دامستیک؛ دوره ۲۰، شماره ۲ (شماره ۱۷ پیاپی): ۲۹-۱۹

نشریهٔ علمی- ترویجی (حرفه‌ای)

دامستیک

انجمن علمی- دانشجویی گروه علوم دامی دانشگاه تهران؛ پاییز ۱۳۹۹

https://domesticsj.ut.ac.ir/article_79151.html

مقاله علمی- ترویجی

کاربرد رویکرد یادگیری ماشین و الگوریتم‌های زیر مجموعهٔ آن در برآورد ارزش‌های اصلاحی ژنومی

فرزاد غفوری[۱]، سمیه علی‌پور[۲*] و صادق محمدیان جشوقانی[۳]

[۱] دانشجوی دکتری تخصصی ژنتیک و اصلاح نژاد دام و طیور، گروه علوم دامی، پردیس کشاورزی و منابع طبیعی دانشگاه تهران، کرج، ایران

[۲] دانشجوی دکتری ژنتیک و اصلاح نژاد دام، گروه علوم دامی، دانشکده کشاورزی دانشگاه تربیت مدرس، تهران، ایران

[۳] کارشناسی ارشد ژنتیک و اصلاح‌نژاد دام، گروه علوم دامی، پردیس کشاورزی و منابع طبیعی دانشگاه تهران، کرج، ایران

https://doi.org/10.22059/domesticsj.2020.310252.1050

چکیده

هدف از انتخاب ژنومی استفاده همزمان از داده‌های ژنوتیپی به همراه داده‌های فنوتیپی است تا بتوان در مدت زمان کوتاه، دام‌ها را ارزیابی نموده و دام‌های برتر از نظر ژنتیکی را گزارش نمود. توسعه الگوریتم‌های داده کاوی مرتبط با آبر داده‌ها در عصر دیجیتال در برآورد ارزش‌های اصلاحی نقش قابل توجهی در اصلاح‌نژاد دام و طیور ایفا می‌کند. اخیراً روش‌های یادگیری ماشین و الگوریتم‌های زیرمجموعهٔ آن مانند یادگیری عمیق، جنگل تصادفی، ماشین بردار پشتیبان و بوستینگ که جزء روش‌های غیرپارامتریک هستند، به مباحث انتخاب ژنومی وارد شده‌اند. یکی از مزایای روش‌های یادگیری ماشین، پتانسیل و کارآیی بسیار بالای آن‌ها به خصوص برای داده‌های با حجم بالا یا به اصلاح آبر داده‌ها و برآورد اثرات غیرافزایشی مانند غالبیت و اپیستازی و همچنین بررسی روابط پیچیده بین متغیرها (مانند اثرات متقابل بین نشانگرها) است. ایده اصلی در این الگوریتم‌ها استفاده از داده‌های آموزشی (در این جا اطلاعات ژنوتیپی و فنوتیپی حیوانات جمعیت مرجع) است تا الگوریتم براساس اطلاعات ژنوتیپی افراد جمعیت کاندید، ارزش‌های اصلاحی ژنومی آن‌ها را پیش‌بینی نماید. برخی از این روش‌ها به طور موفقیت‌آمیزی در ارزیابی‌های ژنومی مورد استفاده قرار گرفته‌اند و نتایج قابل قبولی را با حداقل خطای ممکن ارائه داده‌اند. در واقع هدف از این مطالعه بیانِ تعریفی از رویکرد یادگیری ماشین و الگوریتم‌های زیرمجموعهٔ آن و نیز نقش آن‌ها در پیش‌بینی معماری ژنتیکی صفات با وراثت‌پذیری پیچیده است. در نتیجه، احتمالاً استفاده از رویکرد یادگیری ماشین با هدف یافتن کارآمدترین الگوریتم، همزمان با افزایش حجم داده‌های فنوتیپی و ژنومی تأثیر قابل توجهی در آینده اصلاح‌نژاد دام و طیور، به ویژه پیشرفت ژنتیکی دام‌ها به دنبال خواهد داشت.

کلمات کلیدی: یادگیری ماشین، الگوریتم ژنتیکی، یادگیری عمیق، جنگل تصادفی، ارزش اصلاحی، روش‌های غیرپارامتریک

***نویسنده مسئول:** s_alipour@modares.ac.ir

تاریخ دریافت: ۱۳۹۹/۰۶/۲۷ **تاریخ بازنگری:** ۱۳۹۹/۰۷/۰۳ **تاریخ پذیرش:** ۱۳۹۹/۰۷/۱۶ **تاریخ انتشار آنلاین:** ۱۳۹۹/۰۹/۲۰

رفرنس‌دهی: غفوری، ف.، علی‌پور، س.، محمدیان جشوقانی، ص. کاربرد رویکرد یادگیری ماشین و الگوریتم‌های زیر مجموعهٔ آن در برآورد ارزش‌های اصلاحی ژنومی. *علمی- ترویجی (حرفه‌ای) دامستیک*، ۱۳۹۹؛ ۲۰(۲): ۲۹-۱۹.

مقدمه

با مطرح نمودن مدل‌های مختلط در اصلاح نژاد دام و حل روابط خویشاوندی با تجزیه چولسکی توسط هندرسون، انقلابی عظیم در اصلاح نژاد دام و طیور ایجاد شد. به تدریج علم اصلاح‌نژاد در سه عرصهٔ شاخص انتخاب، برآورد اجزاء واریانس و روش‌های پیش‌بینی ارزش اصلاحی حیوانات گسترش یافت. در نهایت به معرفی روش حداکثر درست‌نمایی (REML: Restricted Maximum Likelihood) و بهترین پیش بینی نااریب خطی (BLUP: Best Linear Unbiased Prediction) منتهی گردید که این دو ابزار هم اکنون نیز در عرصه اصلاح نژاد دام دارای جایگاه ویژه‌ای می‌باشند (Hofer, 1998).

یکی از مسائل مطرح‌شده در انتخاب ژنومی بحث برآورد اثر نشانگرها است، به گونه‌ای که بر این اساس روش‌های مختلفی برای برآورد اثر نشانگرها توسعه یافته‌اند. مشابه با روش‌های مبتنی بر اطلاعات فنوتیپی مانند مدل پدری، مدل حیوانی تک صفته و چند صفته و مدل رگرسیون تصادفی که صحت پیش‌بینی ارزش اصلاحی در آن‌ها متفاوت می‌باشد، صحت پیش‌بینی روش‌های ابداع شده برای برآورد اثر SNPها و پیش‌بینی ارزش اصلاحی ژنومی (GEBV: Genome Estimated Breeding Value) نیز یکسان نبوده و این مسئله به چالش اصلی پیش روی مطالعات انتخاب ژنومی مبدل گشته است (Nejati-Javaremi *et al.*, 1997).

تا همین اواخر نیز بیشتر روش‌های مورد استفاده جهت برآورد اثر نشانگرها، مدل‌های خطی بودند که در آن‌ها اثر SNPها از طریق رگرسیون فنوتیپ یا ارزش اصلاحی روی SNPها به دست می‌آمد. روش‌های حداقل مربعات و انواع روش‌های بیزی از جمله این روش‌ها می‌باشند. بنابراین پیشرفت فناوری‌های توالی‌یابی DNA، به ویژه در عصر جدید موجب ظهور حجم عظیمی از داده‌های ژنتیک مولکولی شده است (Ghafouri *et al.*, 2020).

اخیراً روش‌های یادگیری ماشین (Machine Learning) به مباحث انتخاب ژنومی وارد شده‌اند. استفاده از این روش‌ها گستره وسیعی از علوم، از زمین‌شناسی تا پزشکی و جرم شناسی را در بر می‌گیرد؛ به گونه‌ای که این روش‌ها در تمامی این عرصه‌ها به طور موفقیت‌آمیزی مورد استفاده قرار گرفته‌اند. این روش‌ها در کل با روش‌های رایج در ارزیابی ژنومی متفاوت هستند و ایده اصلی در آن‌ها آموزش یک الگوریتم با استفاده از داده‌های آموزشی (در این جا اطلاعات ژنوتیپی و فنوتیپی حیوانات جمعیت مرجع) است تا الگوریتم یاد بگیرد که براساس اطلاعات ژنوتیپی افراد جمعیت کاندید، ارزش‌های اصلاحی ژنومی آن‌ها را پیش‌بینی نماید. برخی از این روش‌ها به طور موفقیت‌آمیزی در ارزیابی‌های ژنومی مورد استفاده قرار گرفته‌اند و نتایج تحقیقات انجام شده مؤید این مطلب است که عملکرد آن‌ها قابل مقایسه با روش های رایج ارزیابی ژنومی مانند GBLUP یا RR-BLUP است (Pérez-Enciso and Zingaretti, 2019).

یکی از مزیت‌های کلیدی این روش‌ها توانایی آن‌ها در تجزیه و تحلیل داده‌های با ابعاد بالا می‌باشد. در آیندهٔ نزدیک و با در دسترس بودن اطلاعات ژنوتیپی گسترده 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 (González-Recio *et al.*, 2013). کاربرد این روش‌ها در پیش‌بینی ارزش‌های اصلاحی ژنومی به چند سال اخیر محدود می‌شود، امّا به دلیل ویژگی‌های مطلوب و بهینه این روش‌ها، استفاده از آن روز به روز در حال گسترش است.

روش‌های برآورد ارزش‌های اصلاحی به دو دستهٔ پارامتریک و غیرپارامتریک دسته‌بندی می‌شوند. روش‌های پارامتریک شامل BLUP، GBLUP، روش‌های بیزین (Bayesian methods) و ... است و روش‌های غیرپارامتریک نیز شامل شبکه‌های عصبی (Neural Networks)، هوش مصنوعی (Artificial intelligence)، یادگیری ماشین (Machine Learning) هستند. توسعه الگوریتم‌های داده کاوی مرتبط با اَبر داده‌ها در برآورد ارزش‌های اصلاحی نقش قابل توجهی خواهند داشت. مجموعه‌ای از تکنولوژی‌ها همچون یادگیری ماشین، هوش مصنوعی و یادگیری عمیق در عصر جدید فرصت‌های مناسبی را در مقایسه با روش‌های سنتی برای بررسی صفات اقتصادی با معماری پیچیده فراهم ساخته‌اند (Ghafouri *et al.*, 2020). در این مطالعه به بیانِ تعریفی از رویکرد یادگیری ماشین، الگوریتم‌های زیرمجموعه آن و نیز نقش آن‌ها در برآورد ارزش‌های اصلاحی به ویژه برآورد اثرات غیرافزایشی مثل غالبیت و اپیستازی در اصلاح‌نژاد دام و طیور پراخته خواهد شد.

یادگیری ماشین (Machine Learning)

مقوله یادگیری ماشین، شاخه‌ای از هوش مصنوعی است که هدف آن دستیابی به ماشین‌ها یا الگوریتم‌هایی است که قادر به استخراج دانش (یادگیری) از محیط می‌باشند. بنابر تعریف، یادگیری ماشین عبارت است از این که چگونه می‌توان برنامه‌ای نوشت که از طریق تجربه، آموزش ببیند و عملکرد خود را در هر مرحله تصحیح و بهتر نماید.

ماشین زمانی که بتواند تغییراتی در ساختار، برنامه و یا اطلاعات خود ایجاد کند، یاد می‌گیرد؛ بنابراین انتظار می‌رود تا تغییراتی مثبت در عملکرد آینده آن ایجاد شود (Nilsson, 1998). گفته می‌شود یک برنامه کامپیوتری از تجربه E در مورد کار T یادگیری انجام داده است، اگر عملکرد آن در صورت اندازه‌گیری با معیار P بهبود پیدا کند. یک مثال در این باره می‌تواند بازی شطرنج باشد که در آن انجام بازی کار (T) و درصد بازی‌هایی که ماشین بر حریف غلبه می‌کند، معیار اندازه‌گیری یادگیری (P) خواهد بود.

یادگیری ماشین در مباحث مختلفی که در آن‌ها بحث طبقه‌بندی (ماشین یاد می‌گیرد که ورودی‌ها را به دسته‌هایی از پیش تعیین‌شده نسبت دهد)، خوشه‌بندی (ماشین کشف می‌کند که کدام ورودی‌ها با هم در یک دسته جای می‌گیرند) و پیش‌بینی (ماشین مقدار عددی یک ورودی را پیش‌بینی می‌کند) وجود داشته باشد، کاربرد دارد. بنابراین، دامنه کاربرد این روش‌ها بسیار وسیع بوده و در حوزه‌های مختلف علوم از هواشناسی گرفته تا پزشکی، جرم‌شناسی، مطالعات زمین‌شناسی و ... کاربرد دارند (Bishop, 2006). بر این اساس تاکنون رویکرد یادگیری ماشین به طور موفقیت‌آمیزی در مواردی مانند کنترل ربات‌ها، تشخیص چهره، شناسایی گفتار، شناسایی متن، بازی‌های کامپیوتری و بیوانفورماتیک به کار گرفته شده است.

روش‌های یادگیری ماشین با توجه به غیرپارامتریک بودن و تئوری‌هایی که بر مبنای آن‌ها بنا شده‌اند، اساساً با روش‌های مبتنی بر معادلات خطی کاملاً متفاوت هستند. این روش‌ها قادر هستند ارتباطات مختلف بین ژن‌ها مانند اثرات اپیستاتیک و رابطه پیچیده از ژنوتیپ تا فنوتیپ را به نحو مؤثرتری کنکاش نمایند (Heslot *et al.*, 2012).

روش‌های یادگیری ماشین مانند یادگیری عمیق (Deep Learning) (Bellot *et al.*, 2018)، Random Forest (Breiman, 2001)، Support Vector Machine (Boser *et al.*, 1992) و Boosting (Freund and Schapire, 1997) جهت مدل‌سازی واریانس ژنتیکی و عوامل محیطی، مطالعه شبکه‌های ژنی، مطالعات پویش کل ژنوم (GWAS: Genome Wide Association Study)، مطالعه اثرات اپیستاتیک و ارزیابی ژنومی مورد استفاده قرار گرفته‌اند؛ امّا استفاده از آن‌ها در مباحث اصلاح نژاد دام به سال‌های اَخیر محدود می‌شود (جدول ۱). روش‌های یادگیری ماشین شامل موارد به خصوصی است که در ادامه مورد بحث قرار می‌گیرند.

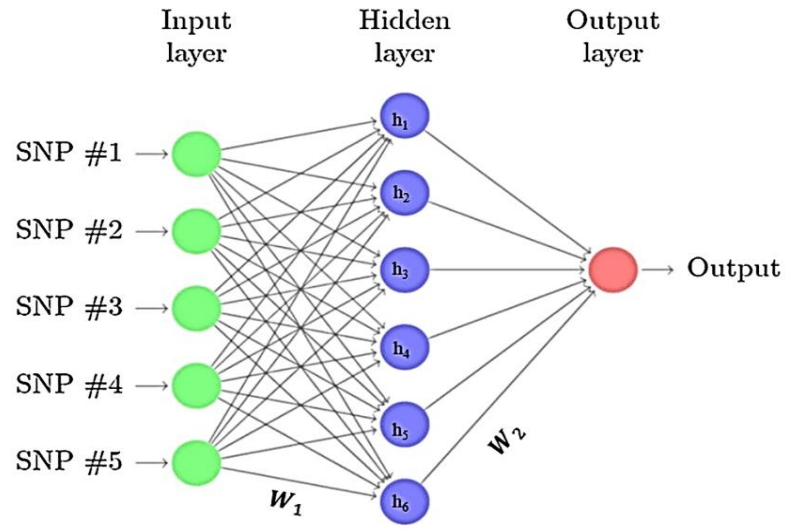
جدول ۱- معرفی الگوریتم‌های زیرمجموعهٔ یادگیری ماشین و بیان تعدادی از ویژگی‌های آن‌ها

	الگوریتم‌های زیرمجموعه	ویژگی‌های مهم
۱	یادگیری عمیق (Deep Learning)	افزایش گسترده قدرت محاسبات دسترسی به داده‌های بزرگ (اَبرداده‌ها) استفاده از لایه‌ها و فیلترهای مختلف پتانسیل بالای برآورد اثرات غیرافزایشی (مانند غالبیت و اپیستازی)
۲	جنگل تصادفی (Random Forest)	استفاده در عرصه‌های مختلف علوم نیاز به فرضیات کمتر دربارهٔ توزیع داده‌ها کاهش خطای پیش‌بینی از طریق استراتژی بگینگ
۳	ماشین بردار پشتیبان (Support Vector Machine)	صحت بالای پیش‌بینی‌های ژنومی عدم نیاز به نحوهٔ توارث جمعیت توانایی بالا در بکارگیری اثرات غیرافزایشی
۴	بوستینگ (Boosting)	توانایی بالا در آنالیز ژنومی صفات آستانه‌ای توانایی بالا در آنالیز داده‌های ژنومی با حجم بالا قدرت بالا در مدل‌سازی بهبود و افزایش صحت پیش‌بینی ژنومی

یادگیری عمیق (Deep Learning)

برخی از الگوریتم‌های پیشرفته یادگیری ماشین مانند الگوریتم‌های یادگیری عمیق (Deep Learning) ممکن است در ارزیابی‌ها و پیش‌بینی‌های ژنومی کارآیی گسترده‌ای داشته باشند. یادگیری عمیق (DL)، زیرمجموعه‌ای از روش‌های یادگیری ماشین است که از ساختار و عملکرد مغز برای طراحی آن الهام گرفته شده است و اساساً مجموعه‌ای از شبکه‌های عصبی با تعداد زیادی از گره و لایه می‌باشد (Abdollahi-Arpanahi *et al.*, 2020).

این الگوریتم‌ها تا حد زیادی با افزایش گسترده قدرت محاسبات و دسترسی به داده‌های بزرگ همراه هستند. الگوریتم‌های یادگیری عمیق مانند پرسپترون چند لایه (MLP: Multilayer Perceptron) و شبکه عصبی همگشتی (CNN: Convolutional Neural Network) ممکن است که بتوانند از الگوریتم‌های ناشناخته عدم تعادل لینکاژی استفاده کنند. MLPها شبکه‌های عصبی ساده‌ای هستند که از لایه‌های مختلفی تشکیل شده است و فاقد هر نوع فیلتری می‌باشند. در حالت کلی، لایه‌های مربوطه را می‌توان به سه لایهٔ ورودی (Input layer)، میانی یا پنهانی (Hidden layers) و خروجی (Output layer) دسته‌بندی کرد. هر مجموعهٔ اطلاعاتی به عنوان مجموعه آموزشی (ورودی) به MLP داده می‌شود، پس از وزن‌یابی برای به حداقل رساندن میزان خطا، در نهایت فنوتیپ‌ها، ارزش‌های اصلاحی و یا همبستگی بین مقادیر واقعی و برآورد شده به عنوان خروجی برآورد می‌گردد (شکل ۱). MLPها نسبت به CNNها قدرت برآورد پایین‌تری دارند؛ امّا CNNها چون علاوه بر لایه‌ها از فیلترهای مخصوص متناسب با مطالعه مورد نظر استفاده می‌کنند، پتانسیل و کارآیی بسیار بالایی به خصوص برای داده‌های با حجم بالا و برآورد اثرات غیرافزایشی مانند غالبیت و اپیستازی دارند (Abdollahi-Arpanahi *et al.*, 2020).

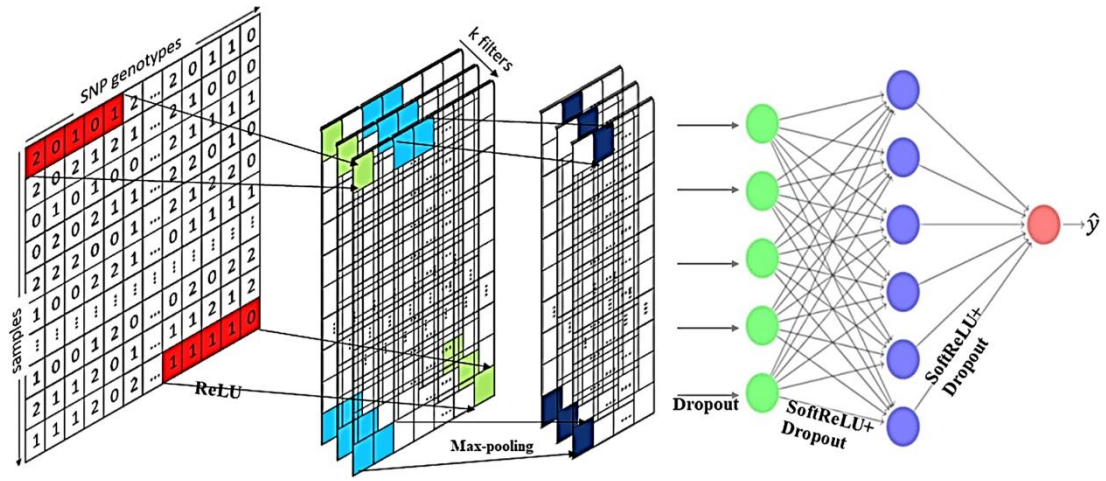


شکل ۱– شبکه عصبی پرسپترون چند لایه (MLP) (Abdollahi-Arpanahi *et al.*, 2020)؛ برگرفته شده از:
(/http://www.texample.net/tikz/examples/neural-network

MLPها شبکه‌های عصبی تقریباً ساده‌ای هستند که امروزه به دلیل سادگی کار بیشتر مورد استفاده قرار می‌گیرند. در این شبکه کدهای SNPها که به صورت سیستم کدینگ ۰-۱-۲ هستند به عنوان ورودی به شبکه معرفی می‌شوند و از یکسری لایه پنهان که با چینش مشخص است، عبور داده می‌شوند. همزمان با عبور از این لایه‌ها (قبل و بعد از Hidden layers)، در ضرایبی که تعریف شده‌اند ضرب می‌شوند. در این مرحله توابعی مانند توابع سیگمایی تعریف می‌شوند که اعداد بین صفر و یک را به یک ارتقا می‌دهند. آن چه که به عنوان خروجی خواهید داشت، ارزش اصلاحی حیوان که اثر مجموع SNPها است. در این شبکه مصنوعی بالطبع هر چه تعداد لایه‌ها بیشتر باشد، قدرت شبکه بیشتر خواهد بود، امّا این میزان قدرت به اندازهٔ شبکه عصبی CNN نخواهد بود (Abdollahi-Arpanahi *et al.*, 2020).

شبکه‌های عصبی CNN، بیشتر برای برآورد ارزش اصلاحی صفت‌هایی که وراثت‌پذیری و معماری پیچیده‌ای دارند، مورد استفاده قرار می‌گیرند. در این نوع از شبکه ابتدا اطلاعات ژنوتیپی (SNP Genotype) به عنوان ورودی به شبکه داده می‌شود. سپس

این اطلاعات از یکسری فیلترهای بخصوص (بر مبنای مباحث پیچیده ریاضی مانند تبدیل فوریه) که با ترتیب خاصی متناسب با هدف مورد نظر چینش شده‌اند، عبور داده می‌شوند. در نهایت اطلاعات خروجی از فیلترها به لایه‌های مختلف مانند شبکه MLP انتقال داده می‌شوند؛ همزمان نیز در ضرایب مشخصی ضرب می‌شوند. این شبکه عصبی تا مرحله‌ای که مقدار خطای موجود بخصوص (خروجی مورد نظر و خروجی پیشبینی شده) به حداقل مقدار ممکن کاهش پیدا کند (شکل ۲). فیلترها هر کدام کارآیی خاص خود را دارند، به عنوان مثال با توجه به توابعی که برای آن‌ها تعریف شده است، ممکن است اعدادی (SNPs) را حذف کنند. بنابراین با عبور اطلاعات اولیه حجم اطلاعات کاهش می‌یابد؛ به گونه‌ای که تنها اطلاعاتی (SNPs) که نقش مهمی در صفت مورد نظر دارند باقی می‌مانند. میزان اطلاعات آموزشی برای شبکه عصبی عمیق باید به طور قابل توجهی بالا باشد، زیرا این نوع از الگوریتم ها خاصیت بسیار بالایی برای حفظ و نه یادگیری اطلاعات دارند (Abdollahi-Arpanahi *et al.*, 2020).



شکل ۲- شبکه عصبی همگشتی (CNN) (Abdollahi-Arpanahi *et al.*, 2020)

جنگل تصادفی (Random Forest)

جنگل تصادفی یکی از روش‌های یادگیری ماشین است که در عرصه‌های مختلف علوم به طور موفقیت‌آمیز مورد استفاده قرار گرفته است. همان‌طور که از اسم این الگوریتم بر می‌آید، در این روش مجموعه یا جنگلی از درختان مورد استفاده قرار می‌گیرند؛ به گونه‌ای که هر درخت از ریشه، گره‌ها و برگ‌ها تشکیل شده است.

جنگل تصادفی یک روش غیرپارامتریک باز نمونه‌گیری است که برخلاف روش‌های بیزی نیازمند فرضیات کمتری درباره توزیع داده‌ها است (Goldstein *et al.*, 2010). این الگوریتم درختانِ زیادی در نمونه‌های بوت استرپ ایجاد می‌کند و از طریق میانگین هر درخت، پیش‌بینی‌های نهایی را محاسبه می‌کند. به طور کلی جنگل تصادفی از طریق استراتژی بگینگ، خطای پیش‌بینی را کاهش و با استفاده از انتخاب ویژگی گزینش متغیر تصادفی جهت ایجاد هر درخت را ایجاد می‌کند (Breiman, 2001). برای اطلاعات با حجم و ابعاد بسیار بالا مانند اطلاعات حاصل از مطالعات پویش ژنومی (GWAS) یک مدل ساده نمی‌تواند پیچیدگی‌های موجود در اطلاعات را پوشش دهد. یک استراتژی برای برطرف نمودن این نقص، استفاده از جنگل یا تجمعی از درخت‌ها است که به روش جنگل تصادفی مشهور است. پارامترهای کلیدی برای مدل جنگل تصادفی، تعداد درختان و تعداد متغیرهای پیش‌گو می‌باشند (Breiman, 2001).

سه پارامتر مهمی که در جنگل تصادفی در مورد کلاسه‌بندی بایستی تنظیم شوند شامل mtry (تعداد SNP یا کوواریت‌های نمونه برداری شده در هر بار نمونه‌گیری تصادفی)، ntree (تعداد بوت استرپ و یا تعداد درختانی است که بایستی رشد کنند و معیاری برای انتخاب بهترین SNP برای تقسیم شدن هر گره است) و Node (تعداد گره یا وزن‌دهی است که نشان‌دهندهٔ تعداد مشاهدات در هر خوشه درخت می‌باشد) است (Naderi, 2018).

به طور کلی جنگل تصادفی، تجمعی از درختان است 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، تحت آموزش قرار می‌گیرد و بر جمعیت تأیید یا کاندید (حیوانات کاندیدای انتخاب) اعمال

می‌شود. هر یک از n نمونه، وارد هر گره از هر درخت می‌شود و از این نمونه (اطلاعات یک SNP) برای تقسیم‌بندی حیوانات استفاده می‌شود؛ به گونه‌ای که حیوانات براساس اطلاعات ژنوتیپی خود برای SNP انتخاب شده دسته‌بندی می‌شوند. این کار در گره‌های متوالی انجام می‌شود تا در نهایت به برگ‌ها و یا همان گره‌های پایانی می‌رسد که در آن‌ها حداکثر یکنواختی وجود خواهد داشت (حیوانات دارای اطلاعات فنوتیپی با ژنوتیپ‌های مشابه برای SNPهای مختلف در یک گره پایانی تجمع می‌یابند) (Ghafouri-Kesbi *et al.*, 2016). در جمعیت تأیید یا کاندیدا، پیش‌بینی جنگل تصادفی برای یک ورودی جدید (حیوان دارای اطلاعات ژنوتیپی xi اما فاقد اطلاعات فنوتیپی yi) با توجه به رابطه (۱) می‌گیرد.

$$\bar{f}_{rf}^{B}(x) - \frac{1}{B}\sum_{b=1}^{B} T\left(x, \Psi_b\right) \qquad \text{رابطه (۱)}$$

در رابطه (۱) Ψb، bاُمین 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 مقادیر مناسب آن‌ها تعیین شود. 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).

در هر بار نمونه‌گیری با جایگزینی اطلاعات، برخی اطلاعات (SNP) هرگز نمونه‌گیری نمی‌شوند و برای برخی دیگر شاید چند بار نمونه‌گیری صورت گیرد. به عبارت دیگر برخی داده‌های ورودی برای تعدادی از درخت‌ها در اصطلاح "نمونه خارج از کیسه" (Out of Bag) خواهند بود؛ یعنی در ایجاد برخی درخت‌ها مشارکت نخواهند داشت. این داده‌ها عملکرد یک اعتبارسنج داخلی برای هر درخت را خواهند داشت که این اعتبارسنجی از طریق برآورد خطای خارج از کیسه انجام می‌شود. اگر 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 می‌شوند که نشان‌دهندهٔ میزان تأثیر نمونه‌های انتخاب نشده بر میزان خطای نتیجهٔ نهایی جنگل تصادفی است.

ماشین بردار پشتیبان (Support Vector Machine (SVM))

از جمله روش‌های یادگیری ماشین می‌توان "ماشین بردار پشتیبان" را نام برد که علاوه بر صحت بالا در پیش‌بینی ژنومی (Ghafouri-Kesbi *et al.*, 2017)، در تشخیص ژن- ژن، پروتئین- پروتئین، اثر متقابل ژن- محیط، ژن‌های مرتبط با بیماری، مدل‌سازی جهت ارتباط میان ترکیب نشانگرها، انتخاب ژن‌های در ارتباط با صفت هدف و شناسایی فاکتورهای تنظیمی در توالی آمینواسیدها نقش برجسته‌ای دارند (Yang et al., 2010).

تحقیقات در مورد استفاده از روش‌های یادگیری ماشین در ارزیابی ژنومی صفاتِ گسسته، نشان از برتری این روش‌ها در مقایسه با روش‌های بیز داشت (González-Recio and Forni, 2011). همچنین مطالعات بر روی صفات پیوسته از قدرت بالای ارزیابی ژنومی ماشین بردار پشتیبان و دیگر روش‌های یادگیری ماشین در مقایسه با بهترین پیش‌بینی ناریب خطی با استفاده از رگرسیون ریج (Ridge Regression) (Ogutu *et al.*, 2011) و بهترین پیش بینی ناریب خطی ژنومی (Ghafouri-Kesbi *et al.*, 2017) را گزارش داده‌اند. با این حال تفاوت عمدهٔ روش ماشین بردار پشتیبان در مقایسه با روش‌های مرسوم از جمله بهترین پیش‌بینی ناریب خطی ژنومی و بیز لاسو (The Bayesian Lasso)، به عدم نیاز به نحوهٔ توارث جمعیت، توانایی بالا در به کارگیری اثرات غیرافزایشی، فرضیات در نظر گرفته شده برای مدل ژنتیکی پشت صحنه آن‌ها و تنظیم و بهینه‌سازی پارامترهای آن‌ها جهت دستیابی به حداکثر صحت پیش‌بینی ژنومی می‌باشد (Ghafouri-Kesbi *et al.*, 2017).

در واقع روش ماشین بردار یک الگوریتم ماشینی است که از طریق اطلاعات آموزشی به دسته‌بندی عوامل و تشخیص و تمایز الگوهای پیچیده در داده‌ها می‌پردازد (Boser *et al.*, 1992). این الگوریتم یک روش رایج در کلاسه‌بندی پروفایل‌های بیان ژن حاصل از ریزآرایه‌ها و مسائل رگرسیون غیرخطی و دو کلاسه کاربرد فراوانی دارد.

بوستینگ (Boosting)

بوستینگ یک روش باز نمونه‌گیری، از نوع غیرپارامتریک و به عنوان یکی از قدرتمندترین روش‌های یادگیری ماشین جهت بهبود عملکرد روش‌های طبقه‌بندی توسعه داده شد (Freund and Schapire, 1997). بعدها از آن برای آنالیز ژنومی صفات آستانه‌ای (González-Recio and Forni, 2011) به دلیل توانایی بالای آن در آنالیز حجم زیاد داده‌های ژنومی، توانایی در تشخیص اثرات متقابل ژن- ژن، ژن- محیط، قدرت بالای مدل‌سازی و ارتباط بین ترکیبات مختلف نشانگرها، تشخیص ژن‌های مرتبط

با بیماری، مشکلات مربوط به نداشتن نمونه کافی و بهبود صحت پیش بینی ژنومی استفاده شد (Yang *et al.*, 2010).

در الگوریتم Boosting توابع پایه مورد استفاده شامل یاد-گیرنده‌های ضعیف مانند درخت رگرسیونی می‌باشند. در این الگوریتم سعی می‌شود تعدادی یادگیر پایه ضعیف (یاد-گیرنده‌های بهتر از حالت تصادفی) که مکمل همدیگر هستند، تولید شود و از طریق آموزش با استفاده از یادگیرنده‌های قبلی، یادگیرنده‌های جدید قوی‌تری ایجاد شود. در این الگوریتم توابع پایه مانند درختان رگرسیونی به صورت سریالی هر یک روی باقی‌مانده درخت قبلی اضافه می‌شوند؛ در نتیجه عدم دسته‌بندی اشتباه در درخت قبلی باعث کاهش مقدار خطا در درخت بعدی می‌شود (Ghafouri-Kesbi *et al.*, 2017). این الگوریتم تا زمانی ادامه می‌یابد که خطای آخرین درخت به حداقل مقدار ممکن برسد. مدل کلی الگوریتم Boosting به صورت رابطه (۲) است.

$$f(x) = \sum_{m=1}^{m} \beta_m b(x; \gamma_m) \quad \text{رابطه (۲)}$$

در روش Boosting، پارامترهای بهینه‌سازی شامل تعداد درخت (ntree)، عمق درخت (Complexity tree) و پارامتر انقباضی یا نرخ یادگیری (Rate learning) می‌باشند.

کاربرد روش‌های یادگیری ماشین در ارزیابی‌های ژنومی

Long و همکاران (۲۰۰۷) یک روش دو مرحله‌ای مبتنی بر فیلترکردن و کلاسه‌بندی را برای انتخاب SNPهای مؤثر بر مرگ‌ومیر جوجه‌های گوشتی یک لاین تجاری توسعه دادند. ایده و روش کار که از مقاله Hoh و همکاران (۲۰۰۰) اخذ شد، این بود که از مجموع چند هزار SNP فقط برخی از آن‌ها به طور معنی‌دار فنوتیپ را تحت تأثیر قرار می‌دهند و حضور سایر SNPهای بی‌اثر یا با اثر کم، فقط منجر به پیچیده‌شدن محاسبات و بروز برخی مشکلات دیگر مانند پاسخ‌های دروغین مثبت (False Positive) خواهند شد.

بر این اساس Long و همکاران (۲۰۰۷) اطلاعات ژنوتیپی ۵۵۲۳ SNP مربوط به ۲۳۱ خروس به همراه اطلاعات مرتبط با میزان مرگ‌ومیر فرزندان را برای مشخص نمودن SNPهای مؤثر بر نرخ مرگ‌ومیر جوجه‌ها آنالیز کردند. ابتدا SNPها فیلتر شده و براساس میزان تأثیر بر واریانس بین نرها رتبه‌بندی شدند. در مرحلهٔ بعد با استفاده از یک روش هوشمند کلاسه‌بندی تحت عنوان Naïve-Bayes، زیر مجموعه‌هایی از SNPهای فیلترشده استخراج گردیدند (کلاسه‌بندی شدند) و در نهایت با مقایسه زیرمجموعه‌های مختلف، ۱۷ SNP انتخاب شد که با هم حدود ۳۱ درصد از واریانس مشاهده شده بین خانواده‌ها را توجیه می‌کردند (Long *et al.*, 2007).

در مطالعه‌ای ارتباطات و ساختار ژنتیکی هشت نژاد از اسب‌های جمهوری چک را با استفاده از اطلاعات ژنوتیپی ۱۷ نشانگر میکروساتلایت مورد استفاده در آزمون انساب، جهت مقایسه الگوریتم‌های مختلف کلاسه‌بندی شامل دو روش مبتنی بر الگوریتم‌های کلاسه‌بندی احتمالی (Naïve-Bayes و Bayes-Net)، دو روش مبتنی بر درخت تصمیم‌گیری (JRip , J4s) و دو روش Instance-Base (JRip , J48) آنالیز شد و در نهایت گزارش گردید که دو روش بیزی با صحت تقریباً برابر ۸۷ درصد در کلاسه‌بندی کردن اسب‌ها در نژاد مربوطه آن‌ها موفق‌تر بودند (Burocziova and Riha, 2009).

Naderi و همکاران (۲۰۱۶) یک مطالعه شبیه‌سازی شده جهت بررسی عملکرد RF و BLUP ژنومی (GBLUP) برای پیش‌بینی ژنومی با استفاده از صفات بیماری باینری مبتنی بر گروه‌های کالیبراسیون گاو انجام دادند. آن‌ها دقت پیش‌بینی ژنومی از طریق تغییر وراثت‌پذیری، تعداد QTL، تراکم نشانگر، ساختار LD از جمعیت تعین ژنوتایپ شده و شیوع گاوهای بیمار در جمعیت آموزش را مقایسه کردند. آن‌ها همچنین تخمین‌های RF را در مورد اثرات و مکان‌های مهمترین SNPها با QTL واقعی بررسی کردند (Naderi *et al.*, 2016). درنهایت نتیجه گرفتند که دقت پیش‌بینی در هنگام استفاده از روش GBLUP بیشتر است و کاهش میزان وراثت‌پذیری و تعداد QTLها با کاهش دقت پیش‌بینی برای تمام سناریوها در جایی که برای روش RF برجسته‌تر است، همراه بود. روشِ RF، تنها در شرایطی بهتر از روش GBLUP انجام می‌شود که بالاترین وراثت‌پذیری، اسنیپ‌چیپ‌های متراکم و بیشترین تعداد QTL برای آنالیزها استفاده شود. علاوه بر این، روش RF می‌تواند با موفقیت SNPهای مهم را در مجاورت یک QTL یا یک ژن کاندیدا شناسایی کند (Naderi *et al.*, 2016).

Li و همکاران (۲۰۱۸a) از سه رویکرد یادگیری ماشین (RF، GBM، XgBoost)، ۳۸۰۸۲ نشانگر SNP و فنوتیپ‌های وزن بدن ۲۰۹۳ گاو برهمن (۱۰۹۷ گاو به عنوان جمعیت مرجع و ۹۹۶ گاو به عنوان جمعیت کاندیدا) برای شناسایی زیر مجموعه‌های SNPها جهت ساخت ماتریس‌های مربوط به روابط ژنومی (GRMs) برای برآورد ارزش اصلاحی ژنومی (GEBV)

استفاده کردند. از میان سه روش ذکر شده، روش GBM بهترین عملکرد را داشت و رتبه‌های دیگر از نظر عملکرد مربوط به RF و XgBoost بود (Li *et al.*, 2018a). همچنین در مطالعهٔ Li و همکاران (۲۰۱۸b)، از روش جنگل تصادفی به عنوان ابزاری برای شناسایی زیر مجموعه‌های SNP جهت پیش‌بینی ژنومی کل ارزش‌های ژنتیکی وزن سالانه در گاوهای گوشتی استفاده شد. هدف از مطالعه آن‌ها بررسی تأثیر عوامل غیرافزایشی ناشناخته (به عنوان مثال اثرات اپیستازی SNPها) در PAC از ارزش‌های ژنتیکی کل با استفاده از روش‌های یادگیری ماشین بود (Li *et al.*, 2018b).

در مطالعهٔ Nayeri و همکاران (۲۰۱۹) شرح مختصری از نحوه استفاده از روش‌های یادگیری ماشین در رشته‌های مختلف دامپروری ارائه شده است. آن‌ها مطالعه خود را در شش حوزهٔ کاربردی مرتبط با علوم دامی یعنی بهداشت، پرورش، باروری، مرگ و میر، تغذیه و اصلاح‌نژاد گروه‌بندی کردند. به طور خلاصه، این مطالعه انتقال رویکرد در صنعت دامپروری از استراتژی‌های پیش‌بینی سنتی مانند GBLUP تک مرحله‌ای، MAS و GWAS گرفته تا رویکردهای پیشرفته‌تر یادگیری ماشین مانند ANN (Artificial Neural Network)، DL (Deep Learning)، BN (Bayesian Network) را نشان می‌دهد. همچنین اشاره شده است که اتخاذ روش‌های یادگیری ماشین در تحقیقات دامپروری باید با اتخاذ معیارهای مربوطه و در صورت لزوم معرفی معیارهای ارزشیابی جدید که کیفیت نتایج را بهبود می‌بخشند، همراه باشد (Nayeri *et al.*, 2019).

عوامل مؤثر بر صحت ارزش اصلاحی ژنومی

علاوه بر تأثیر مدل‌های آماری و نوع صفت مورد مطالعه، عوامل مختلفی می‌توانند صحت ارزش‌های اصلاحی ژنومی و ارزیابی‌های ژنومی را تحت تأثیر قرار دهند. این عوامل شامل تعداد QTL (González-Recio and Forni, 2011)، توزیع اثرات QTL (Ghafouri-Kesbi *et al.*, 2017)، مقدار عدم تعادل پیوستگی (Yin *et al.*, 2014)، تراکم نشانگرها (Badke *et al.*, 2014)، وراثت‌پذیری، تعداد داده‌های فنوتیپی در جمعیت مرجع (Meuwissen *et al.*, 2001) و فاصله زمانی (تعداد نسل) بین جمعیت مرجع و جمعیت تأیید (Gorgani Firozjah *et al.*, 2014) می‌باشند. با توجه به این که در انتخاب ژنومی، ارزش‌های اصلاحی حیوانات جمعیت کاندیدا (دارای ژنوتیپ) از طریق برآورد میزان اثر هر کدام از نشانگرها بر صفت در جمعیت مرجع (دارای ژنوتیپ و فنوتیپ) برآورد می‌شود، برای صفات آستانه‌ای علاوه بر عوامل فوق‌الذکر، نسبت فنوتیپی جمعیت مرجع (نسبت شیوع بیماری) یکی از عوامل تأثیرگذار در برآورد ارزش‌های اصلاحی حیوانات جمعیت کاندیدا است (Pimentel *et al.*, 2013). تحقیقات در این زمینه نشان داد که نسبت شیوع بیماری در جمعیت مرجع از عوامل تأثیرگذار بر صحت و قابلیت اطمینان ژنومی می‌باشد (Naderi *et al.*, 2016). همچنین در تحقیقات دیگر افزایش نسبت رکوردهای فنوتیپی حیوانات ماده نسبت به نرها در جمعیت مرجع از عوامل مؤثر بر صحت ژنومی عنوان شده است (Buch *et al.*, 2012).

Ghafouri-Kesbi و همکاران (۲۰۱۷)، در یک مطالعه شبیه‌سازی شده نشان دادند که روش جنگل تصادفی عملکرد بهتری در تعداد بالای QTL (۱۰۰۰) نسبت به تعداد پایین QTL (۱۰۰) برای صفات با وراثت‌پذیری پایین دارد. همچنین در مطالعه‌ای دیگر شبیه‌سازی اثر مثبت وراثت‌پذیری بر صحت پیش‌بینی ژنومی روش‌های Boosting و جنگل تصادفی صورت گرفت، در ادامه اثبات شد که افزایش وراثت‌پذیری از ۰/۱ به ۰/۵، افزایش ۷۲/۴ و ۷۵/۵ درصدی صحت در این روش‌ها را به دنبال خواهد داشت (Ghafouri-Kesbi *et al.*, 2017).

نتیجه‌گیری کلی

در این مطالعه به رویکردهای پیشرفته برآورد ارزش‌های اصلاحی ژنومی به ویژه یادگیری ماشین و الگوریتم‌های زیرمجموعهٔ آن (یادگیری عمیق، جنگل تصادفی، ماشین بردار پشتیبان و بوستینگ) پرداخته شد. روش‌های مورد استفاده در اصلاح‌نژاد دام و طیور باید با اتخاذ و انتخاب دقیق معیارهای مناسب و همگن برای برآورد کیفیت نتایج پیش‌بینی همراه باشد. توسعهٔ تکنیک‌ها و رویکردهای پیشرفته متناسب با افزایش ثبت اطلاعات فنوتیپی و ژنوتیپی و به دنبال آن پیدایش اَبر داده‌ها در عصر جدید فرصت‌های زیادی را فراهم ساخته است تا با کمک الگوریتم‌های خاص و رایانه‌های پیشرفته، تجزیه و تحلیل مجموعه‌های پیچیده انجام شود؛ این مسئله مطمئناً تأثیر قابل توجهی در آینده اصلاح‌نژاد دام و طیور، به ویژه پیشرفت و بهبود ژنتیکی دام‌ها خواهد داشت.

سپاسگزاری

بدینوسیله از زحمات جناب آقای دکتر رسول واعظ ترشیزی، عضو هیئت علمی گروه علوم دامی دانشگاه تربیت مدرس که ایدهٔ نوشتن مقاله را مطرح و ما را تشویق به نوشتن آن کردند، بسیار تشکر و قدردانی می‌شود.

منابع

Abdollahi-Arpanahi, R., Gianola, D., and Peñagaricano, F. (2020). "Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes." *Genetics Selection Evolution*, 52(1): 1-15.

Badke, Y.M., Bates, R.O., Ernst, C.W., Fix, J. and Steibel, J.P. (2014). "Accuracy of estimation of genomic breeding values in pigs using low-density genotypes and imputation." *Genes Genomes Genetics*, 4(4): 623-631.

Bellot, P., de Los Campos, G., and Pérez-Enciso, M. (2018). "Can deep learning improve genomic prediction of complex human traits?" *Genetics*, 210(3): 809-819.

Bishop, C.M. (2006). "Pattern recognition and machine learning." Springer, Vol. 1, New York.

Boser, B.E., Guyon, I.M. and Vapnik, V.N. (1992). "A training algorithm for optimal margin classifiers. Proceedings of the fifth annual workshop on computational learning theory." *Association for Computing Machinery*, 144-152.

Boser, B., Guyon, I. and Vapnik, V. (1992). "A training algorithm for optimal margin classifiers." In 'Proceedings of the fifth annual workshop on computational learning theory. Pittsburgh (USA). 27–29.

Breiman, L. (2001). "Random forests." *Machine Learning*, 45: 5-32.

Buch, L.H., Kargo, M., Berg, P., Lassen, J., and Sørensen, A.C. (2012). "The value of cows in reference populations for genomic selection of new functional traits." *Animal*, 6(6): 880-886.

Burocziova, M. and Riha, J. (2009). "Horse breed discrimination using machine learning methods." *Journal of Applied Genetics*, 50(4): 375-77.

Freund, Y. and Schapire, R.E. (1997). "A decision-theoretic generalization of on-line learning and an application to boosting." *Journal of Computer and System Sciences*, 55(1): 119–139.

Ghafouri, F., Mehrabani Yeganeh, H., Mohamadian Jeshvaghani, S. (2020). "Big data and the role of high-throughput technologies in livestock and poultry breeding." *Professional Journal of Domestic*, 20(1): 34-40.

Ghafouri-Kesbi, F., Rahimi-Mianji, G., Honarvar, M. and Nejati-Javaremi, A. (2016) Tuning and application of random forest algorithm in genomic evaluation. *Research on Animal Production*, 7 (13): 178-185 (In Persian).

Ghafouri-Kesbi, F., Rahimi-Mianji, G., Honarvar, M. and Nejati-Javaremi, A. (2017). "Predictive ability of Random Forests, Boosting, Support Vector Machines and Genomic Best Linear Unbiased Prediction in different scenarios of genomic evaluation." *Journal of Animal Production Science*, 57(2): 229-36.

Goldstein, B.A., Hubbard, A.E., Cutler, A. and Barcellos, L.F. (2010). "An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings." *Journal of BMC Genetics*, 11(1): 49.

González-Recio, O. and Forni, S. (2011). "Genome-wide prediction of discrete traits using Bayesian regressions and machine learning." *Journal of Genetics Selection Evolution*, 43(1): 7.

González-Recio, O., Jiménez-Montero, J.A. and Alenda, R. (2013). "The gradient boosting algorithm and random boosting for genome-assisted evaluation in large data sets." *Journal of Dairy Science*, 96: 614–624.

Gorgani Firozjah, N., Atashi, H., Dadpasand, M. and Zamiri, M. (2014). "Effect of marker density and trait heritability on the accuracy of genomic prediction over three generations." *Journal of Livestock Science and Technologies*, 2(2): 53-58.

Heslot, N., Yang, H.P., Sorrells, M.E. and Jannink, J.L. (2012). "Genomic selection in plant breeding: a comparison of models." *Crop Science*, 52: 146-160.

Hofer, A. (1998). "Variance component estimation in animal breeding: a review." *Journal of Animal Breeding and Genetics*, 115(1-6), 247-265.

Hoh, J., Wille, A., Zee, R., Cheng S., Reynolds R., and et al. (2000). "Selecting SNPs in two-stage analysis of disease association data: a model-free approach." *Ann Hum Genet*, 64: 413– 417.

Li, B., Zhang, N., Wang, Y.G., George, A.W., Reverter, A. and et al. (2018a). "Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods." *Frontiers in Genetics*, 9: 237–256.

genomic selection." BMC proceedings. *BioMed Central*, 5(3): 11.

Pérez-Enciso, M., and Zingaretti, L.M. (2019). "A guide on deep learning for complex trait genomic prediction." *Genes*, 10(7), 553.

Pimentel, E.C., Wensch-Dorendorf, M., König, S., and Swalve, H.H. (2013). "Enlarging a training set for genomic selection by imputation of un-genotyped animals in populations of varying genetic architecture." *Genetics Selection Evolution*, 45(1): 12.

Yang, P., Hwa Yang, Y., Zhou, B.B. and Zomaya, Y A. (2010). "A review of ensemble methods in bioinformatics." *Current Bioinformatics*, 5(4): 296-308.

Yin, T., Pimentel, E. Borstel, U.K.V. and Konig, S. (2014). "Strategy for the simulation and analysis of longitudinal phenotypic and genomic data in the context of a temperature× humidity-dependent covariate." *Journal of Dairy Science*, 97(4): 2444-2454.

Publisher Note
Animal Science Students Scientific Association, Campus of Agriculture and Natural Resources at the University of Tehran

Submit Your Manuscript:
https://domesticsj.ut.ac.ir/contacts?_action=loginForm

Li, Y., Raidan, F.S.S., Li, B., Vitezica, Z.G. and Reverter, A. (2018b). "Using Random Forests as a prescreening tool for genomic prediction: impact of subsets of SNPs on prediction accuracy of total genetic values." Proceedings of the 11th World Congress on Genetics Applied to Livestock Production (WCGALP). 248.

Long, N., Gianola, D., Rosa, G.J.M., Weigel, K.A. and Avendaño, S. (2007). "Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers." *Journal of Animal Breeding and Genetics*, 124: 377–389.

Meuwissen, T.H., Hayes, B.J. and Goddard, M.E. (2001). "Prediction of total genetic value using genome-wide dense marker maps." *Genetics*, 157:1819–1829.

Mitchell, T.M. (1997). "Machine learning." Boston, McGraw-Hill.

Naderi, S., Yin, T. and König, S. (2016). "Random forest estimation of genomic breeding values for disease susceptibility over different disease incidences and genomic architectures in simulated cow calibration groups." *Journal of Dairy Science*, 99(9): 7261-7273.

Naderi, Y. (2018). "Evaluation of genomic prediction accuracy in different genomic architectures of quantitative and threshold traits with the imputation of simulated genomic data using random forest method." *Research on Animal Production*, 9(20): 129-138 (In Persian).

Nayeri, S., Sargolzaei, M. and Tulpan, D. (2019). "A review of traditional and machine learning methods applied to animal breeding." *Animal Health Research Reviews*, 20: 31-46.

Nejati-Javaremi, A., Smith, C. and Gibson, J. (1997). "Effect of total allelic relationship on accuracy of evaluation and response to selection." *Journal of Animal Science*, 75: 1738-1745.

Nilsson, N.J. (1998). "Introduction to Machine Learning." Stanford University. Stanford, USA. 412.

Ogutu, J.O., Piepho, H.P. and SchulzStreeck, T. (2011). "A comparison of random forests, boosting and support vector machines for

Domestic; 20 (2): 19-29

Scientific-Extensional (Professional) Journal

Domestic

Animal Science Students Scientific Association at the University of Tehran, Autumn 2020

https://domesticsj.ut.ac.ir/article_79151.html

Scientific-Extensional

Application of machine learning approach and its subset algorithms in estimating genomic breeding values

Farzad Ghafouri[1], Somayeh Alipour[2*] and Sadegh Mohamadian Jeshvaghani[3]

[1] Ph.D. Student of Animal and Poultry Breeding & Genetics, Department of Animal Science, College of Agriculture and Natural Resources, University of Tehran, Karaj, Iran

[2] Ph.D. Student of Animal Breeding and Genetics, Department of Animal Science, Faculty of Agriculture, University of Tarbiat Modares, Tehran, Iran

[3] M.Sc. of Animal Breeding and Genetics, Department of Animal Science, College of Agriculture and Natural Resources, University of Tehran, Karaj, Iran

https://doi.org/10.22059/domesticsj.2020.310252.1050

Abstract

Genomic selection strives to make use of genotypic and phenotypic data, simultaneously, in order to evaluate animals genetically in a short period of time to opt superior ones. The development of data mining algorithms related to big data analysis in the digital era makes a great contribution to estimating breeding values in livestock and poultry breeding. Recently, machine learning procedures and their sub-algorithms such as Deep Learning (DL), Random Forest (RF), Support Vector Machine (SVM), and boosting, which are categorized as non-parametric animal evaluation methods, have been introduced to the realm of genomic selection. Machine learning algorithms not only provide breeders with much more potential and efficiency but also they are more adapted with big data. These algorithms enable breeders to estimate non-additive effects such as dominance and epistasis, as well as studying of complex relationships between variables (such as marker interactions). The punch line of these algorithms is to use training data (here the genotypic and phenotypic information of the animals in reference population) to predict their genomic breeding values based on the genotypic information of the candidate population. Some of these methods have been used successfully in animal genomic evaluations and they have provided acceptable results with low error. In fact, the purpose of this study is to define machine learning approaches and their sub-algorithms besides their role in predicting the genetic architecture of traits with complex heritability. As a result, it is likely that using machine learning approach to find the most efficient algorithm, along with increasing the volume of phenotypic and genomic data, will have a significant impact on the future of livestock and poultry breeding.

Keyword(s): Machine learning, Genetic algorithm, Deep learning, Random forest, Breeding value, Non-parametric methods

***Corresponding Author E-mail:** s_alipour@modares.ac.ir

Received: 17 Sep 2020 **Revised:** 24 Sep 2020 **Accepted:** 07 Oct 2020 **Published online:** 10 Dec 2020

Citation: Ghafouri, F., Alipour, S., Mohamadian Jeshvaghani, S. Application of machine learning approach and its subset algorithms in estimating genomic breeding values. *Professional Journal of Domestic*, 2020; 20(2): 19-29.