نشریه پژوهش در نشخوارکنندگان
جلد هفتم، شماره چهارم، ۱۳۹۸
http://ejrr.gau.ac.ir
۱۷-۳۲

ارزیابی ژنومی روش ماشین بردار پشتیبان و روش‌های رایج پیش‌بینی ژنومی در بروز متفاوت فنوتیپ آستانه‌ای مطالعه شبیه‌سازی

یوسف نادری

استادیار، گروه علوم دامی، باشگاه پژوهشگران جوان و نخبگان، دانشگاه آزاد اسلامی، واحد آستارا، آستارا، ایران

تاریخ دریافت: ۹۸/۰۶/۱٤؛ تاریخ پذیرش: ۹۸/۰۹/۰٦

چکیده

سابقه و هدف: بسیاری از صفات برجسته در دام‌های اهلی شامل: مقاومت به بیماری‌ها و سختی زایش مشمول یک توزیع طبقه بندی از فنوتیپ هستند. این صفات به علت اهمیت در آسایش حیوان و گرایشات انسانی به تولیدات با کیفیت بالا و سالم از اهمیت ویژه‌ای در اصلاح دام برخوردار می‌باشند. بنابراین شناسایی و تشخیص واریانت های ژنتیکی موثر بر صفات آستانه‌ای اعم از مقاومت به بیماری یکی از اهداف اصلی در ژنتیک حیوانی است. در این رأستا گزینش ژنومی می‌تواند نقش مهمی در افزایش پیشرفت ژنتیکی صفات آستانه‌ای ایفا کند. هدف از تحقیق حاضر، ارزیابی سطح زیر منحنی مشخصه عملکرد ژنومی روش‌های ماشین بردار پشتیبان، بهترین پیشبینی ناریب خطی ژنومی و بیز لاسو برای نرخ مختلف توزیع فنوتیپ دودویی در جمعیت مرجع بود.

مواد و روش‌ها: یک جمعیت پایه ۱۰۰۰ رأسی برای ۱۰۰۰ نسل با استفاده از نرم افزار QMSim شبیه‌سازی شد. جمعیت‌های ژنومی برای سطوح مختلف وراثت‌پذیری (۰/۰۵ و ۰/۲)، عدم تعادل پیوستگی (۰/۲۲۱ و ۰/٤۳۵) و تعداد متفاوت جایگاه صفات کمی (۱۰۰ و ۱۰۰۰) بر روی ۲۹ کروموزوم شبیه‌سازی شدند. جهت ایجاد نسبت‌های مختلف فنوتیپ آستانه‌ای دودویی، فنوتیپ افراد جمعیت مرجع وابسته به این که باقی‌مانده آنها کمتر از میانگین باقی‌مانده ($\bar{e}$)، $\bar{e}-1SD_e$ یا $\bar{e}+1SD_e$ باشد کد یک (فنوتیپ نامطلوب) و سایر حیوانات کد صفر (فنوتیپ مطلوب) اختصاص داده شدند. جهت آنالیز داده‌های شبیه‌سازی شده سه مدل آماری شامل: ماشین بردار پشتیبان، بهترین پیشبینی ناریب خطی ژنومی و بیز لاسو به کار گرفته شد.

نتایج: بهترین نرخ آستانه جمعیت مرجع هنگامی بود که فنوتیپ نامطلوب این مجموعه نسبتی نزدیک به شرایط واقعی داشت ($\bar{e}-1SD_e$) و منجر به ایجاد بیشترین سطح زیر منحنی مشخصه عملکرد در روش‌های ماشین بردار پشتیبان، بیز لاسو و بهترین پیشبینی ناریب خطی ژنومی شد. بیشترین (۰/۸۱۳) و کمترین (۰/۵۲۱) میزان سطح زیر منحنی مشخصه عملکرد برای روش ماشین بردار پشتیبان مشاهده شد. به طور کلی وراثت پذیری صفت عاملی مؤثر بر سطح زیر منحنی مشخصه عملکرد ژنومی روش‌های آماری ماشین بردار پشتیبان، بیز لاسو و بهترین پیشبینی ناریب خطی ژنومی بود. به‌طوری که با افزایش وراثت پذیری سطح زیر منحنی مشخصه عملکرد ژنومی در هر سه روش آماری افزایش یافت. میانگین عدم تعادل پیوستگی برای

*نویسنده مسئول: yousefnaderi@gmail.com

جمعیت‌های با عدم تعادل پیوستگی پایین و بالا در فاصله ۰/۰۵ سانتی مورگان به ترتیب ۰/۲۲۱ و ۰/٤۳۵ بود و سطح زیر منحنی مشخصه عملکرد ناشی از روش‌های بهترین پیشبینی ناریب خطی ژنومی، بیز لاسو و ماشین بردار پشتیبان با افزایش سطح عدم تعادل پیوستگی افزایش یافت. نتایج این تحقیق نشان داد که سطح بالای عدم تعادل پیوستگی میان جایگاه‌های صفت کمی و نشانگرها، باعث افزایش احتمال نمونه‌گیری نشانگرهای مجاور در روش‌های باز نمونه‌گیری می‌شود، که این امر عملکرد مثبت ماشین بردار پشتیبان را به همراه داشت.با وجود سطح زیر منحنی مشخصه عملکرد ژنومی بالاتر بیز لاسو و بهترین پیشبینی ناریب خطی ژنومی در جمعیت‌های مختلف، هنگامی که صفات گسسته توسط تعداد زیادی جایگاه صفت کمی کنترل شدند، روش ماشین بردار پشتیبان عملکرد بهتری داشت.

نتیجه‌گیری: علی‌رغم نقش مهم نرخ توزیع فنوتیپ دودویی در جمعیت مرجع، بهترین پیش‌بینی سطح زیر منحنی مشخصه عملکرد ژنومی صفات گسسته دودویی روش ماشین بردار پشتیبان به ساختار ژنتیکی جمعیت مورد آنالیز و پارامتر جریمه وابسته بود.

واژه‌های کلیدی: بیز لاسو، سطح زیر منحنی مشخصه عملکرد، صحت ژنومی، یادگیری ماشین

مقدمه

در سالهای أخیر، به کارگیری انتخاب ژنومی در نتیجه توسعه تکنولوژی فن‌آوری تعیین ژنوتیپ منجر به تسهیل پیشرفت ژنتیکی در برنامه‌های اصلاح نژادی شده است. در حقیقت صحت پیش‌بینی ژنومی از طریق انتخاب ژنومی افزایش و به سرعت در برنامه‌های اصلاح نژادی و خصوصاً برای صفات کمی گسترش یافته است. از دیدگاه اصلاح نژادی پرداختن محض به این نوع صفات به علت همبستگی منفی با برخی صفات گسسته از جمله مقاومت به بیماری‌ها، درجه سختی زایش و صفات رفتاری منجر به کاهش شایستگی ژنتیکی حیوان خواهد شد (۲٦). در نتیجه، پیشرفت قابل توجه در افزایش سوددهی اقتصادی در برنامه‌های اصلاح نژادی امروزی نیازمند فهم بهتر و ورود مستقیم به صفات با بروز فنوتیپی گسسته دارد (۱۸).

ماهیت گسسته این نوع صفات، اثر پذیری به وسیله ژن‌های متعدد و عدم تطابق با توزیع نرمال و وراثت مندلی چالش جدیدی را از منظر آماری در این نوع صفات مطرح کرده است در نتیجه، استفاده از انتخاب ژنومی و کاربرد آن در اصلاح‌نژاد دام به فهم بالای روش‌های آماری مورد استفاده در انتخاب ژنومی وابسته است. در سالهای أخیر روش‌ها و مدل‌های متنوعی برای انتخاب ژنومی پیشنهاد شده است تا از یک سوی مشکلات مربوط به ابعاد بالای ماتریس ضرایب و تنوع در میزان اثر نشانگرها بر صفت اصلاحی را حل کنند و از سوی دیگر با عملکرد بالای خود در ارزیابی صحت پیش‌بینی ژنومی راه را برای دستیابی هر چه سریعتر به پیشرفت ژنتیکی در اصلاح دام میسر سازند(۲۰).

بهترین پیشبینی ناریب خطی ژنومی[2] و روش‌های بیز از متداول‌ترین روش‌های آماری در ارزیابی ژنومی هستند که به ترتیب از ساختار واریانس – کواریانس و رگرسیون مستقیم فنوتیپبه نشانگرها برای رسیدن به تابعی از روابط ژنومی بهره می‌گیرد (٦). با این حال در سال‌های اخیر روش‌های یادگیری ماشین به طور گسترده‌ای جهت حل چالش‌های ارزیابی ژنومی صفات گسسته مطرح شده‌اند (۷). این الگوریتم‌ها

[2] Genomic best linear unbiased prediction

توانایی پیشبینی را در مجموعه‌ای از داده ها بدون نیاز به تصحیح الگوی خاصی از وراثت، بهینه کرده و در ارزیابی‌های ژنومی صفات گسسته نیز وارد شده‌اند (۱۳). از جمله این روش‌ها می‌توان ماشین بردار پشتیبان[3] را نام برد که علاوه بر صحت بالا در پیش‌بینی ژنومی (۷)، 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 آمینو اسیدها نقش برجسته‌ای دارند (۳۲).

تحقیقات در مورد استفاده از روش‌های یادگیری ماشین در ارزیابی ژنوم یصفات گسسته نشان از برتری این روش‌ها در مقایسه با روش‌های بیز داشت (۸). 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 (۲۲) و بهترین پیشبینی ناریب خطی ژنومی (۷) گزارش شده‌اند. با این حال، تفاوت عمده روش ماشین بردار پشتیبان در مقایسه با روش‌های مرسومی از جمله بهترین پیشبینی ناریب خطی ژنومی و بیز لاسو[4] عدم نیازبه نحوه توارث، توانایی بالا در به کارگیری اثرات غیر افزایشی، فرضیات در نظرگرفته شده برای مدل ژنتیکی پشت صحنه آن‌ها و تنظیم و بهینه سازی پارامترهای آنها جهت دستیابی به حداکثر صحت پیش‌بینی ژنومی می‌باشد (۷).

مطالعات مختلف در زمینه انتخاب ژنومی دال بر اهمیت مدل آماری ومعماری ژنومی صفات (تعداد جایگاه های صفات کمی، سطح عدم تعادل پیوستگی و وراثت‌پذیری صفت و غیره) بر صحت ارزش‌های اصلاحی ژنومی و ارزیابی ژنومی دارد (۱۵ و ۲۶). با این حال با توجه به اینکه صفات گسسته (از جمله بیماری‌ها) در شرایط و موقعیت‌های مختلف احتمال وقوع متفاوتی دارند. به نظر می‌رسد نسبت وقوع و توزیع فنوتیپ گسسته در جمعیت مرجع عواملی تأثیرگذار در برآورد ارزش‌های اصلاحی حیوانات جمعیت کاندیدا باشد (۲۴). لذا در پژوهش حاضر با الهام گرفتن از شرایط واقعی و از طریق یک مطالعه شبیه‌سازی سعی بر آن شده است این عامل (توزیع مختلف فنوتیپ گسسته در جمعیت مرجع) بر ارزیابی ژنومی روش‌های ماشین بردار پشتیبان، بهترین پیشبینی ناریب خطی ژنومی و بیز لاسو با در نظر گرفتن ساختار ژنتیکی متفاوت صفات مورد ارزیابی قرار گیرد.

مواد و روش‌ها

در تحقیق حاضر از نرم افزار QMSim (نسخه ۱/۱) جهت شبیه‌سازیاستفاده شد (۲۷). ابتدا یک جمعیت اولیه ۱۰۰۰ رأسی طی ۱۰۰۰ نسل شبیه سازی شد. سپس این جمعیت پایه در دو مسیر جهت تولید عدم تعادل پیوستگی پایین و بالا شبیه‌سازی و تکثیر یافت. در مسیر اول، جمعیت اولیه ۲۰۰ نسل دیگر (تا نسل ۱۲۰۰) تکثیر یافت و جمعیتی با عدم تعادل پیوستگی پایین شبیه‌سازی شد. در مسیر دوم، تعداد افراد جمعیت اولیه از طریق ایجاد یک گلوگاه ژنتیکی به ۱۰۰ رأس در نسل ۱۱۰۰ کاهش یافت. سپس این تعداد افراد (۱۰۰ رأس) برای۱۰۰ نسل دیگر (تا نسل ۱۲۰۰) تکثیریافته و به تعداد اولیه خود یعنی۱۰۰۰ رأس افزایش داده شدند. برای ایجاد جمعیت مرجع و تأیید، همه افراد (۱۰۰۰ رأس) آخرین نسل جمعیت پایه برای تولید مثل درجمعیت حاضرمورد استفاده قرارگرفتند که در این بین۴۰ رأس

[3] Support Vector Machine

[4] Least Absolute Shrinkage and Selection Operator

نر درنظرگرفته شد تا منعکس کننده ی نسبت نر به ماده ی (حدود ٤ صدم و اندازه مؤثرجمعیت ١٥٤) موجود درگله های گاو شیری باشد و بتوان اثر روش تلقیح مصنوعی برنسبت نر به ماده را تقلید کرد.

نوع سیستم آمیزشی تصادفی بود و برای پنج نسل دیگر (تانسل ۱۲۰۵) جمعیت تکثیر شدند. شانس آمیزش درهمه ی حیوانات برابر (در هر دو جنس) و یک فرزند برای هر زایش در نظرگرفته شد. درصد جایگزینی برای نر و ماده به ترتیب ۵۰ و ۲۰ درصد در نظرگرفته شد. در جمعیت اخیر، انتخاب حیوانات برتر برای نسل بعد بر اساس ارزش اصلاحی بالاو معیار حذف براساس ارزش اصلاحی پایین و سن صورت گرفت. نشانگرها به صورت دو آللی و به صورت تصادفی برای هر یک از کروموزوم ها با توجه به نقشه ژنتیکی گاو شیری (۲۹ کروموزوم در دامنه ٤٢ تا ۱۵۸ سانتی مورگان) توزیع شدند (۱۲). به ازای هر کروموزوم تعداد متفاوت نشانگر (دامنه ی ۱۸۰ تا ٦٥٣ با توجه به نوع کروموزوم) جهت تولید پنل های ۱۰ K شبیه‌سازی شد.

در مجموع دو سطح مختلف تعداد جایگاه صفات کمی ۱۰۰ (با دامنه یک تا هشت با توجه به نوع کروموزوم) و ۱۰۰۰ (دامنه ۱۰-۸۰ با توجه به نوع کروموزوم) در طول کروموزم ها توزیع شدند. نرخ جهش برای نشانگرها و تعداد جایگاه‌های صفات کمی در هر جایگاه و در هر نسل $۲/۵\times۱۰^{-۵}$ فرض شد. دو سطح مختلف وراثت پذیری (۰/۰۵ و ۰/۲) برای هرصفت در نظرگرفته شد. در طراحی جمعیت نهایی، افراد آخرین نسل (نسل ۱۲۰۵) به عنوان جمعیت تأیید (۱۰۰۰ رأس) در نظرگرفته شدکه این افراد اطلاعات ژنوتیپی داشته اما فاقد اطلاعات فنوتیپی بودند. همچنین افراد ٤ نسل ما قبل جمعیت تأیید (نسل ۱۲۰۱تا ۱۲۰٤) در گروه جمعیت‌های مرجع (٤۰۰۰ رأس) که این افراد هم اطلاعات ژنوتیپی داشته و هم ارزش‌های اصلاحی ژنومی آنها مشخص می‌باشد طبقه‌بندی شدند. با توجه به این که از دیدگاه آماری توزیع احتمال تعداد جایگاه‌های صفات کمی صفات مهم اقتصادی توسط شمار اندک ژن های دارای اثر عمده و درصد بالایی از ژن‌های کوچک اثر هستند و این فرضیه به توزیع گاما نزدیک است (۹). توزیع احتمال تعداد جایگاه‌های صفات کمی، گاما (۰/٤) فرض شد.

در این تحقیق فراوانی آللی اولیه برای نشانگرها ۰/۵ در نظر گرفته شد. در هر نسل و هر جایگاه، کل میزان واریانس افزایشی توسط تعداد جایگاه‌های صفات کمی توجیه شد. در مجموع هشت سناریو درتحقیق حاضر شبیه‌سازی شدند. برای ایجاد نسبت‌های مختلف توزیع فنوتیپ دودویی تغییراتی در فایل فنوتیپ خروجی QMSim ایجاد شد. به‌طوری‌که فنوتیپ پیوسته حیوانات به‌عنوان متغیر پاسخ (y) از طریق عوامل ثابت یا مستقل (اثر نسل =x1 و اثر جنسیت =x2) تجزیه واریانس شد تا اثر عوامل ثابت و تأثیرگذار بر فنوتیپ کنار گذاشته شود و در نهایت مقادیر تصادفی باقی‌مانده برای هر حیوان محاسبه شدند. در نتیجه برای شبیه‌سازی فنوتیپ دودویی در جمعیت مرجع، ابتدا باقی‌مانده‌ها از بیشترین به کم‌ترین مرتب شدند. در مرحله بعد با توجه نرخ توزیع فنوتیپ، باقی‌مانده‌های پیوسته از طریق سه رویکرد به فنوتیپ دودویی تبدیل شدند.

در رویکرد اول: حیواناتی از جمعیت مرجع که مقادیر باقی‌مانده‌ی آنها از میانگین کلی باقی‌مانده‌ها کمتر بود، کد یک (یا فنوتیپ غیر مطلوب: حدود ۵۰ درصد) و سایر حیوانات (که مقادیر باقی‌مانده‌ی آنها از میانگین کلی باقی‌مانده‌ها بیشتر بود)، کد صفر (یا فنوتیپ مطلوب) اختصاص داده شد (حدود ۵۰ درصد). در رویکرد دوم: حیواناتی که باقی‌مانده آنها پایین‌تر از $\bar{e}-1SD_e$ بود، کد یک (یا فنوتیپ

نامطلوب: حدود ۱۶ درصد) و سایر حیوانات کد صفر (یا فنوتیپ مطلوب: حدود ۸۴ درصد) اختصاص داده شد. در رویکرد سوم: حیواناتی که باقی‌مانده آنها پایین‌تر از $\bar{e}+1SD_e$ بود، کد یک (یا فنوتیپ نامطلوب: حدود ۸۴ درصد) و سایر حیوانات کد صفر (یا فنوتیپ مطلوب: حدود ۱۶ درصد) اختصاص داده شد.

در این تحقیق توجه به این که در بروز صفات گسسته از جمله بیماری‌ها ممکن است نشانگرهای بزرگ اثر و جهش یافته کنترل بیماری را تحت شعاع قرار دهند فراوانی آللی کمیاب کمتر از ۰/۰۰۵ در نظر گرفته شده است. برای ارزیابی صحت مدل‌ها ۱۰ تکرار برای هر سناریو در نظر گرفته شد. سطح عدم تعادل پیوستگی برای سناریوهای مختلف شبیه‌سازی شده با استفاده از محاسبه‌ی توان دوم ضریب همبستگی (r^2) بین همه‌ی جفت نشانگرهای ممکن ارزیابی گردد(۱۰). نرم افزار PLINK 1.9 برای برآورد عدم تعادل پیوستگی بین جفت نشانگرهای مختلف در ژنوم همه حیوانات موجود در آخرین نسل مورد استفاده قرار گرفت (۲۵).

مدل آماری

بهترین پیش‌بینی ناریب خطی ژنومی: این روش مشابه ارزیابی بهترین پیشبینی ناریب خطی در مدل‌های هندرسون است با این تفاوت که ماتریس روابط ژنومی ساخته شده از اطلاعات مولکولی (G) جایگزین ماتریس روابط شجره‌ای (A) می‌شود. در روش بهترین پیشبینی ناریب خطی ژنومی اثر نشانگرها به صورت تصادفی با واریانس یکسان در نظر گرفته می‌شود و برای برآورد اثر نشانگرها از معادلات مختلط هندرسون کمک می‌گیرد (۱۵). از الگوریتم AI-REML و بسته نرم افزاری DMU (۱۴) که در آن امکان کلاس بندی مدل مختلط خطی تعمیم یافته را با تابع لوجیک برای داده‌های دودویی ممکن می‌سازد استفاده شد (رابطه ۱).

$$logit\ (\pi_r) = \log\left[\frac{\pi_r}{1-\pi_r}\right] = \emptyset + \gamma_r \quad \text{رابطه ۱)}$$

که در آن π_r احتمال بروز فنوتیپ دودویی در حیوان r، $\emptyset$ اثر میانگین کل و γ_r اثر تصادفی حیوانات است که بر گیرنده ارتباط ژنومی میان افراد بر اساس اطلاعات نشانگرها می‌باشد. ماتریس خویشاوندی ژنومی **G** (۲۹) بر اساس روش مرسوم (۳۱) و جهت اجتناب از تشکیل ماتریس تکین مقدار ۰/۰۱ به عناصر قطری ماتریس روابط ژنومی (**G**) اضافه شد.

بیز لاسو: در سال ۱۹۹۶روش بیز لاسو (۳۰) پیشنهاد و مدل آستانه‌ای آن در سال ۲۰۰۸ تشریح شد (۲۳). در سال ۲۰۰۹ نسخه ژنومیک پیوسته (۶) و آستانه‌ای (۸) آن ترویج داد شد. این روش از نوع روش‌های جریمه‌ای توأم با انتخاب متغیراست، در این روش توزیع‌های پیشینی برای نشانگرها تعریف می‌شود که به اجبار تأثیر درصدی از نشانگرها را صفر فرض می‌کند. در این روش توزیع اثرات نشانگرها به صورت توزیع نمایی دوگانه است. این توزیع دارای تعداد بیشتری از اثرات غیرصفر کوچک است. روش لاسو از توزیع دو نمایی (مشروط به پارامتر تنظمی لاندا) برای توزیع اثرات تعداد جایگاه صفات کمی تحت مدل بیزین استفاده می‌کند. تخمین‌های لاسو از طریق توزیع‌های پسین بیزی، تحت مقادیر پیشین دونمایی مستقل برای اثرات تعداد جایگاه صفات کمی استنباط و استنتاج شود.

این مدل از قدرت زیرمجموعه‌های کاهشی و رگرسیون حاشیه‌ای از طریق مساوی قرار دادن برخی متغیرها با صفر، حاصل می‌شود. مدل لاسو را می‌توان بصورت مدل خطی تصادفی و نیز مدل خطی مختلط نمایش داد. این مدل روشی از حداقل مربعات

می‌باشد، که مجموع مربعات باقیمانده را حداقل می‌کند (رابطه ۲).

$$argmin\beta\left\{\sum_{i=1}^{n}(y_i-\sum_{j=1}^{p}x_{ij}\beta_j)^2\right\}subjectto\sum_j|\beta_j|\le t\,\hat{\beta}=argmin_\beta(y-X\beta)^T(y-X\beta)+\lambda||\beta||_1$$

رابطه ۲)

که در آن t پارامتری است که مقادیر پیشین را طوری از طریق ضرب لانگراژ گزینش می‌کند که λ≤0 باشد. روش بیز آستانه‌ای لاسو از طریق پارامترهای انقباضی روی نشانگرها متاثر می‌شود. مدل کلی بیز آستانه‌ای لاسو به صورت رابطه ۳ بیان شده است.

$$\lambda=\mu 1+X\hat{\beta}+e$$

رابطه ۳)

در اینجا λ متغیر پاسخ بوده، μ میانگین جمعیت، 1 ماتریسی با ابعاد n×1 به صورت بردار یکه می‌باشد. $\widehat{\beta}$ در بر گیرنده‌ی برآوردهای لاسو و در ارتباط با ماتریس X می‌باشد. در بیز آستانه‌ای لاسو باقی مانده‌ها (e) با فرض میانگین صفر و واریانس یک در نظر گرفته شدند. خصوصیت لاپلاس شرطی پسین برای برآوردهای لاسو به صورت زیر بیان شده است (رابطه ٤).

$$P\left(\boldsymbol{\beta}|\sigma_e^2\right)=\prod_{i=1}^{n}\frac{\gamma}{2\sqrt{\sigma_e^2}}e^{-\gamma|\beta_j|/\sqrt{\sigma_e^2}}$$

رابطه ٤)

در اینجا σ_e^2 واریانس باقی مانده، γ پارامتر کنترل کننده‌ی انقباض توزیع می‌باشد. توزیع گاما برای $\gamma 2$ با نرخ مشخص (r) و شکل فزونی پارامتر (δ) در نظر گرفته شد .برای اجرای روش‌های بیزی، از بسته نرم افزاری BGLR والگوریتم نمونه‌گیری گیبس برای نمونه‌گیری توزیع پسین شرطی اثرات نشانگری استفاده شد.

ماشین بردار پشتیبان: روش ماشین بردار پشتیبان یک الگوریتم ماشینی است که از طریق اطلاعات آموزشی به دسته بندی عوامل و تشخیص و تمایز الگوهای پیچیده در داده‌ها می‌شود (۳). این الگوریتم یک روش رایج در کلاس بندی پروفایل‌های بیان ژنحاصل از ریزآرایه‌ها و مسائل رگرسیون غیر خطی و دو کلاسه کاربرد فراوانی دارد. در ماشین بردار پشتیبان ابتدا اطلاعات نشانگری به فضای n بعدی به واسطه توابع کرنل نگاشت می‌شود (xi→Θxi, Θmap). در اینجا Θ ضریبی است که هر نشانگر (x) دریافت می‌کند و سپس در فضای Hilbert رگرسیون خطی اعمال می‌شود (۷). خطای ε-insensitive و در نهایت برآوردها از طریق رابطه ۵ تخمین زده می‌شوند.

$$y = w\Theta(x) + b$$

رابطه ۵)

ضرائب w و b با حداقل کردن رابطه ٦ باعث حداقل کردن خطا و پیچیدگی مدل می‌شوند.

$$R\,(C) = (1/2)||w||^2 + C\,\frac{1}{n}\sum_{i=1}^{n}L_\varepsilon\,(d_i.y_i)$$

رابطه ٦)

در اینجا $L\varepsilon(d_i.y_i)$ خطایی است که توسط تابع ε-insensitive اندازه‌گیری می‌شود.

$$L\varepsilon(d.y)=\begin{cases}|d-y|-\varepsilon.\,if\;|d-y|\ge 0\\ 0.\,othervise\end{cases}$$

رابطه ۷)

رابطه ۷ برآوردی از ε-insensitive است و با چشم پوشی از خطاهای کمتر از ε، مقدار پارامتر جریمه توازن بین خطای تقریب وزنی و میزانی که تا آن حد انحراف از ε تحمل می‌شود تا جواب برای مساله رگرسیونی حاصل شود را کنترل می‌کند. مقدار ε که توسط کاربر تعیین می‌شود و تخمینی از صحت داده‌های آموزشی است. بر این اساس تابع حل به صورت رابطه ۸ می‌باشد.

$$y=f(x.a_i.\boldsymbol{a}_i.)=\sum_{i=1}^{n}(a_i-\boldsymbol{a}_i)\,K(x.\mathrm{x}_i)+b$$

رابطه ۸ ه)

در اینجا a_i و $\boldsymbol{a_i}$ وزن‌های مثبتی هستند که به هر مشاهده داده می‌شود و خود از اطلاعات برآورد می‌شوند و ضریب کرنل $K(x.\mathrm{x}_i)$ یک ماتریس مثبت قطعی معین n×n است و به صورت $K(x, x_i)$ برآورد می‌شود. برای اجرای این روش از بسته نرم افزاری e1071 نسخه (۲-۱,۷) استفاده شد

(۱۶). همچنین برای هر سناریو بهترین میزان پارامتر جریمه جهت بهینه مدل استفاده شد.

برای ارزیابی صحت پیش‌بینی ارزش‌های اصلاحی ژنومی، سطح زیر منحنی مشخصه عملکرد[5] به‌عنوان معیار ارزیابی و با استفاده از ۱۰ تکرار شبیه‌سازی برای هر سناریو انجام گرفت (رابطه ۹). برای محاسبه مقدار سطح زیر منحنی مشخصه عملکرد از پکیج pROCR در محیط R استفاده شد.

$$AUROC=\frac{True\ Positive-False\ Positive}{1-False\ Positive}+(1-\frac{True\ Positive-False\ Positive}{1-False\ Positive})\times False\ Positive \quad \text{رابطه ۹}$$

هر چه سطح زیر منحنی مشخصه عملکرد مربوط به یک دسته بند بزرگتر باشد کارایی نهایی دسته بند مطلوب‌تر ارزیابی می‌شود. نمودار ROC روشی برای بررسی کارایی دسته بندها می‌باشد. که در آن از نرخ تشخیص صحیح دسته مثبت (True Positive Rate - TPR) و نرخ تشخیص غلط دسته منفی (False Positive Rate - FPR) استفاده می‌شود.

نتایج

اثر روش آماری بر سطح زیر منحنی مشخصه عملکرد: میانگین سطح زیر منحنی مشخصه عملکرد روش‌های بهترین پیش‌بینی ناریب خطی ژنومی و بیز لاسو و ماشین بردار پشتیبان برای نسبت‌های مختلف فنوتیپ دودویی جمعیت مرجع در هر یک از جمعیت‌های شبیه‌سازی با معماری‌های مختلف ژنومی در جدول ۱ نشان داده شده است. بیشترین و کمترین میزان سطح زیر منحنی مشخصه عملکرد برای روش ماشین بردار پشتیبان مشاهده شد. با این حال میانگین کلی سطح زیر منحنی مشخصه عملکرد در تمامی سناریوها نشان از برتری جزئی بیز لاسو(۰/۶۶۳) نسبت به بهترین پیش‌بینی ناریب خطی ژنومی (۰/۶۵۸) و ماشین بردار پشتیبان (۰/۶۴۵) داشت. تحقیقات شبیه‌سازی برای نرخ توزیع فنوتیپ دودویی برابر (۵۰ درصدی برای هر کدام از فنوتیپ‌ها) نشان داد که علارغم صحت قابل قبول در برخی سناریوها توسط روش‌های یادگیری ماشین، با این حال روش‌های بیزی عملکرد قدرتمندتری در ارزیابی ژنومی داشتند (۲۶) که تأییدی بر نتایج مطالعه حاضر دارد. همچنین مطالعات دیگر برای سطوح مختلف نرخ توزیع فنوتیپ در جمعیت شبیه‌سازی(۲۰) و واقعی (۱۷) نشان از برتری جزئی روش بهترین پیش‌بینی ناریب خطی ژنومی بر روش‌های یادگیری ماشین دارد که به‌طورکلی در تحقیق حاضر این روند مشاهده شد.

در تحقیق حاضر، با وجود برتری کلی روش بهترین پیش‌بینی ناریب خطی ژنومی در مقایسه با روش ماشین بردار پشتیبان برای میانگین کلی سناریوها، روش ماشین بردار پشتیبان در سناریوهای با تعداد بالای جایگاه صفات کمی عملکرد بهتری نسبت به بهترین پیش‌بینی ناریب خطی ژنومی داشت که موافق با نتایج سایر محققین بود (۱۱). برخلاف نتایج تحقیق حاضر، مطالعات در مورد صفات پیوسته نشان از برتری فاحش روش‌های بهترین پیش‌بینی ناریب خطی ژنومی(۷) و بیزی بر روش‌های یادگیری ماشین داشت. که دلیل این امر بکارگیری روش یادگیری ماشین با ماهیت ناپارامتری در ارزیابی صفات پارامتری (پیوسته) بود.

[5] The area under the receiver operating characteristic curve (AUROC)

جدول ۱: میانگین سطح زیر منحنی مشخصه عملکرد روش‌های بهترین پیش‌بینی ناریب خطی ژنومی، بیز لاسو و ماشین بردار پشتیبان برای نسبت‌های مختلف فنوتیپ دودویی جمعیت مرجع در هریک از جمعیت‌های شبیه‌سازی شده با معماری‌های مختلف ژنومی

Table 1. Average of AUROC using of GBLUP, Bayes LASSO and SVM methods for different proportions of the threshold phenotypes rate in training set for each of simulated population with different genomic architecture

مدل	نوع معماری ژنتیکی Type of genomic architecture			سطح زیر منحنی مشخصه عملکرد AUROC		
	تعداد QTL No. QTL	وراثت‌پذیری Heritability	عدم تعادل پیوستگی Linkage disequilibrium	$\bar{e}-1SD_e$[1]	$\bar{e}$[2]	$\bar{e}+1SD_e$[3]
GBLUP	100	0.05	H	0.642(0.008)	0.62(0.036)	0.579(0.021)
	1000	0.05	H	0.651(0.012)	0.622(0.021)	0.576(0.009)
	100	0.2	H	0.777(0.031)	0.749(0.001)	0.726(0.001)
	1000	0.2	H	0.809(0.013)	0.747(0.017)	0.72(0.022)
	100	0.05	L	0.596(0.029)	0.575(0.009)	0.55(0.027)
	1000	0.05	L	0.587(0.017)	0.58(0.028)	0.552(0.002)
	100	0.2	L	0.695(0.019)	0.665(0.014)	0.636(0.017)
	1000	0.2	L	0.738(0.009)	0.712(0.013)	0.678(0.015)
LASSO	100	0.05	H	0.652(0.041)	0.622(0.016)	0.6(0.017)
	1000	0.05	H	0.644(0.016)	0.624(0.002)	0.609(0.022)
	100	0.2	H	0.795(0.016)	0.772(0.025)	0.737(0.024)
	1000	0.2	H	0.781(0.031)	0.751(0.016)	0.733(0.011)
	100	0.05	L	0.617(0.002)	0.608(0.012)	0.552(0.018)
	1000	0.05	L	0.582(0.031)	0.568(0.011)	0.536(0.026)
	100	0.2	L	0.737(0.021)	0.699(0.015)	0.66(0.012)
	1000	0.2	L	0.706(0.014)	0.679(0.033)	0.642(0.017)
SVM	100	0.05	H	0.637(0.004)	0.605(0.018)	0.617(0.024)
	1000	0.05	H	0.623(0.011)	0.616(0.017)	0.563(0.017)
	100	0.2	H	0.719(0.016)	0.671(0.014)	0.638(0.012)
	1000	0.2	H	0.813(0.025)	0.762(0.002)	0.759(0.015)
	100	0.05	L	0.576(0.013)	0.553(0.004)	0.521(0.014)
	1000	0.05	L	0.597(0.022)	0.554(0.019)	0.56(0.012)
	100	0.2	L	0.685(0.009)	0.641(0.018)	0.635(0.033)
	1000	0.2	L	0.747(0.001)	0.693(0.014)	0.702(0.029)

[1, 2, 3] به ترتیب حدود ۸٤، ۵۰ و ۱٦ درصد افراد جمعیت مرجع فنوتیپ نامطلوب دارند. عدم تعادل پیوستگی بالا (H) و پایین (L)

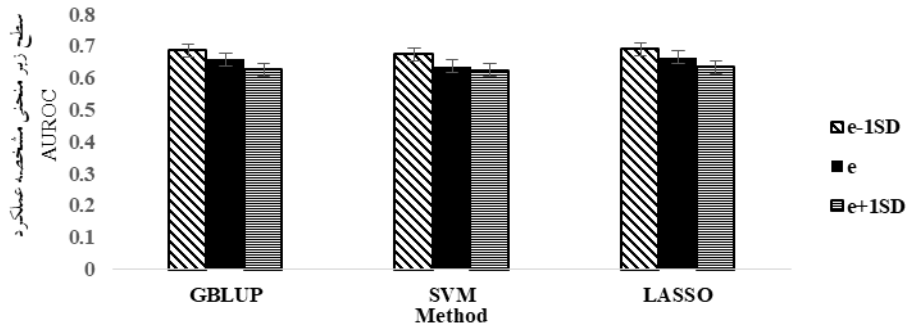
[1, 2, 3] showed that proportion of unpleasant phenotype in training set are 84, 50 and 16 percentage, respectively. High (H) and Low (L) linkage disequilibrium.

تحقیقات نشان داد که استفاده از روش‌های یادگیری ماشین برای صفات گسسته نسبت به صفات پیوسته افزایش ۱٤ تا ۱۹ درصدی سطح زیر منحنی مشخصه عملکرد را به همراه خواهد داشت (۱۹). در مجموع تفاوت محسوسی بین میانگین عملکرد کلی سطح زیر منحنی مشخصه عملکرد مدل‌های مختلف وجود نداشت. با این حال بسته به نوع معماری ژنتیکی صفت و نرخ توزیع فنوتیپ دودویی جمعیت مرجع، عملکرد سطح زیر منحنی مشخصه عملکرد تحت تأثیر مدل آماری قرار گرفت که در ادامه به آن پرداخته خواهد شد.

اثر نرخ توزیع فنوتیپ دودویی جمعیت مرجع بر سطح زیر منحنی مشخصه عملکرد: به‌طورکلی نرخ توزیع فنوتیپ دودویی میزان سطح زیر منحنی مشخصه عملکرد را تحت تأثیر قرار داد به‌طوری‌که در هر سه روش آماری با افزایش نرخ فنوتیپ نامطلوب

از ۱۶ بـه ۸٤ درصـد سـطح زیـر منحنـی مشخصـه عملکردکاهش یافت (شکل ۱). تحقیقات محدودی در زمینه ارزیابی نرخ فنوتیپ دودویی جمعیت مرجع بـر ارزیابی ژنومی انجام شده است. بـا ایـن حـال برخـی مطالعات نشـان دادنـد کـه نـرخ فنوتیـپ دودویـی از عوامـل مـؤثر بـر صـحت ژنـومی روش‌هـای جنگـل تصادفی و بهترین پیش‌بینی ناریب خطی ژنومی اسـت (۱۷). در یـک مطالعـه نشـان داده شـد کـه عملکـرد روش‌های بیز و بهترین پیش‌بینی ناریب خطی ژنـومی متاثر از نرخ فنوتیپ دودویی بوده و عملکـرد صـحت را در دامنه‌ی ۳۰ الـی ٤۰ درصـدی تحـت تـأثیر قـرار می‌دهد (۲۸). در مطالعـه‌ای روی صـفات بـا فنوتیـپ دودویی گاوهای هلشتاین آلمان نسـبت‌هـای مختلـف نرخ فنوتیپ دودویی در جمعیت مرجع مورد ارزیـابی قرار گرفت و نشـان داده شـد کـه صـحت پـیش‌بینـی ژنومی روش‌های مختلف آماری بـا افـزایش فنوتیـپ نـامطلوب در جمعیـت مرجـع (از ۵) بـه ۲۰ درصـد افـزایش و پـس رونـدی نزولـی دارد و بهتـرین نـرخ فنوتیپ دودویی یک جمعیت، نـرخ آسـتانه‌ای واقعـی جامعه (۲۰ درصد فنوتیپ نامطلوب) بود (۱۷).

در مطالعه حاضر بهترین عملکرد سطح زیر منحنی مشخصه عملکرد مربوط بـه جمعیـت هـایی بـا نـرخ فنوتیپ نامطلوب حدود ۱٦ درصـد ($\bar{e}$-$1SD_e$) بـود که نزدیک به نتایج سایر محققین برای داده‌های واقعی است (۱۷). در مقایسه با داده‌های واقعی، در مطالعات شبیه‌سازی عـلاوه بـر اثـر نـرخ فنوتیـپ دودویـی بـر صـحت ژنـومی، اشـتباه کلاسـبندی فنوتیـپ گسسـته می‌تواند باعث بـروز خطـا در نتـایج ارزیـابی ژنـومی شود. در تحقیق حاضر، نرخ فنوتیپ دودویی از طریق کد گذاری فنوتیپ پیوسته (باقیمانده) که دارای توزیع نرمـال بـود و از خروجـی QMSim اسـتخراج شـدند اعمال شد که این عمل به نوبه خـود موجـب افـزایش خطای کد گذاری در گروه $\bar{e}$ نسبت به دو گروه دیگر شد. دلیل این امر ایـن اسـت کـه در گـروه $\bar{e}$، تعـداد بیشتری از باقی‌مانده‌های حول محور میانگین (یـا مـد باقی‌مانده‌ها حول محور میانگین) بوده در نتیجه افـراد بیشتری بدون در نظر گرفتن شایستگی‌شـان و تنهـا بـا استفاده از فنوتیپ شان دسته بندی می‌شـوند کـه ایـن امر منجر به افزایش خطای کلاس بندی در گـروه $\bar{e}$ و کاهش سطح زیر منحنی مشخصه عملکرد شد.



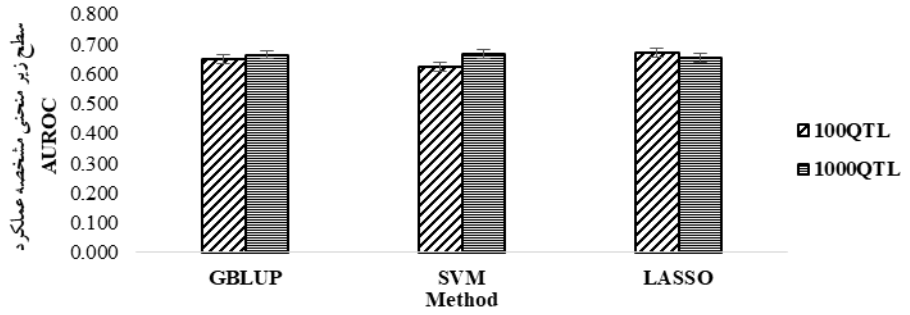
شکل ۱: میانگین سطح زیر منحنی مشخصه عملکرد روش‌های بهترین پیش‌بینی ناریب خطی ژنومی، بیز لاسو و ماشین بردار پشتیبان برای سطوح مختلف نرخ فنوتیپ دودویی

Figure 1. Average of AUROC using GBLUP, bayes LASSO and SVM methods for different levels of binary phenotype rate

اثر تعداد جایگاه صفات کمی بر سطح زیر منحنـی مشخصه عملکرد روش‌های آماری: شکل ۲، میانگین سطح زیر منحنی مشخصه عملکرد روش‌های بهتـرین پیش‌بینی ناریب خطی ژنومی، بیز لاسو و ماشین بردار پشتیبان را برای تعداد مختلـف جایگـاه صـفات کمـی نشان می‌دهد. به‌طورکلی تعداد جایگـاه صـفات کمـی عاملی مـؤثر بـر میـزان سـطح زیـر منحنـی مشخصـه عملکرد هر یک از روش‌های مورد مطالعه بـود. روش

ماشین بردار پشتیبان با عملکردی نسبتا مشابهی با بهترین پیش‌بینی ناریب خطی ژنومی و عملکرد بهتری نسبت به بیز لاسو، میزان سطح زیر منحنی مشخصه عملکرد را برای صفات گسسته با تعداد زیاد جایگاه صفات کمی بهبود بخشید. این در حالی بود که برای صفات تحت تأثیر تعداد پایین جایگاه صفات کمی روش بیز لاسو عملکرد بهتری داشت. برای سطوح مختلف توزیع فنوتیپ دودویی در جمعیت مرجع، بیشترین و کمترین میزان سطح زیر منحنی مشخصه عملکرد به‌ترتیب در روش ماشین بردار پشتیبان و بیز لاسو در تعداد بالای جایگاه صفات کمی مشاهده شد. این درحالی بود که برای صفات تحت تأثیر تعداد پایین جایگاه صفات کمی، بیشترین و کمترین میزان سطح زیر منحنی مشخصه عملکرد به‌ترتیب در روش بیز لاسو و ماشین بردار پشتیبان مشاهده شد.

تحقیقات در زمینه اثر تعداد جایگاه صفات کمی بر صحت ارزیابی ژنومی نشان داد که صحت روش‌های آماری به شدت متأثر از تعداد جایگاه صفات کمی است (۸). در تحقیق حاضر، اثر مثبت افزایش سطح زیر منحنی مشخصه عملکرد ژنومی با افزایش تعداد جایگاه صفات کمی برای روش‌های بهترین پیش‌بینی ناریب خطی ژنومی و ماشین بردار پشتیبان هم‌راستا با سایر مطالعات بود (۷). مطالعات نشان داد که با افزایش تعداد جایگاه‌های صفات کمی سهم هر جایگاه صفات کمی در ارزش ژنتیکی کل کاهش یافته و به نوعی قدرت مدل‌های بیز در پیش‌بینی این اثرات کوچک کاهش خواهد یافت. از طرف دیگر برای روش‌های بهترین پیش‌بینی ناریب خطی ژنومی و ماشین بردار پشتیبان در تعداد پایین جایگاه صفات کمی، احتمال شکل‌گیری توزیع اثرات ژنی کم بوده و توزیع آماری مورد نظر با تعداد ژن‌های بزرگ اثر و کوچک اثر به خوبی بیان و نمایان نمی‌شود که می‌تواند محتمل‌ترین دلیل برای نتایج به دست آمده باشد (۱). همچنین به خاطر این که توزیع اثرات جایگاه صفات کمی در این تحقیق گاما بود (که این توزیع به پیش فرض‌های روش بیزی سازگاری بیشتری دارد) و این مطلب در تحقیقات مختلف اذعان شده است که این توزیع در مقایسه با توزیع نرمال به پیش فرض‌های بهترین پیش‌بینی ناریب خطی ژنومی سازگاری کمتری دارد در نتیجه در تعداد اندک جایگاه صفات کمی نتایج متفاوتی برای روش‌های بیز لاسو در مقایسه با بهترین پیش‌بینی ناریب خطی ژنومی مشاهده شد (۲۰). در این تحقیق روش بیز لاسو نسبت به روش بهترین پیش‌بینی ناریب خطی ژنومی ازسطح زیر منحنی مشخصه عملکرد بالاتری برخوردار بود. این برتری در صفتی که تحت تأثیر تعداد کمی ژن بزرگ أثر قرار داشت، محسوس‌تر بود وزمانی که صفت تحت تاثیر تعداد زیادی ژن با اثرات کم قرار داشت، کاهش یافت. روش بهترین پیش‌بینی ناریب خطی ژنومیکمتر به معماری ژنتیکی صفت وابسته است و زمانی که صفت تحت تأثیر تعداد زیادی ژن قراردارد، از نظر سطح زیر منحنی مشخصه عملکرد تفاوت چندانی با روش لاسو ندارد. درنتیجه زمانی که در انتخاب ژنومی، معماری ژنتیکی صفت از نظر تعداد و میزان اثر جایگاه‌های صفت کمی آن کاملاً مشخص نباشد، روش مناسبی خواهد بود (۱۷). همچنین در روش باز نمونه‌گیری ماشین بردار پشتیبان، افزایش تعداد جایگاه صفات کمی، منجر به تولید عدم پیوستگی قوی بین برخی نشانگرها با جایگاه‌های صفات کمی کنترل کننده‌ی صفت، نزدیک‌تر شدن فاصله نشانگرها با جایگاه‌های صفات کمی و افزایش شانس نمونه‌گیری شده، در نتیجه افزایش سطح زیر منحنی مشخصه عملکرد را به همراه خواهد داشت (۱۸) که این اثر مثبت در نتایج تحقیق حاضر صادق بود.



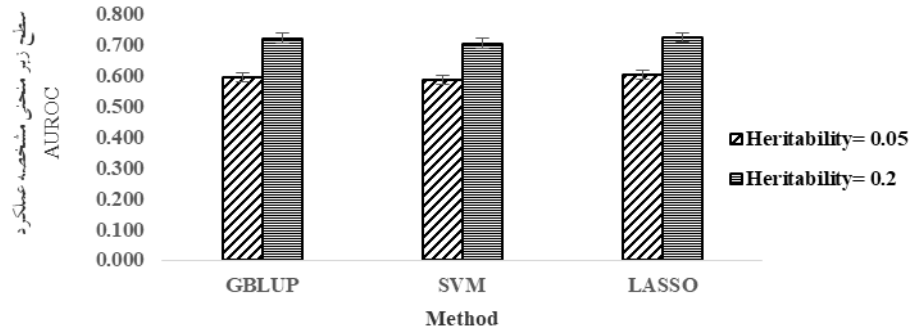
شکل ۲: میانگین سطح زیر منحنی مشخصه عملکرد روش‌های بهترین پیش‌بینی ناریب خطی ژنومی، بیز لاسو و ماشین بردار پشتیبان برای تعداد مختلف جایگاه صفات کمی

Figure 2. Average of AUROC using GBLUP, bayes LASSO and SVM methods for different number of QTL

شکل ۳، میانگین سطح زیر منحنی مشخصه عملکردروش‌های بهترین پیش‌بینی ناریب خطی ژنومی،بیز لاسو و ماشین بردار پشتیبان را برای سطوح مختلف وراثت پذیری نشان می‌دهد. بطور کلی وراثت پذیری عاملی مؤثر بر میزان سطح زیر منحنی مشخصه عملکرد هر یک از روش‌های مورد مطالعه بود بطوری که با افزایش وراثت پذیری میزان سطح زیر منحنی مشخصه عملکرد در هر سه روش آماری افزایش یافت که این نتایج با تئوری‌های ارائه شده در مورد ارتباط مستقیم بین وراثت‌پذیری و صحت پیش‌بینی ارزش‌های اصلاحی ژنومی مطابق بود (۲). با این حال در بررسی اثر سطوح مختلف وراثت‌پذیری بر صحت پیش‌بینی ژنومی جمعیت موش تفاوت محسوسی در صحت پیش‌بینی ژنومی روش ماشین بردار پشتیبان مشاهده نشد (۲۱).

در چندین مطالعه اثر مطلوب افزایش وراثت پذیری بر صحت پیش‌بینی ارزش های اصلاحی ژنومی ناشی از مدل‌های بیز (۲۰)، بهترین پیش‌بینی ناریب خطی ژنومی(۱۸) و ماشین بردار پشتیبان (۷) به اثبات رسیده است. این تأثیر مثبت و مطلوب وراثت‌پذیری منجر به تغییرات بالای ژنتیکی و در نتیجه کمک به پیش‌بینی بهتر اثرات نشانگری شد. به طور کلی دلیل افزایش صحت پیش‌بینی ژنومی با افزایش وراثت‌پذیری را می‌توان این گونه عنوان کرد که هر چه وراثت‌پذیری صفت بیشتر باشد، فنوتیپ فرد به ارزش ژنتیکی فرد نزدیک‌تر بوده و اثر نشانگرها و به دنبال آن ارزش‌های اصلاحی ژنومی افراد به طور صحیح‌تر پیش‌بینی می‌شود. مطالعات جهت پیش‌بینی صحت ژنومی از طریق فرمول $r=\sqrt{N_p h^2/N_p h^2+M_e}$ (که در اینجا N_p: تعداد افراد جمعیت مرجع، M_e: تعداد سیگمنت‌های کروموزوم مستقل و h^2: وراثت‌پذیری) نشان داد که صحت پیش‌بینی ژنومی ارتباطی مستقیمی با وراثت‌پذیری دارد (۵).

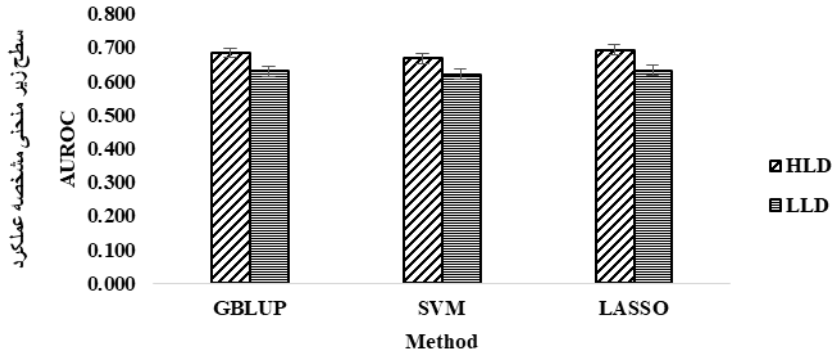
اثر عدم تعادل پیوستگی بر سطح زیر منحنی مشخصه عملکرد روش‌های آماری: شکل ٤، میانگین سطح زیر منحنی مشخصه عملکردروش‌های بهترین پیش‌بینی ناریب خطی ژنومی، بیز لاسو و ماشین بردار پشتیبان را برای سطوح مختلف عدم تعادل پیوستگی نشان می‌دهد. میانگین عدم تعادل پیوستگی برای جمعیت‌های با عدم تعادل پیوستگی پایین و بالا در فاصله ۰/۰۵ سانتی مورگان به‌ترتیب ۰/۲۲۱ و ۰/٤۳۵ بود و سطح زیر منحنی مشخصه عملکردناشی از روش‌های بهترین پیش‌بینی ناریب خطی ژنومی، بیز لاسو و ماشین بردار پشتیبان با افزایش سطح عدم تعادل پیوستگی افزایش یافت. به‌عنوان یک اصل کلی، وجود عدم تعادل پیوستگی بین نشانگرها و



شکل ۳: میانگین سطح زیر منحنی مشخصه عملکردروش‌های بهترین پیش‌بینی ناریب خطی ژنومی، بیز لاسو و ماشین بردار پشتیبان برای سطوح مختلف وراثت‌پذیری

Figure 3. Average of AUROC using GBLUP, bayes LASSO and SVM methods for different levels of heritability

جایگاه‌های صفات کمی منبع اصلی اطلاعات است و نقش عمده‌ای در صحت پیش بینی ارزش‌های اصلاحی ژنومی ایفا می‌کنند. تحقیقات نشان دادند که اگر سطح عدم تعادل پیوستگی بین نشانگرها افزایش ۱٦ درصدی داشته باشد، صحت برآورد ارزش‌های اصلاحی ژنومی از افزایشی ۱٤ درصدی خواهد یافت (٤). نتایج این تحقیق نشان داد که سطح بالای عدم تعادل پیوستگی میان جایگاه‌های صفات کمی و نشانگرها، باعث افزایش احتمال نمونه‌گیری نشانگرهای مجاور و دارای عدم تعادل پیوستگی بالا در روش‌های باز نمونه‌گیری می‌شود، که این امرعملکرد مثبت ماشین بردار پشتیبان را به همراه داشت.



شکل ٤: میانگین سطح زیر منحنی مشخصه عملکرد روش‌های بهترین پیش‌بینی ناریب خطی ژنومی، بیزلاسو و ماشین بردار پشتیبان برای سطوح مختلفعدم تعادل پیوستگی

Figure 4. Average of AUROC using GBLUP, bayes LASSO and SVM methods for different levels of LD

نتیجه‌گیری

نرخ توزیع فنوتیپ دودویی جمعیت مرجع از مهمترین عوامل مؤثر بر سطح زیر منحنی مشخصه عملکرد مدل‌های بهترین پیش‌بینی ناریب خطی ژنومی، بیز لاسو و ماشین بردار پشتیبان بود. هنگام که فنوتیپ نامطلوب در جمعیت نرخی نزدیک به شرایط واقعی و حدود $\bar{e}$-1SD_e داشت، مقدار سطح زیر منحنی مشخصه عملکرد حداکثر بود. با وجود برتری روش‌های رایج ارزیابی ژنومی از جمله

بیز لاسو و بهترین پیش‌بینی ناریب خطی ژنومی، روش ناپارامتری ماشین بردار پشتیبان یک روش قدرتمند در ارزیابی صفات گسسته و تحت تآثیر تعداد بالای جایگاه صفات کمی می‌باشد. با این حال استفاده از روش ماشین بردار پشتیبان در ارزیابی ژنومی منوط به تشخیص و شناسایی پارامتر C که نقش تعیین کننده‌ای جهت بیشینه کردن سطح زیر منحنی مشخصه عملکرد دارد خواهد داشت.

منابع

1. Abdollahi-Arpanahi, R., Pakdel, A., Nejati-Javaremi, A. and Shahrbabak, M. M. 2013. Comparison of genomic evaluation methods in complex traits with different genetic architecture. Journal of Animal Production. 15: 65-77.
2. Bo, Z., Zhang, J. J., Hong, N., Long, G., Peng, G., Xu, L.-Y., Yan, C., Zhang, L. P., Gao, H. J. and Xue, G. 2017. Effects of marker density and minor allele frequency on genomic prediction for growth traits in Chinese Simmental beef cattle. Journal of Integrative Agriculture. 16(4): 911-20.
3. 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.
4. Calus, M., De Roos, A. and Veerkamp, R. 2008. Accuracy of genomic selection using different methods to define haplotypes. Genetics. 178(1): 553-61.
5. Daetwyler, H. D., Calus, M. P., Pong-Wong, R., de los Campos, G. and Hickey, J. M. 2013. Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics. 193(2): 347-65.
6. De Los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K. and Cotes, J.M. 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics. 182(1): 375-85.
7. 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.
8. González-Recio, O. and Forni, S. 2011. Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. Genetics Selection Evolution. 43(1): 7.
9. Hayes, B. and Goddard, M. E. 2001. The distribution of the effects of genes affecting quantitative traits in livestock. Genetics Selection Evolution. 33(3): 209.
10. Hill, W. and Robertson, A. 1968. Linkage disequilibrium in finite populations. TAG Theoretical and Applied Genetics. 38(6): 226-231.
11. Honarvar, M. and Ghiasi, H. 2013. A comparison of genomic predictions using support vector machines (SVMs) and GBLUP methods. Agrochimica Research. 57: 3-21.
12. Kappes, S. M., Keele, J. W., Stone, R. T., McGraw, R. A., Sonstegard, T. S., Smith, T., Lopez-Corrales, N. L. and Beattie, C.W. 1997. A second-generation linkage map of the bovine genome. Genome Research. 7(3): 235-49.
13. Long, N., Gianola, D., Rosa, G.J., Weigel, K., and Avendano, 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(6): 377-89.
14. Madsen, P. and Jensen, J. 2010. A users guide to DMU. A package for analysing multivariate mixed models, Version 6.
15. Meuwissen, T., Hayes, B. and Goddard, M. 2001. Prediction of total genetic

value using genome-wide dense marker maps. Genetics. 157(4): 1819-29.
16. Meyer, D. 2014. Support Vector Machines—the Interface to libsvm in package. 1-8.
17. Naderi, S., Bohlouli, M., Yin, T. and König, S. 2018. Genomic breeding values, SNP effects and gene identification for disease traits in cow training sets. Animal Genetics. 49(3): 178-92.
18. 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): 7261-73.
19. 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): 129-39. (In Persian).
20. Naderi, Y. and Sadeghi, S. 2019. Assessment of the genomic prediction accuracy of discrete traits with imputation of missing genotypes. Animal Science Papers and Reports. 37(2): 149-68.
21. Neves, H. H., Carvalheiro, R. and Queiroz, S. A. 2012. A comparison of statistical methods for genomic selection in a mice population. BMC Genetics. 13(1):100.
22. Ogutu, J. O., Piepho, H. P. and Schulz-Streeck, T. 2011. A comparison of random forests, boosting and support vector machines for genomic selection. BMC proceedings. BioMed Central. 5(3): 11.
23. Park, T. and Casella, G. 2008. The Bayesian LASSO. Journal of the American Statistical Association. 103(482): 681-6.
24. 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.
25. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I. and Daly, M. J. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics. 81(3): 559-75.
26. Sadeghi, S., Rafat, S. A. and Alijani, S. 2018. Evaluation of imputed genomic data in discrete traits using Random forest and Bayesian threshold methods. Acta Scientiarum Animal Sciences. 40: 39007.
27. Sargolzaei, M. and Schenkel, F. S. 2009. QMSim: a large-scale genome simulator for livestock. Bioinformatics. 25(5): 680-1.
28. Shirali, M., Ashtiani, S., Pakdel, A., Hilli, K. and Vanoog, R. 2012. Comparison between Bayesc and GBLUP in estimating genomic breeding values under different QTL variance distributions. Iranian Journal of Animal Science (IJAS). 43(2): 261-8.
29. Su, G. and Madsen, P. User's Guide for GMATRIX version 2, a Program for Computing Genomic Relationship Matrix. 2013.
30. Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological). 267-88.
31. VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. Journal of Dairy Science. 91(11): 4414-23.
32. Yang, P., Hwa Yang, Y., B Zhou, B. and Y Zomaya, A. 2010. A review of ensemble methods in bioinformatics. Current Bioinformatics. 5(4): 296-308.

Gorgan University of Agricultural Sciences and Natural Resources
J. of Ruminant Research, Vol. 7(4), 2019
http://ejrr.gau.ac.ir

Genomic evaluation of support vector machine and common genomic prediction methods in different prevalence of threshold phenotype- A simulation study

Y. Naderi
Assistant Prof., Dept. of Animal Science, Young Resaerchers and Elite Club, Astara Branch, Islamic Azad University, Astara, Iran
Received: 09/05/2019; Accepted: 11/27/2019

Abstract

Background and objectives: Many prominent traits in livestock including disease resistance and dystocia, present a classification distribution of phenotypes. These traits are important in animal breeding due to importance of animal welfare and human tendency for healthy and high quality products. Therefore, identifying and characterizing the genetic variants that impact threshold traits, ranging from disease susceptibility, is one of the central objectives of animal genetics. In this regard, genomic selection can have an important role in increasing the genetic progress of the threshold traits. The objective of current study was genomic evaluation of area under receiver operating characteristic curve (AUROC) of support vector machine (SVM), GBLUP and Bayes LASSO methods for different rates of binary phenotype distribution in training set.

Materials and methods: A population of 1000 animals genotyped for 10,000 markers was simulated using QMSim software. Genomic population were simulated to reflect variations in heritability (0.05 and 0.2), number of QTL (100 and 1000) and linkage disequilibrium (low and high) for 29 chromosomes. In order to create different rates of discrete phenotype, the animal's phenotype of training set was coded as 1 (inappropriate phenotype) depending on whether their phenotype residuals was less than the average of residuals ($\bar{e}$), $\bar{e}$- $1SD_e$or $\bar{e}$+ $1SD_e$for the first, second and third approaches, respectively, and other individuals was defined as code 0 (appropriate phenotype). Three statistical models were implemented to analyze the simulated data including SVM, GBLUP and Bayes LASSO methods.

Results: Optimal training sets were characterized by inappropriate phenotype rate that were similar to the population real, leading to the highest AUROC in SVM, GBLUP and Bayes LASSO methods, in which concluded for $\bar{e}$- $1SD_e$ threshold point to the training set. The highest (0.813)and lowest(0.521) AUROC were observed for SVM method. Generally, heritability of trait was a factor affecting on genomic AUROC of SVM, GBLUP and Bayes LASSO methods; so that we recognized an increase in genomic AUROC with increase in heritability in all three statistical methods. Average r^2 in the low and high LD scenarios was 0.221 and 0.435 at distances of 0.05 cM and the results showed an increase in genomic AUROC using GBLUP, Bayes LASOO and SVM methods with increasing in linkage disequilibrium. The result of current study showed that high level of LD between SNP and QTLs increased the probability of adjacent markers sampling for re-sampling methods. Therefore, this resulted in a positive performance of SVM. Despite of the higher AUROC of GBLUP and Bayes LASSO

*Corresponding author; yousefnaderi@gmail.com

methods at different scenarios, SVM method showed a better performance when discrete traits were controlled by a large number of QTLs.

Conclusions: Despite the important role of different rates of binary phenotype distribution in training set, SVM method to predict genomic AUROC of discrete traits depends on genetic basis of the population analyzed and cost parameter.

***Keywords*:** Bayes LASSO, Area under receiver operating characteristic, Genomic accuracy, Machine learning