحث ارزیابی ژنومی روز به روز در حال گسترش است.

جنگل تصادفی یکی از روش‌های یادگیری ماشین است که در عرصه‌های مختلف علوم به طور موفقیت آمیز مورد استفاده قرار گرفته است. همان‌طور که از اسم این الگوریتم بر می‌آید در این روش مجموعه یا جنگلی از درختان مورد استفاده قرار می‌گیرد. هر درخت از ریشه[1]، گره‌ها[2] و برگ‌ها[3] تشکیل شده است. مجموعه‌ای از مثال‌های آموزشی $\{x_i,y_i\}$ که در این‌جا x_i بردار ژنوتیپی هر حیوان و y_i فنوتیپ آن است (با فرض دو کلاس سالم و بیمار برای y) برای آموزش هر درخت مورد استفاده قرار می‌گیرد و درخت یاد می‌گیرد که یک مثال جدید (حیوان دارای اطلاعات ژنوتیپی x_i اما فاقد اطلاعات فنوتیپی y_i) بر اساس اطلاعات ژنوتیپی (x_i) به کدام دسته فنوتیپی تعلق دارد. اگرچه روش‌های درختی از برای تفسیر نتایج ساده بوده اما محدودیت‌هایی نیز دارند؛ برای مثال، میزان اندکی آشفتگی در داده‌های آموزشی منجر به درختی با مدل کاملاً متفاوتی خواهد شد. از این برای مدل‌های درختی چندان با ثبات نیستند. در ضمن برای اطلاعات با حجم و ابعاد بسیار بالا مانند اطلاعات حاصل از مطالعات پویش ژنومی (GWAS)[4]، یک مدل ساده نمی‌تواند پیچیدگی‌های موجود در اطلاعات را پوشش دهد. یک استراتژی برای برطرف نمودن این نقائص استفاده از جنگل یا تجمعی از درخت‌ها است که به روش جنگل تصادفی مشهور است. این کار صحت طبقه‌بندی و پیش‌بینی را افزایش خواهد داد در حالی که بقیه خصوصیات مطلوب یک درخت مثل سادگی تفسیر نتایج حفظ خواهد شد (۸). در مورد روش‌های یادگیری ماشین بحثی تحت عنوان تنظیم کردن وجود دارد که بسیار حائز اهمیت است مانند ماشینی که نیاز به تنظیم موتور دارد تا بهترین عملکرد را داشته باشد. در این تحقیق چگونگی تأثیر تنظیم روش جنگل تصادفی بر عملکرد آن در پیش‌بینی ارزش‌های اصلاحی ژنومی مورد بررسی قرار گرفته است.

1- Root 2- Node 3- Leaf 4- Genome Wide Association Study

مواد و روش‌ها

شبیه‌سازی جمعیت

با استفاده از بسته نرم افزاری hypred (۲۳) ژنومی متشکل از ۵ کروموزوم هر یک به طول ۱ مورگان شبیه‌سازی شد که روی آن ۱۰۰۰۰ نشانگر تک نوکلئوتیدی دو آللی (SNP) با فراوانی اولیه یکسان ۰/۵ به همراه ۱۰۰۰ QTL[1] به طور یکنواخت پخش شدند. در مطالعات شبیه‌سازی ارزیابی ژنومی از سه توزیع نرمال، یکنواخت و گاما برای مدل‌سازی توزیع اثرات QTL موثر بر صفات استفاده می‌شود که معمولاً نتایج ارزیابی ژنومی در هر سه حالت تقریباً برابر است و از این نظر توزیع نرمال حد وسط دو توزیع دیگر است (۱).

در تحقیق حاضر فرض اولیه از توزیع اثرات QTL این بود که این اثرات از توزیع نرمال پیروی می‌کنند. بنابراین اثر جایگزینی QTLها با استفاده از توزیع نرمال استاندارد (با میانگین صفر و واریانس ۱) مدل‌سازی شد. در ضمن به طور مشابه با تحقیقات دیگر (۲۲،۱۱،۸،۳) برای QTLها اثرات ژنتیکی افزایشی منظور شد و از اثرات غالبیت و اپیستازی صرف‌نظر شد. وراثت‌پذیری صفت به میزان ۰/۳ در نظر گرفته شد و از فنوتیپ تصحیح شده به عنوان متغیر پاسخ در بردار **y** استفاده شد. به هر جایگاه SNP با ژنوتیپ AA کد ۲، با ژنوتیپ Aa کد ۱ و با ژنوتیپ aa کد ۰ اختصاص داده شد. جمعیت پایه به تعداد ۱۰۰ فرد (۵۰ نر و ۵۰ ماده) شبیه سازی شده و اجازه داده شد تا برای ۵۰ نسل به طور تصادفی در آن آمیزش صورت گیرد. در این حالت به طور تصادفی از هاپلوتایپ‌های پدری و مادری نمونه‌گیری شده و از آنها برای تولید نتاج استفاده شد. از هر دو والد فقط دو فرزند ایجاد شد که در نتیجه اندازه جمعیت در طی ۵۰ نسل در تعداد ۱۰۰ فرد ثابت باقی ماند. به عبارت دیگر در طی این نسل‌ها اندازه موثر جمعیت[2] ۱۰۰ بود. تحت شرایط ثابت بودن اندازه جمعیت، آمیزش تصادفی به ایجاد LD[3] بین نشانگرها و QTLها منجر خواهد شد. در نسل ۵۱ اندازه جمعیت به ۱۰۰۰ فرد افزایش داده شد که این افراد هم اطلاعات ژنوتیپی داشته و هم ارزش‌های اصلاحی ژنومی آنها مشخص می‌باشد که این افراد جمعیت مرجع[4] را تشکیل دادند.

در ادامه نسل‌های ۵۲ تا ۵۶ از افراد نسل ۵۱ ایجاد شدند که این افراد دارای اطلاعات ژنوتیپی بوده اما اطلاعات فنوتیپی نداشتند. در واقع این نسل‌ها جمعیت‌های تأیید[5] را تشکیل داده بودند که ارزش‌های اصلاحی ژنومی آنها باید پیش‌بینی شود.

تجزیه و تحلیل اطلاعات

در این تحقیق از الگوریتم یادگیر جنگل تصادفی در قالب بسته نرم‌افزاری *randomForest* (۱۵) برای پیش‌بینی ارزش‌های اصلاحی ژنومی استفاده شد. جنگل تصادفی از تجمعی از درختان که هر کدام با استفاده از n نمونه از اطلاعات ورودی که شامل اطلاعات ژنوتیپی و فنوتیپی افراد جمعیت مرجع است ایجاد می‌شود. مدل در جمعیت مرجع آموزش می‌بینند و بر جمعیت تأیید یا کاندید (حیوانات کاندیدای انتخاب) اعمال می‌شود. یکی از n نمونه وارد هر گره از هر درخت می‌شود و از این نمونه اطلاعات یک SNP برای تقسیم‌بندی حیوانات مورد استفاده قرار می‌گیرد به طوری که حیوانات بر اساس اطلاعات ژنوتیپی خود برای SNP انتخاب شده دسته بندی می‌شوند. این کار در گره‌های متوالی انجام می‌شود تا در نهایت به برگ‌ها و یا همان گره‌های پایانی می‌رسیم که در آن‌ها حداکثر یکنواختی وجود خواهد داشت (حیوانات دارای اطلاعات فنوتیپی با ژنوتیپ‌های مشابه برای SNPهای مختلف در یک گره پایانی تجمع می‌یابند). در جمعیت تأیید، پیش‌بینی جنگل تصادفی برای یک مثال ورودی جدید (حیوان دارای اطلاعات ژنوتیپی $\mathbf{x_i}$ اما فاقد اطلاعات فنوتیپی y_i)، $\hat{f}_{rf}^{B}(x)$ ، از طریق میانگین‌گیری از B درخت، $\{T(x,\Psi_b)\}_1^B$ و به صورت زیر انجام می‌شود:

$$\hat{f}_{rf}^{B}(x)=\frac{1}{B}\sum_{b=1}^{B}T(x,\Psi_b),$$

در رابطه فوق Ψ_b، bامین درخت در جنگل را نشان می‌دهد. پارامترهای بسیار مهم در روش جنگل تصادفی تعداد متغیر انتخاب شده در هر گره درخت، تعداد درخت و حداقل اندازه یا حداقل تعداد مشاهدات در گره‌های پایانی یا برگ‌ها می‌باشند که قبل از انجام آنالیزها باید مقدار مناسب آن‌ها تعیین شود. برای تنظیم مدل، ترکیبات مختلفی از تعداد متغیر انتخاب شده در هر گره درخت، تعداد درخت و حداقل اندازه گره‌های پایانی آزمون شد. در ارتباط با داده‌های پیوسته مقدار پیشنهاد شده برای تعداد متغیر انتخاب شده در هر گره درخت برابر $p/3$ است (p تعداد SNP است) که در این تحقیق مقادیر برابر، دو برابر و نصف این مقدار در نظر گرفته شد. بنابراین، مقادیر ۱۵۰۰، ۳۰۰۰ و ۶۰۰۰ برای پارامتر تعداد متغیر انتخاب شده در هر گره درخت، مقادیر ۱۰۰۰، ۱۵۰۰ و ۲۰۰۰ برای تعداد درخت و مقادیر ۱، ۵ و ۱۰ برای حداقل اندازه گره‌های پایانی در نظر گرفته شد و میزان خطای خارج از کیسه برای هر ترکیب محاسبه شد (جدول ۱). در

1- Quantitative Trait Loci
2- Effective Population Size (N_e)
3- Linkage Disequilibrium
4- Reference, Training
5- Test

هر بار نمونه‌گیری با جایگزینی از اطلاعات، برخی اطلاعات (SNPها) هرگز نمونه‌گیری نمی‌شوند و برخی دیگر شاید چند بار نمونه‌گیری شوند. به عبارت دیگر برخی داده‌های ورودی برای برخی درخت‌ها در اصطلاح نمونه خارج از کیسه[1] خواهند بود یعنی در ایجاد برخی درخت‌ها مشارکت نخواهد داشت. این داده‌ها عمل یک اعتبارسنج داخلی برای هر درخت را دارند که این اعتبارسنجی از طریق برآورد خطای خارج از کیسه انجام می‌شود. اگر خود داده‌های خارج از کیسه از طریق درختان پیش‌بینی شوند، برای این پیش‌بینی‌ها خطا وجود خواهد داشت و میانگین این خطاها، خطای خارج از کیسه نامیده می‌شود که نشان‌دهنده میزان تأثیر نمونه‌های انتخاب نشده بر میزان خطای نتیجه نهایی جنگل تصادفی است.

صحت ارزش‌های اصلاحی ژنومی پیش‌بینی شده نیز با مدل‌های مختلف برآورد شده و با هم مقایسه شد. شاخص صحت پیش‌بینی عبارت بود از هم‌بستگی پیرسون بین ارزش‌های اصلاحی ژنومی پیش‌بینی شده و ارزش‌های اصلاحی ژنومی واقعی (شبیه‌سازی شده).

نتایج و بحث

در جدول ۱، ترکیبات مختلف از تعداد متغیر انتخاب شده در هر گره درخت، تعداد درخت و حداقل اندازه گره‌های پایانی به همراه خطای خارج از کیسه مربوط به هر ترکیب آورده شده است. همان‌طور که مشاهده می‌شود ترکیبی شامل تعداد متغیر انتخاب شده در هر گره درخت برابر ۶۰۰۰، تعداد درخت برابر ۱۰۰۰ و حداقل اندازه گره‌های پایانی برابر ۵ حداقل خطا را ایجاد کرده است. اوگوتو و همکاران (۲۲) برای تعداد متغیر انتخاب شده در هر گره درخت، تعداد درخت و حداقل اندازه گره‌های پایانی به ترتیب مقادیر ۳۰۰۰، ۱۰۰۰ و ۱ را منظور نمودند. البته این محققین مقدار خطای خارج از کیسه را برآورد نکرده و با تکرار آنالیزها با ترکیبات مختلف از سه پارامتر مذکور و برآورد صحت ارزش‌های اصلاحی، مقدار بهینه این پارامترها را از مدلی که حداکثر صحت پیش‌بینی را ایجاد کرده بود، استخراج کردند. در مطالعات پویش ژنومی پیشنهاد شده است که مقدار پارامتر تعداد متغیر انتخاب شده در هر گره درخت باید از ۰/۱ تعداد SNP بیشتر باشد ($mtry>0.1p$) (۱۰). در مسایل رگرسیون مانند تحقیق حاضر مقدار پیشنهاد شده برای این پارامتر مقداری نزدیک به $p/3$ است و در مسایل کلاسه‌بندی مقداری نزدیک به میزان $\sqrt{p}$ (۵). البته زمانی که اطلاعات حاوی داده‌های پرت[2] است، $p/3$ کفایت نمی‌کند و باید مقادیر بزرگ‌تر در نظر گرفته شود چرا که در این حالت نمونه‌های گرفته شده از اطلاعات اصلی ممکن است حاوی اطلاعات پرت زیادی باشند که این اطلاعات پرت به ایجاد درخت‌هایی با دقت پیش‌بینی کم منجر می‌شوند (۴)، بنابراین، در حالت اخیر هر چه تعداد متغیر انتخاب شده در هر گره درخت بزرگ‌تر باشد، نسبت داده‌های پرت در نمونه گرفته شده کمتر شده و تأثیر آنها نیز کمتر خواهد شد که این مسأله خصوصاً زمانی که با داده‌های واقعی سر و کار داریم مشهودتر است. در ضمن مقدار این پارامتر تحت تأثیر دو پارامتر دیگر است و با تغییر در دو پارامتر دیگر مقدار بهینه این پارامتر تغییر خواهد کرد. برای مثال در مطالعه اوگوتو و همکاران (۲۲) مقدار این پارامتر برابر ۳۰۰۰ بود اما این مقدار در کنار مقادیر ۱۰۰۰ برای تعداد درخت و ۱ برای حداقل اندازه گره‌های پایانی انتخاب شده بود.

تعداد درخت در جنگل معمولاً رابطه مستقیمی با تعداد متغیر پیش‌بینی‌کننده[3] (در اینجا SNPها) دارد به نحوی که با افزایش تعداد SNP لازم است تعداد درختان نیز افزایش یابد. مقدار پیش فرض برنامه برای تعداد درخت در جنگل ۵۰۰ می‌باشد، اما این مقدار برای ۱۰۰۰۰ SNP و بیشتر کافی نیست.

با افزایش تعداد درخت احتمال اینکه بیشتر SNPها نمونه‌گیری شوند و در ارزیابی ژنومی مشارکت کنند افزایش می‌یابد و به عبارت دیگر هر SNP این فرصت را خواهد داشت که حداقل یک بار نمونه‌گیری شود. زمانی که تعداد درخت ۵۰۰ باشد این بدان معنی است که ۵۰۰ بار نمونه تصادفی با جایگزینی از ۱۰۰۰۰ SNP گرفته خواهد شد. در حالی که وقتی تعداد درخت ۱۰۰۰ باشد ۱۰۰۰ بار نمونه تصادفی با جایگزینی از مجموع SNPها گرفته خواهد شد. به هر حال زمانی که اطلاعات شامل چندین هزار SNP باشد، نباید از مقدار پیش فرض برنامه که ۵۰۰ درخت است استفاده شود و افزایش در تعداد درخت لازم است (۴). البته ممکن است تعداد ۵۰۰ درخت صحتی برابر ۱۰۰۰ یا ۲۰۰۰ درخت ایجاد کند اما قابلیت اطمینان این صحت کم خواهد بود و صحت پیش‌بینی حاصل شده ممکن است در آنالیزهای بعدی به دست نیاید و به عبارت دیگر نتیجه تکرارپذیر نخواهد بود (۴). در ضمن همیشه با افزایش تعداد درخت صحت پیش‌بینی افزایش نمی‌یابد چرا که پارامترهای دیگر مانند تعداد متغیر انتخاب شده در هر گره درخت و حداقل اندازه گره‌های پایانی نیز اثر تعیین‌کننده بر صحت پیش‌بینی دارند. همان‌گونه که در جدول ۱ مشاهده می‌شود با افزایش درخت از ۱۰۰۰ به ۱۵۰۰ و سپس به ۲۰۰۰ نه تنها خطا افزایش پیدا کرده بلکه صحت پیش‌بینی نیز

1- Out Of Bag
2- Outliers
3- Predictor variable

کاهش پیدا کرده است. در این حالت زمان محاسبات نیز به طور محسوس (تا دو برابر) افزایش می‌یابد.

مقدار مناسب حداقل اندازه گره‌های پایانی برابر ۵ تعیین شد. هرچه مقدار این پارامتر بزرگ‌تر باشد، درختان کوچک‌تری تولید خواهد شد و زمان محاسبات کاهش می‌یابد، اما نکته منفی این است که ممکن است برخی متغیرهای پیش‌بینی کننده (در این‌جا SNPها) دو گروه (آلل) با فراوانی بالا و پایین داشته باشند و در این حالت تعداد مشاهدات در گره برای آلل با فراوانی کم ممکن است از مقدار از پیش تعیین شده کمتر شده و در نتیجه SNP مورد نظر در ارزیابی ژنومی مشارکت داده نشود. پیشنهاد شده است که از مقادیر کوچک این پارامتر در مطالعات بیوانفورماتیک و ژنومی استفاده شود (۴). همان‌گونه که مشاهده می‌شود با ثابت نگه داشتن یک پارامتر مانند تعداد درخت، با تغییر در پارامترهای تعداد متغیر انتخاب شده در هر گره درخت و حداقل اندازه گره‌های پایانی هر دو شاخص خطای خارج از کیسه و صحت پیش‌بینی ارزش‌های اصلاحی ژنومی تغییر می‌کند. نتیجه اخیر مؤید این مطلب است که یک رابطه داخلی بین پارامترها برقرار است و ترکیب بهینه از این پارامترها باید جستجو شود و برای تجزیه و تحلیل اطلاعات استفاده شود.

در جدول ۲ هم‌بستگی بین ارزش‌های اصلاحی ژنومی پیش‌بینی شده و ارزش‌های اصلاحی ژنومی واقعی (شبیه‌سازی شده) مربوط به حیوانات نسل ۵۲ تا ۵۶ (حیوانات فاقد مقادیر فنوتیپی) آورده شده است. در این‌جا نیز حداکثر صحت پیش‌بینی ارزش‌های اصلاحی ژنومی مربوط به ترکیبی از سه پارامتر تعداد متغیر انتخاب شده در هر گره درخت، تعداد درخت و حداقل اندازه گره‌های پایانی است که حداقل خطا را ایجاد کرده‌اند (به ترتیب ۶۰۰۰، ۱۰۰۰ و ۵). در این مدل صحت پیش‌بینی در مقایسه با مدل‌های دیگر بالاتر است. نوس و همکاران (۲۰) با تجزیه و تحلیل برخی صفات در موش‌های آزمایشگاهی و برای صفات مختلف با وراثت‌پذیری متفاوت، صحت پیش‌بینی ارزش‌های اصلاحی ژنومی از طریق الگوریتم جنگل تصادفی را در دامنه ۰/۲ تا ۰/۸ گزارش نمودند.

جدول ۱- ترکیبات مختلف از تعداد متغیر انتخاب شده در هر گره درخت (mtry)، تعداد درخت (ntree) و حداقل اندازه گره‌های پایانی (nodesize) به همراه خطای خارج از کیسه (OOB error) مربوط به هر ترکیب

ntree	*mtry*	*nodesize*	*OOB error*
۱۰۰۰	۱۵۰۰	۱	۲۴۱/۲۴۲
۱۰۰۰	۱۵۰۰	۵	۲۴۰/۹۴۱
۱۰۰۰	۱۵۰۰	۱۰	۲۴۲/۵۶۰
۱۰۰۰	۳۰۰۰	۱	۲۳۷/۳۴۵
۱۰۰۰	۳۰۰۰	۵	۲۳۶/۴۵۲
۱۰۰۰	۳۰۰۰	۱۰	۲۳۶/۸۰۰
۱۰۰۰	۶۰۰۰	۱	۲۳۵/۸۳۴
۱۰۰۰	۶۰۰۰	۵	۲۳۴/۶۶۹
۱۰۰۰	۶۰۰۰	۱۰	۲۳۷/۲۱۰
۱۵۰۰	۱۵۰۰	۱	۲۳۹/۶۵۴
۱۵۰۰	۱۵۰۰	۵	۲۳۶/۱۴۰
۱۵۰۰	۱۵۰۰	۱۰	۲۳۷/۳۸۶
۱۵۰۰	۳۰۰۰	۱	۲۳۹/۸۵۶
۱۵۰۰	۳۰۰۰	۵	۲۳۶/۰۷۳
۱۵۰۰	۳۰۰۰	۱۰	۲۳۷/۵۴۳
۱۵۰۰	۶۰۰۰	۱	۲۳۶/۱۵۴
۱۵۰۰	۶۰۰۰	۵	۲۳۵/۶۸۳
۱۵۰۰	۶۰۰۰	۱۰	۲۳۸/۴۹۳
۲۰۰۰	۱۵۰۰۰	۱	۲۴۰/۶۵۴
۲۰۰۰	۱۵۰۰	۵	۲۳۷/۷۸۴
۲۰۰۰	۱۵۰۰	۱۰	۲۳۹/۳۵۴
۲۰۰۰	۳۰۰۰	۱	۲۴۳/۵۲۲
۲۰۰۰	۳۰۰۰	۵	۲۴۱/۵۴۰
۲۰۰۰	۳۰۰۰	۱۰	۲۴۲/۳۶۱
۲۰۰۰	۶۰۰۰	۱	۲۴۴/۷۶۵
۲۰۰۰	۶۰۰۰	۵	۲۴۲/۲۱۵
۲۰۰۰	۶۰۰۰	۱۰	۲۴۴/۶۵۱

جدول ۲- صحت پیش‌بینی (هم‌بستگی بین ارزش‌های اصلاحی ژنومی پیش‌بینی شده و واقعی) ارزش‌های اصلاحی ژنومی پیش‌بینی شده در مدل‌های مختلف در حیوانات کاندیدای انتخاب مربوط به نسل‌های ۵۲ تا ۵۶

ntree	*mtry*	*nodesize*	G 52	G 53	G 54	G 55	G 56
۱۰۰۰	۱۵۰۰	۵	۰/۵۷۳	۰/۴۸۷	۰/۴۲۶	۰/۴۰۷	۰/۴۰۰
۱۰۰۰	۳۰۰۰	۵	۰/۵۵۴	۰/۵۰۷	۰/۴۳۱	۰/۴۱۴	۰/۳۹۵
۱۰۰۰	۶۰۰۰	۵	۰/۵۹۷	۰/۵۳۶	۰/۴۵۰	۰/۴۳۴	۰/۴۱۸
۱۵۰۰	۱۵۰۰	۵	۰/۵۴۴	۰/۵۲۲	۰/۴۱۸	۰/۳۸۶	۰/۳۸۱
۱۵۰۰	۳۰۰۰	۵	۰/۵۶۶	۰/۵۱۵	۰/۴۳۹	۰/۴۱۰	۰/۴۱۴
۱۵۰۰	۶۰۰۰	۵	۰/۵۳۴	۰/۵۰۹	۰/۴۱۷	۰/۳۹۵	۰/۳۷۹
۲۰۰۰	۱۵۰۰	۵	۰/۵۷۷	۰/۵۲۵	۰/۴۳۱	۰/۳۸۶	۰/۳۷۱
۲۰۰۰	۳۰۰۰	۵	۰/۵۷۱	۰/۵۲۰	۰/۴۲۷	۰/۴۲۱	۰/۴۱۷
۲۰۰۰	۶۰۰۰	۵	۰/۵۷۹	۰/۵۱۳	۰/۴۲۴	۰/۳۹۳	۰/۳۸۶

G 52 تا G 56 نسل‌های ۵۲ تا ۵۶ می‌باشند که حیوانات کاندیدای انتخاب فاقد ارزش‌های اصلاحی ژنومی هستند و ارزش‌های اصلاحی آنها بر اساس مدل به دست آمده از جمعیت مرجع به دست می‌آید.

هرچه از جمعیت مرجع دورتر شویم کاهش صحت پیش‌بینی ارزش‌های اصلاحی ژنومی قابل انتظار است چرا که با افزایش فاصله بین جمعیت رفرنس و جمعیت تأیید فاز LD بین نشانگر و QTL به هم می‌خورد و نشانگرها به خوبی نمی‌توانند اثرات QTL را منعکس کنند (۹). باستیانسن و همکاران (۲) گزارش کردند که صحت پیش‌بینی ارزش‌های اصلاحی ژنومی برآورد شده افراد کاندیدای انتخاب تحت تأثیر فاصله از جمعیت تأیید قرار گرفته و از ۰/۴۷ در نسل اول به ۰/۰۷ در نسل ۱۰ کاهش می‌یابد. نتایج مشابه از سوی عبداللهی و همکاران (۱) و بوستان و همکاران (۳) گزارش شده است. در جمعیت‌های حیوانات اهلی، تغییر در فاز LD با گذشت زمان و افزایش فاصله از جمعیت مرجع عمدتاً به دلایل نوترکیبی، انتخاب و مهاجرت بروز می‌نماید (۲۵). در صورتی که کاهش در صحت پیش‌بینی ارزش‌های اصلاحی چشم‌گیر باشد باید مجدداً جمعیت مرجع تشکیل شود و اثرات نشانگرها برآورد شود (۱۸). افزایش در اندازه جمعیت مرجع و استفاده از پنل‌های نشانگر با تراکم بالاتر نیز برای جلوگیری از کاهش صحت پیش‌بینی ارزش‌های اصلاحی ژنومی در نتیجه افزایش فاصله از جمعیت مرج با گذشت زمان پیشنهاد شده‌اند (۲۵).

به طور کلی نتایج این تحقیق نشان داد که برای کاربرد الگوریتم جنگل تصادفی در ارزیابی ژنومی، برای حصول نتایج قابل اعتماد، ترکیب مناسبی از تعداد متغیر انتخاب شده در هر گره درخت، تعداد درخت و حداقل اندازه گره‌های پایانی باید مورد استفاده قرار گیرد چرا که ترکیبات متفاوت از این پارامترها نتایج متفاوتی را ایجاد می‌کنند. مقادیر انتخاب شده برای پارامترهای فوق در این تحقیق خاص تحقیق حاضر بوده و در تحقیقات دیگر خصوصاً در صورتی که ساختار جمعیت (تعداد افراد در جمعیت رفرنس، تعداد موثر جمعیت و ...) و معماری ژنتیکی صفت (تعداد نشانگر، تعداد QTL و ...) متفاوت باشد، باید برای جمعیت مورد نظر مقادیر مناسب آنها جست‌وجو و انتخاب شود؛ برای مثال در این تحقیق برای پارامتر تعداد متغیر انتخاب شده در هر گره درخت مقدار ۶۰۰۰ در نظر گرفته شد، حال آن‌که اگر در تحقیق دیگری اطلاعات ۵۰۰۰ SNP شبیه‌سازی شود نمی‌توان به مقدار استفاده شده در این تحقیق که برابر ۶۰۰۰ است استناد شود چرا که اگر مقدار پارامتر تعداد متغیر انتخاب شده در هر گره درخت از تعداد SNPها بیشتر باشد تجزیه و تحلیل‌ها قابل انجام نخواهد بود و برنامه متوقف می‌شود.

تشکر و قدردانی

پیشنهادات ارزنده دوست گرانقدر جناب آقای دکتر رستم عبداللهی برای شبیه‌سازی ساختار جمعیت بسیار راهگشا بود. بدین وسیله از زحمات ایشان کمال قدردانی می‌شود.

منابع

1. Abdollahi Arpahahi, R., A. Pakdel, A. Nejati-Javaremi and M. Moradi Shahrbabak. 2013. Comparison of different genomic evaluation methods for traits with different genetic architecture. Animal Production, 15: 65-77.
2. Bastiaansen, J.W.M., A. Coster, M.P.L. Calus, J.A.M. van Arendonk and H. Bovenhuis. 2011. Long-term response to genomic selection: effects of estimation method and reference population structure for different genetic architectures. Genetics Selection Evolution, 44: 3 pp.
3. Boustan, A., A. Nejati-Javaremi, M. Moradi Shahrbabak and M. Saatchi. 2012. Effect of using different number and type of records from different generations as reference population on the accuracy of genomic evaluation. Archiv Tierzucht, 56: 68 pp.
4. Boulesteix, A.L., S. Janitza, J. Kruppa and I.R. König. 2012. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Technical Report. Department of Statistics, University of Munich.
5. Breiman, L. 2001. Random forests. Machine Learning, 45: 5-32.
6. Bureau, A., J. Dupuis, B. Hayward, K. Falls and P. Van Eerdewegh. 2003. Mapping complex traits using random forests. BMC Genetics, 4: 64 pp.
7. Bureau, A., J. Dupuis, K. Falls, K. Lunetta, B. Hayward and T. Keith. 2005. Identifying SNPs predictive of phenotype using random forests. Genetic Epidemiology, 28: 171-182.
8. Chen, X., M. Wang and H. Zhang. 2011. The use of classification tree for bioinformatics. WIREs Data Mining and Knowledge Discovery, 1: 55-63.
9. Erbe, M., B.J. Hayes, L.K. Matukumalli, S. Goswami, P.J. Bowman, C.M. Reich, B.A. Mason and M.E. Goddard. 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. Journal of Dairy Science, 95: 4114-4129.
10. Goldstein, B.A., A.E. Hubbard, A. Cutler and L.F. Barcellos. 2010. An application of random forests to a genome-wide association dataset: Methodological considerations and new findings. BMC Genetics, 11: 49 pp.
11. González-Recio, O. and S. Forni. 2011. Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. Genetics Selection Evolution, 43: 7 pp.
12. Habier, D., R.L. Fernando and J.C. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics, 177: 2389-2397.
13. Hill, W.G. 2008. Estimation, effectiveness and opportunities of long term genetic improvement in animals and maize, Lohmann Information, 43: 3-20.
14. Khatkar, M.S., M. Moser, B.J. Hayes and H.W. Raadsma. 2012. Strategies and utility of imputed SNP genotypes for genomic analysis in dairy cattle. BMC Genomics, 13: 538 pp.
15. Liaw, A. and M. Wiener. 2002. Classification and regression by random forest. R News. 2:18-22.
16. Meuwissen, T.H.E., B.J. Hayes and M.E. Goddard. 2001. Prediction of total genetic value using genome wide dense marker maps. Genetics, 157: 1819-1829.
17. Moser, G., B. Tier, R.E. Crump, M.S. Khatkar and H.W. Raadsma. 2009. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genetic Selection Evolution, 41: 56 pp.
18. Muir, W.M. 2007. Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. Journal of Animal Breeding and Genetics, 124: 342-355.
19. Nejati-Javaremi, A., C. Smith and J. Gibson. 1997. Effect of total allelic relationship on accuracy of evaluation and response to selection. Journal of Animal Science, 75: 1738-1745.
20. Neves, H.H.R., R. Carvalheiro and S.A. Queiroz. 2012. A comparison of statistical methods for genomic selection in a mice population. BMC Genetics, 13:100.
21. Nilsson, N.J. 1998. Introduction to Machine Learning. Stanford University. Stanford, USA. 412 pp.
22. Ogutu, J.O., H.P. Piepho and T. Schulz-Streeck. 2011. A comparison of random forests, boosting and support vector machines for genomic selection. BMC Proc. 5(Suppl. 3): S11.
23. Technow, F. 2014. hypred: Simulation of genomic data in applied genetics. Available at: http://cran.r-project.org/web/packages/hypred/hypred.pdf
24. Visscher, P.M. and C.S. Haley. 1998. Power of a chromosomal test to detect genetic variation using genetic markers. Heredity, 81: 317-326.
25. Zhe, Z., Q. Zhang and X.D. Ding. 2011. Advances in genomic selection in domestic animals. Chinese Science Bulletin, 56: 2655-2663.

Tuning and Application of Random Forest Algorithm in Genomic Evaluation

Farhad Ghafouri Kesbi[1], Ghodrat Rahimi Mianji[2], Mahmoud Honarvar[3] and Ardeshir Nejati Javaremi[4]

1- Ph.D. Student, Sari Agricultural Sciences and Natural Resources University,
(Corresponding Author: email:farhad_ghy@yahoo.com)
2- Professor, Sari Agricultural Sciences and Natural Resources University
3- Assistant Professor, of Animal Science, Islamic Azad University, Shahre Qods Branch, Karaj
4- Assistant Professor, College of Agriculture and Natural Resources, University of Tehran
Received: June 29, 2014 Accepted: September 21, 2014

Abstract

One of the most important issues in genomic selection is using a decent method for estimating marker effects and genomic evaluation. Recently, machine learning algorithms which are members of non-parametric and non-linear methods have been extended to genomic evaluation. One of these methods is Random Forest (RF) on which this research was focused. Important parameters in RF algorithm are the number of SNPs selected randomly at each tree node (*mtry*), the number of trees to grow` (*ntree*) and the minimum size of terminal nodes of trees (*node size*) which need to be pre-defined before analyses and for them the model should be tuned. A genome comprised of five chromosomes, one Morgan each, on which 10000 bi-allelic SNP were arrayed was simulated and the efficiency of different combinations of *mtry*, *ntree* and *node size* was tested and the best combination was selected based on comparison of accuracy of predicted genomic value as well as OOB error estimates. For the simulated data in the current study the least OOB error as well as the maximum prediction accuracy was related to a model with 6000 *mtry*, 1000 *ntree* and 5 *node size*. Other combinations did not increase the accuracy of prediction while led to an increase in time of analyses for those which used more trees. Since the accuracy of prediction is a function of *mtry*, *ntree* and *node size*, in genomic evaluation, different combinations of these parameters should be used and the combination which caused the maximum prediction accuracy should be used for genomic evaluation.

Keywords: Genomic Breeding Value, Genomic Evaluation, Random Forest, Single Nucleotide Marker, Tree