دانشگاه علوم کشاورزی و منابع طبیعی ساری
پژوهشهای تولیدات دامی

"مقاله پژوهشی"

بازیابی ژنوتیپ‌های از دست رفته با استفاده از روش هوشمند K-نزدیکترین همسایگی

فاطمه ونایی[۱]، فرهاد غفوری کسبی[۲]، احمد احمدی[۳] و پویا زمانی[۴]

۱- دانشجوی کارشناشی ارشد، گروه علوم دامی، دانشکده کشاورزی، دانشگاه بوعلی سینا، همدان، ایران
۲- استادیار، گروه علوم دامی، دانشکده کشاورزی، دانشگاه بوعلی سینا، همدان، ایران، (نویسنده مسوول: farhad_ghy@yahoo.com)
۳- استادیار، گروه علوم دامی، دانشکده کشاورزی، دانشگاه بوعلی سینا، همدان، ایران
۴- دانشیار، گروه علوم دامی، دانشکده کشاورزی، دانشگاه بوعلی سینا، همدان، ایران
تاریخ دریافت: ۱۴۰۰/۲/۱۵ تاریخ پذیرش: ۱۴۰۰/۹/۱۶
صفحه: ۱۳۰ تا ۱۳۸

چکیده مبسوط

مقدمه و هدف: بازیابی ژنوتیپ در طرح‌های انتخاب ژنومی به دلیل آنکه می‌تواند هزینه‌های انتخاب ژنومی را کاهش دهد بدون اینکه تأثیر منفی بر صحت انخاب ژنومی داشته باشد در طی سال‌های اخیر مورد توجه محققین قرر گرفته است. در فرآیند بازیابی ژنوتیپ، ژنوتیپ نشانگرهایی که به هر دلیل اطلاعات ژنوتیپی آن‌ها از دست رفته است با استفاده از روش‌های مختلف آماری بازیابی می‌شود.

مواد و روش‌ها: جهت ایجاد ماتریس ژنوتیپی، ژنومی متشکل از ۱ کروموزوم به طول یک مورگان برای ۲۵۰ و ۱۰۰۰ فرد شبیه‌سازی شد که روی آنها در سناریوهای مختلف، ۲۵۰، ۵۰۰، ۷۵۰، ۱۰۰۰، ۱۵۰۰ و ۲۰۰۰ نشانگر چند شکلی تک نوکلئوتیدی (SNP) توزیع شد. جهت ایجاد فایل اطلاعات حاوی ژنوتیپ های از دست رفته، اطلاعات ژنوتیپی به ترتیب ۵، ۱۰، ۲۵، ۵۰، ۷۵ و ۹۰ درصد SNPها از ماتریس ژنوتیپی حذف شده تا مجدداً توسط روش KNN بازیابی شوند. درصد ژنوتیپ‌های به درستی بازیابی شده (نسبت تعداد ژنوتیپ‌های به درستی بازیابی شده به کل ژنوتیپ‌های از دست رفته) و همبستگی بین ماتریس ژنوتیپی اولیه (فاقد اطلاعات ژنوتیپی از دست رفته) و ماتریس ژنوتیپی بازیابی شده به عنوان شاخص‌های صحت بازیابی ژنوتیپ مورد استفاده قرار گرفت.

یافته‌ها: در جمعیت شامل ۲۵۰ فرد صحت بازیابی ژنوتیپ در سناریوهای ۵، ۱۰، ۲۵، ۵۰، ۷۵ و ۹۰ درصد، صحت بازیابی ژنوتیپ به ترتیب برابر ۰/۸۲، ۰/۸۲، ۰/۸۰، ۰/۷۶، ۰/۶۲ و ۰/۴۰ بود اما با افزایش جمعیت به ۱۰۰۰ فرد، صحت بازیابی ژنوتیپ برابر ۰/۸۳، ۰/۸۳، ۰/۸۲، ۰/۸۲، ۰/۷۱ و ۰/۵۴ حاصل شد که بویژه تاثیر افزایش جمعیت در دو سناریو ۷۵ و ۹۰ درصد ژنوتیپ از دست رفته قابل توجه بود. همبستگی بین ماتریس ژنوتیپی اولیه و ماتریس ژنوتیپی بازیابی شده نیز با افزایش درصد حذف کاهش یافت. با افزایش تعداد SNPها از ۲۵۰ به ۲۰۰۰، صحت بازیابی ژنوتیپ از ۰/۶۷ به ۰/۸۴ افزایش یافت. همچنین یک رابطه معکوس بین فراوانی آلل نادر (MAF) با صحت بازیابی ژنوتیپ مشاهده شد به صورتیکه با افزایش MAF از ۰/۰۱ به ۰/۵، صحت بازیابی ژنوتیپ ۱۵ درصد کاهش نشان داد. مدت زمان بازیابی ژنوتیپ نیز با افزایش ابعاد ماتریس ژنوتیپی به صورت تصاعدی افزایش یافت. با افزایش درصد ژنوتیپ‌های بازیابی شده، صحت پیش‌بینی ارزش‌های اصلاحی ژنومی کاهش یافت. در سناریوهای ۵ و۱۰ درصد ژنوتیپ بازیابی شده تغییری در صحت مشاهده نشد اما در دو سناریو ۷۵ و ۹۰ درصد ژنوتیپ بازیابی شده، صحت پیش‌بینی ارزش‌های اصلاحی ژنومی به ترتیب ۱۶ و ۳۲ درصد کاهش یافت.

نتیجه‌گیری: به طور کلی صحت بازیابی ژنوتیپ‌های از دست رفته با استفاده از روش KNN قابل قبول بود به طوری که با افزایش درصد ژنوتیپ‌های از دست رفته تا ۵۰ درصد، KNN با صحتی در حدود ۸۰ درصد ژنوتیپ‌های از دست رفته را بازیابی نمود و بنابراین می‌توان این روش را برای طرح‌های انتخاب ژنومی پیشنهاد نمود.

واژه‌های کلیدی: بازیابی ژنوتیپ، فراوانی آلل نادر، نشانگر چند شکلی تک نوکلئوتیدی، K -نزدیکترین همسایگی

مقدمه

انتخاب ژنومی (۱۵) فرم پیشرفته انتخاب به کمک نشانگر[۱] است (MAS). در انتخاب ژنومی فرایند انتخاب از طریق تعیین ژنوتیپ افراد برای تعداد زیادی نشانگر چند شکلی تک نوکلئوتیدی SNP[۲] صورت می‌گیرد که بر حسب گونه تعداد SNPها متفاوت است (۱، ۱۶). برای مثال، در حال حاضر از تراشه‌های متراکم SNP که در برگیرنده اطلاعات ژنوتیپی ۵۴۰۰۰ SNP است به صورت تجاری برای تعیین ژنوتیپ گاوهای شیری استفاده می‌شود. اگر چه تراشه‌های بسیار متراکم‌تر که حاوی اطلاعات بیش از ۷۰۰۰۰۰ SNP هستند نیز به صورت محدودتر استفاده می‌شوند (۲۵). در هنگام تعیین ژنوتیپ حیوانات با تراشه‌های SNP معمولاً به طور تصادفی اطلاعات برخی SNPها از دست می‌رود. این SNPها در اصطلاح SNPهای ناخوانا[۳] نامیده می‌شوند. این مسأله تصادفی است و از حیوانی به حیوان دیگر تعداد این SNPها متفاوت است ولی در یک حالت کلی و بسته به گونه و تکنولوژی مورد استفاده معمولاً اطلاعات ۰/۵ درصد تا ۲۰ درصد از SNPها در هنگام تعیین ژنوتیپ کردن از دست می‌رود (۳۰). یک راه حل در مواجهه با چنین اطلاعاتی حذف اطلاعات ژنوتیپی SNPهایی است که برای بخشی از جمعیت اطلاعات ژنوتیپی آنها در دسترس نیست. این کار با تعیین یک آستانه انجام می‌شود و SNPهایی که فراوانی افرادی که برای آن SNP فاقد ژنوتایپ هستند از آستانه تعیین شده بالاتر باشد از فایل ژنوتیپی حذف می‌شوند. راه حل دوم این است که قبل از انجام ارزیابی ژنومی اطلاعات از دست رفته به نحوی بازیابی شوند چرا که برخی از این SNPها ممکن است بزرگ اثر باشند و فقدان اطلاعات مربوط به آنها منجر به کاهش صحت ارزیابی ژنومی شود. در مواردی نیز افراد جمعیت مرجع با پنل‌های SNP با تراکم بالا[۴] (HDM) برای مثال با پنل‌های ۵۴۰۰۰ SNP تعیین ژنوتیپ می‌شوند، اما اطلاعات ژنوتیپی افراد کاندید انتخاب با پنل‌های SNP با تراکم پایین[۵] (LDM) برای مثال با ۱۰۰۰۰ SNP به دست آمده است (۱۸،۱۷). در این حالت اطلاعات ۴۴۰۰۰ SNP دیگر افراد جمعیت تأیید با در نظر گرفتن اطلاعات ژنوتیپی جمعیت مرجع بازیابی می‌شود. اخیراً به منظور توسعه انتخاب ژنومی استفاده از اطلاعات توالی‌یابی با ژنوتایپینگ[۶] (GBS) که یکی از پروتکل‌های خانواده توالی‌یابی نسل جدید[۷] (NGS) است مورد توجه قرار گرفته است چرا که هزینه تعیین

1- Marker assisted selection
2- Single nucleotide polymorphism
3- Uncall
4- High density marker
5- Low density marker
6- Genotype by sequencing
7- Next generation sequencing

[DOR: 20.1001.1.22518622.1401.13.35.9.5]
[DOI: 10.52547/rap.13.35.130]

ژنوتیپ با استفاده از این روش به مراتب از روش‌های رایج کمتر است (۶). اما یک اشکال اطلاعات GBS این است که اطلاعات ژنوتیپی بخشی زیادی از SNPها در آنها وجود ندارد چرا که همه SNPها تعیین ژنوتیپ نمی‌شوند. در این روش، SNPهای فاقد اطلاعات ژنوتیپی در حیوانات مختلف مشابه نبوده و حالت تصادفی دارند و تا ۹۰ درصد از SNPها را شامل می‌شود. ژنوتیپ‌های از دست رفته در اطلاعات GBS در صورتیکه ژنوم مرجع و یا نقشه‌های کامل پیوستگی[1] در اختیار نباشد با استفاده از روش‌های رایج بازیابی ژنوتیپ قابل بازیابی نیستند (۱۸،۴). تحت چنین شرایطی، برخی روش‌های هوشمند که عمدتاً جزء روش‌های ناپارامتری می‌باشند برای بازیابی اطلاعات از دست رفته فایل‌های اطلاعات GBS چه با استفاده از ژنوم مرجع یا بدون آن قابل استفاده هستند. از آن جمله می‌توان به روش‌های K–نزدیکترین همسایگی[2] (KNN) (۲۷)، تجزیه مقدار تکین[3] (SVD) (۸) و جنگل تصادفی[4] (RF) (۹) اشاره کرد. KNN یک روش ناپارامتری و جزو خانواده روش‌های یادگیری ماشین است که در داده‌کاوی و تشخیص الگو مورد استفاده قرار می‌گیرد. بر اساس آمارهای ارائه شده، الگوریتم K–نزدیک‌ترین همسایگی یکی از ده الگوریتمی است که بیشترین استفاده را در پروژه‌های گوناگون یادگیری ماشین و داده‌کاوی داشته است. روش KNN برای بازیابی داده‌های از دست رفته در تحقیقات مختلف از مرتعداری گرفته (۷) تا پزشکی (۱۳) و مطالعات پویش ژنومی در انسان (۲۲) مورد استفاده قرار گرفته است. استفاده از این روش در علوم دامی به چند سال اخیر محدود می‌شود. از این روش برای دسته‌بندی نژادهای مختلف گاو شیری بر اساس اطلاعات ابعاد بدن (۱۴)، بررسی میزان لنگش در گاوهای شیری (۳۳)، مطالعه رفتار تغذیه‌ای و نشخوار در گاوهای شیری (۲۴)، امتیازدهی به قسمت‌های مختلف بدن در گاوهای شیری (۲۱)، مطالعه رفتار گاوهای شیری در گله (۳) و تشخیص فحلی در گاوهای شیری (۳۱) استفاده شده است. به هر حال تا کنون این روش در مطالعات انتخاب ژنومی در دام‌ها مورد استفاده قرار نگرفته است و یا اینکه اطلاعات مکتوب آن در دسترس نیست. در این تحقیق عملکرد روش K–نزدیکترین همسایگی در بازیابی ژنوتیپ‌های از دست رفته مورد بررسی قرار خواهد گرفت و در ضمن عملکرد آن با نتایج گزارش شده برای دو روش هم خانواده KNN شامل تجزیه مقدار تکین (۸) و جنگل تصادفی (۹) مقایسه خواهد شد. همچنین تاثیر بازیابی ژنوتیپ بر صحت پیش‌بینی ارزش‌های اصلاحی ژنومی نیز بررسی می‌شود.

مواد و روش‌ها

شبیه‌سازی اطلاعات

برای تشکیل ماتریس ژنوتیپی از بسته نرم‌افزاری *hypred* نسخه 0.4 (۲۶) استفاده شد. جمعیت پایه به تعداد ۱۰۰ فرد (۵۰ نر و ۵۰ ماده) شبیه‌سازی شده و اجازه داده شد تا برای ۵۰ نسل به طور تصادفی در آن آمیزش صورت بگیرد. 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 ۱۰۰ بود. در نسل ۵۱ اندازه جمعیت به ترتیب به ۲۵۰ و ۱۰۰۰ فرد افزایش داده شد که این افراد هم اطلاعات ژنوتیپی داشته و هم اطلاعات فنوتیپی و همچنین ارزش‌های اصلاحی ژنومی آنها مشخص بود که این افراد جمعیت مرجع را تشکیل دادند. در ادامه نسل ۵۲ از افراد نسل ۵۱ ایجاد شد که افراد این نسل دارای اطلاعات ژنوتیپی بوده اما اطلاعات فنوتیپی نداشتند. در واقع نسل ۵۲ جمعیت تأیید را تشکیل داده بودند. ژنومی متشکل از ۱ کروموزوم به طول یک مورگان برای به ترتیب ۲۵۰ و ۱۰۰۰ فرد شبیه‌سازی شد که بر روی آن در سناریوهای مختلف، ۲۵۰، ۵۰۰، ۷۵۰، ۱۰۰۰، ۱۵۰۰ و ۲۰۰۰ SNP پخش گردید (جدول ۱). به هر جایگاه SNP با ژنوتیپ AA کد ۲، ژنوتیپ Aa کد ۱ و ژنوتیپ aa کد صفر اختصاص داده شد. برای بررسی حداقل فراوانی آللی[5] (MAF)، سطوح مختلف MAF شامل ۰/۰۱، ۰/۰۵، ۰/۱۰، ۰/۲۰، ۰/۳۰، ۰/۴۰ و ۰/۵۰ در نظر گرفته شد. جهت ایجاد فایل‌های داده حاوی ژنوتیپ‌های از دست رفته به ترتیب ۵، ۱۰، ۲۵، ۵۰، ۷۰ و ۹۰ درصد از اطلاعات ژنوتیپ‌ها حذف شدند تا مجدد بازیابی شوند. انتخاب این مقادیر بر اساس مقالات مرتبط (۱۱،۹،۸) و به منظور بررسی تاثیر درصد ژنوتیپ از دست رفته بر صحت بازیابی ژنوتیپ و صحت پیش‌بینی ارزش‌های اصلاحی ژنومی انجام گرفت.

روش KNN

در روش KNN (۲۷) که یک روش غیرپارامتری است ژنوتیپ از دست رفته از طریق جایگزینی آن با میانگین وزنی ژنوتیپ‌های معلوم نشانگر مربوطه در افراد دیگر بازیابی می‌شوند. به این منظور فاصله اقلیدسی (d در رابطه پایین) به عنوان معیاری از فاصله نشانگری استفاده می‌شود که این فاصله برای بردارهای حاوی ژنوتیپ‌های نشانگری که دارای طول m (تعداد نشانگر) باشند به صورت زیر تعریف خواهد شد:

$$d = \sqrt{\frac{1}{(G_{Imp} - G)^2}}$$

در رابطه بالا G_{Imp} ژنوتیپ بازیابی شده (به صورت یک کد عددی) برای نشانگر فاقد ژنوتیپ است که به صورت مقدماتی توسط روش MNI[6] برآورد می‌شود و G ژنوتیپ جایگاه مشابه در افراد دیگر با ژنوتیپ معلوم هستند و به صورت یک کد عددی می‌باشد. به طور خلاصه بازیابی ژنوتیپ در طی مراحل زیر صورت می‌گیرد. ۱) ابتدا ژنوتیپ‌های از دست رفته به وسیله روش MNI که روشی حد واسط است بازیابی می‌شوند. روش MNI که یک روش بازیابی نسبتا ضعیف است یک بازیابی اولیه از ژنوتیپ از دست رفته نشانگر j (ژامین SNP) را به دست می‌دهد که به عنوان یک پیش برآورد برای KNN قلمداد می‌شوند. برای مثال اگر در یک جمعیت ۱۰۰ نفره، در یک جایگاه یک ژنوتیپ از دست رفته وجود داشته باشد و در همین جایگاه در افراد دیگر ۱۰ ژنوتیپ A1A1، ۶۰ ژنوتیپ A1A2 و ۹ ژنوتیپ A2A2 وجود داشته باشد، پیش‌بینی اولیه

1- Linkage map
2- K-nearest neighbor
3- Singular value decomposition
4- Random Forest
5- Minor Allele Frequency
6- Mean Neighbor Imputation

MNI برای ژنوتیپ از دست رفته برابر خواهد بود با میانگین ۹۹ جایگاه دیگر که ژنوتیپ معلوم دارند. ۲) در مرحله بعد همه نشانگرها در ماتریس ژنوتیپی وارد شده و ماتریس ژنوتیپی کامل می‌شود و در ادامه فاصله اقلیدسی بین نشانگرj فاقد ژنوتیپ با جایگاه‌های مشابه که ژنوتیپ معلوم دارند برآورد می‌شود. نشانگرهای با ژنوتیپ معلوم بر اساس فاصله اقلیدسی با نشانگرj رتبه‌بندی می‌شوند. برای نشانگرj، میانگین وزنی تعداد k نزدیکترین نشانگر با ژنوتیپ معلوم به عنوان برآوردی از ژنوتیپ برای آن نشانگر در نظر گرفته می‌شود (مقدار k توسط کاربر مشخص می‌شود). در اینجا وزنی که به هر نشانگر داده می‌شود برابر $1/d^2$ که در این رابطه d فاصله اقلیدسی بین تک تک نشانگرهای دارای ژنوتیپ معلوم و نشانگرj است.

تجزیه و تحلیل‌های KNN در محیط نرم‌افزار R و با استفاده از تابع kNNI انجام شد. هر سناریو ۱۰ بار تکرار شد و میانگین ۱۰ تکرار برآورد و گزارش گردید.

دو شاخص برای ارزیابی صحت بازیابی ژنوتیپ استفاده شدند. ۱) درصد ژنوتیپ‌های به درستی بازیابی شده (تعداد تعداد ژنوتیپ‌های به درستی بازیابی شده تقسیم بر تعداد ژنوتیپ از دست رفته ضرب در ۱۰۰) و ۲)همبستگی بین ماتریس ژنوتیپی اولیه (فاقد اطلاعات ژنوتیپی از دست رفته) و ماتریس ژنوتیپی بازیابی شده.

جدول ۱- پارامترهای مورد استفاده در شبیه‌سازی ماتریس ژنوتیپی

Table 1. Parameters used in simulation of genotypic matrix

اندازه ژنوم	۱ مورگان
تعداد کروموزوم	۱
تعداد SNP	۲۰۰۰، ۱۵۰۰، ۱۰۰۰، ۷۵۰، ۵۰۰، ۲۵۰
تعداد افراد	۱۰۰۰، ۲۵۰
تعداد مؤثر جمعیت	۱۰۰
فراوانی آلل نادر	۰/۵، ۰/۴، ۰/۳، ۰/۲، ۰/۱، ۰/۰۵، ۰/۰۱

برای بررسی تأثیر بازیابی ژنوتیپ بر صحت پیش‌بینی ارزش‌های اصلاحی ژنومی، ابتدا ارزش‌های اصلاحی ژنومی با استفاده از ماتریس ژنوتیپی کامل و به وسیله روش GBLUP[1] (۲۸) در قالب بسته نرم‌افزاری BGLR (۵) پیش‌بینی شده و صحت پیش‌بینی محاسبه گردید. در مرحله بعد، اطلاعات ژنوتیپی مربوط به ۵، ۱۰، ۲۵، ۵۰ ۷۰ و ۹۰ درصد از SNPها از ماتریس ژنوتیپی حذف شده و ژنوتیپ‌های از دست رفته توسط KNN بازیابی شده و مجدداً ارزش‌های اصلاحی ژنومی با استفاده از ماتریس ژنوتیپی حاوی ژنوتیپ‌های بازیابی شده برآورد شد و کاهش در صحت پیش‌بینی ارزش‌های اصلاحی ژنومی در هر یک از سناریوهای حذف (۵، ۱۰، ۲۵، ۵۰ ۷۰ و ۹۰ درصد) بررسی گردید.

نتایج و بحث

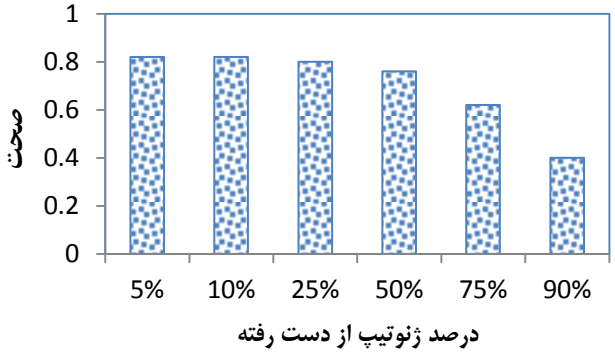
اگرچه استفاده از روش KNN در علوم دامی به چند سال اخیر محدود می‌شود، نتایج عملکرد مطلوب این روش را نشان می‌دهند. برای مثال وانگ و همکاران (۳۱) روش‌های مختلفی را در تشخیص فحلی در گاوهای شیری مورد مقایسه قرار دادند و گزارش نمودند که نتایج روش KNN و شبکه عصبی پسا انتشار[2] (BPNN) نسبت به بقیه روش‌ها از صحت بالاتری برخوردار بود. همچنین وو و همکاران (۳۳) از الگوریتم‌های مختلفی برای مشخص نمودن لنگش در گاوهای شیری با استفاده از تصاویر ویدیویی استفاده نمودند. این محققین نشان دادند که روش KNN با صحتی برابر ۹۵ درصد گاوهای لنگ را از گاوهای سالم تفکیک می‌کند. بعلاوه شِن و همکاران (۲۴) روش‌های یادگیری ماشین شامل ماشین بردار پشتیبان[3] (SVM)، روش KNN و شبکه عصبی را در بررسی رفتار تغذیه‌ای و نشخوار گاوهای شیری را از طریق نصب سنسورهای زیستی بر روی فک گاوها مورد مقایسه قرار دادند و گزارش کردند که با استفاده از اطلاعات حاصل از سنسورها، روش KNN اطلاعات صحیح‌تری از رفتار نشخوار را در اختیار قرار می‌دهد.

در ارتباط با بازیابی ژنوتیپ، اگرچه روش KNN توانایی بازیابی داده‌های از دست رفته را دارد، تا کنون مطالعه‌ای در ارتباط با بازیابی ژنوتیپ‌های از دست رفته با استفاده از این روش در اصلاح نژاد دام انجام نشده است. لذا این تحقیق به منظور بررسی عملکرد این روش در بازیابی ژنوتیپ‌های از دست رفته انجام شد. تأثیر درصد حذف بر صحت بازیابی ژنوتیپ در جمعیت شامل ۲۵۰ فرد که هر کدام برای ۱۰۰۰ SNP ژنوتیپ شده‌اند در شکل ۱ نشان داده شده است. در سناریوهای ۵، ۱۰، ۲۵، ۵۰، ۷۵ و ۹۰ درصد ژنوتیپ از دست رفته، روش KNN با صحتی برابر به‌ترتیب ۰/۸۲، ۰/۸۲، ۰/۸۰، ۰/۷۶، ۰/۶۲ و ۰/۴۰ ژنوتیپ‌های از دست رفته را بازیابی نمود. همانگونه که مشاهده می‌شود با افزایش درصد حذف از ۵ به ۹۰ درصد، صحت بازیابی ژنوتیپ کاهش یافت. کاهش در صحت بازیابی ژنوتیپ در زمان افزایش درصد ژنوتیپ گمشده از ۵ به۵۰ درصد چندان قابل ملاحظه نبود (۷ درصد) ولی پس از آن و با افزایش درصد ژنوتیپ‌های از دست رفته به ۷۵ و ۹۰ درصد، میزان کاهش درصحت بازیابی ژنوتیپ بسیار قابل توجه بود به صورتیکه در سناریو ۹۰ درصد ژنوتیپ از دست رفته صحت بازیابی ژنوتیپ نسبت به سناریو ۵ درصد ژنوتیپ ازدست رفته به میزان ۵۰ درصد کاهش یافت. همچنین همبستگی بین ماتریس ژنوتیپی اولیه و ماتریس ژنوتیپی بازیابی شده نیز با افزایش درصد ژنوتیپ از دست رفته کاهش یافت به طوریکه بیشترین همبستگی در زمان بازیابی ۵ درصد ژنوتیپ از دست رفته (۰/۹۹۸) و کمترین آن در زمان بازیابی ۹۰ درصد ژنوتیپ از دست رفته (۰/۷۲۴) مشاهده شد (جدول ۲). در تمامی گزارشات مربوط به بازیابی ژنوتیپ با روش‌ها و نرم افزارهای متفاوت کاهش در صحت بازیابی ژنوتیپ در نتیجه افزایش میزان ژنوتیپ‌های از دست رفته مشاهده شده است. برای مثال هیکی و همکاران

1- Genomic best linear unbiased prediction 2- Back-propagation neural network 3- Support Vector

(۱۱) درصدهای مختلف شامل ۵، ۵۰، ۷۵، ۸۷، ۹۴، ۹۸ و ۹۹ درصد از اطلاعات ژنوتیپی یک پنل حاوی اطلاعات ژنوتیپی ۵۴۰۰۰ SNP را حذف و سپس بازیابی نموده و گزارش کردند که با افزایش حذف اطلاعات ژنوتیپی و بازیابی آن میزان صحت بازیابی ژنوتیپ از حدود ۱ (بازیابی ۵ درصد از ژنوتیپ‌ها) به ۰/۲۰ (بازیابی ۹۹ درصد از ژنوتیپ‌ها) کاهش یافت. غفوری کسبی و گودرز تله‌جردی (۸) اطلاعات ژنوتیپی به ترتیب ۵، ۱۰، ۲۵، ۵۰، ۷۵ و ۹۰ درصد از SNPهای ۱۰۰۰ فرد را از ماتریس ژنوتیپی افراد حذف نموده و مجدداً توسط روش تجزیه مقدار تکین (SVD) بازیابی نمودند و گزارش کردند که با افزایش درصد حذف ژنوتیپ‌ها از ۵ درصد به ۹۰ درصد، صحت بازیابی ژنوتیپ از ۸۰ درصد به حدود ۵۰ درصد کاهش یافت. علت این مساله این است که با افزایش درصد حذف ژنوتیپ‌ها، میزان اطلاعات قابل بهره برداری توسط الگوریتم برای بازیابی ژنوتیپ‌های نامعلوم کاهش می‌یابد. برای مثال در سناریو ۱۰۰۰ SNP و ۱۰۰۰ فرد ماتریس ژنوتیپی حاوی ۱ میلیون ژنوتیپ خواهد بود. اگر ۵ درصد از این ژنوتیپ‌ها حذف شوند، اطلاعات ۹۹۵ هزار ژنوتیپ دیگر معلوم است اما در زمان حذف ۹۰ درصد از ژنوتیپ‌ها، فقط اطلاعات ۱۰۰۰۰ ژنوتیپ معلوم و قابل استفاده است. بری و کنی (۲) در گاوهای هلشتاین و وِریکن و همکاران (۲۹) در جوجه‌های گوشتی با استفاده از یک الگوریتم بازیابی ژنوتیپ مبتنی بر مدل مخفی مارکوف[1] نشان دادند که هرچه تعداد SNPهایی که باید ژنوتیپ آنها بازیابی شود کمتر باشد، صحت بازیابی ژنوتیپ نیز افزایش می‌یابد چرا که به دلیل درصد بالای SNPهای با ژنوتیپ معلوم، میزان خطای بازیابی ژنوتیپ کاهش می‌یابد و تعداد ژنوتیپ‌هایی که به درستی بازیابی می‌شوند افزایش می‌یابد.

از نظر عملکرد بازیابی ژنوتیپ، روش KNN با روش‌های مبتنی بر یادگیری ماشین مانند جنگل تصادفی (RF) و تجزیه مقدار تکین (SVD) قابل مقایسه است. در سناریو مشابه از تعداد فرد، تعداد کروموزوم و تعداد نشانگر، غفوری کسبی و و همکاران (۹) اطلاعات ژنوتیپی به ترتیب ۵، ۱۰، ۲۵، ۵۰، ۷۵ و ۹۰ درصد از SNPهای ۲۵۰ فرد را از ماتریس ژنوتیپی افراد حذف نموده و مجدداً توسط روش RF بازیابی نمودند و صحت بازیابی ژنوتیپ را در سناریوهای فوق به ترتیب ۸۵، ۸۴، ۷۸، ۷۱، ۶۱ و ۵۶ درصد گزارش کردند که این نتایج عملکرد بهتر روش RF را نسبت به KNN نشان می‌دهد. همچنین غفوری کسبی و گودرز تله جردی (۸) اطلاعات ژنوتیپی به ترتیب ۵، ۱۰، ۲۵، ۵۰، ۷۵ و ۹۰ درصد از SNPها را از ماتریس ژنوتیپی حذف نموده و مجدداً توسط روش SVD بازیابی نمودند و صحت بازیابی ژنوتیپ را به ترتیب ۸۰، ۸۰، ۷۹، ۷۷، ۷۰ و ۴۸ درصد گزارش نمودند که در سناریو حذف ۵ تا ۵۰ درصد به نتایج حاصل از روش مورد بررسی در این تحقیق نزدیک هستند (به‌ترتیب ۸۲، ۸۲، ۸۰، ۷۶، ۶۲ و ۴۰ درصد). با توجه به این گزارشات و نتیجه تحقیق حاضر، از نظر عملکرد کلی می‌توان روش RF را در رتبه اول و دو روش SVD و KNN را در رتبه دوم قرار داد.به طور کلی برای یک جایگاه هر چه تعداد افرادی که ژنوتیپ آنها معلوم است بیشتر باشد، اطلاعات ژنوتیپ‌های نامعلوم با صحت بالاتری بازیابی می‌شوند. این امر در شکل ۲ نشان داده شده است که در آن تعداد افراد از ۲۵۰ فرد به ۱۰۰۰ فرد افزایش یافته است. در این حالت، در سناریوهای ۵، ۱۰، ۲۵، ۵۰، ۷۵ و ۹۰ درصد ژنوتیپ از دست رفته روش KNN با صحتی برابر به ترتیب ۰/۸۳، ۰/۸۳، ۰/۸۲، ۰/۸۲، ۰/۷۱ و ۰/۵۴ درصد ژنوتیپ‌های از دست رفته را بازیابی نمود.



شکل ۱- صحت بازیابی ژنوتیپ در درصدهای مختلف از ژنوتیپ‌های از دست رفته (جمعیت شامل ۲۵۰ فرد و ۱۰۰۰ نشانگر)
Figure 1. Imputation accuracy in different percent of missing genotypes (population includes 1000 individuals and 1000 markers).

جدول ۲- همبستگی بین ماتریس ژنوتیپی اولیه و ماتریس ژنوتیپی بازیابی شده در درصدهای مختلف از ژنوتیپ‌های از دست رفته
Table 2. Correlation between original and imputed genotypic matrix in different percent of missing genotypes

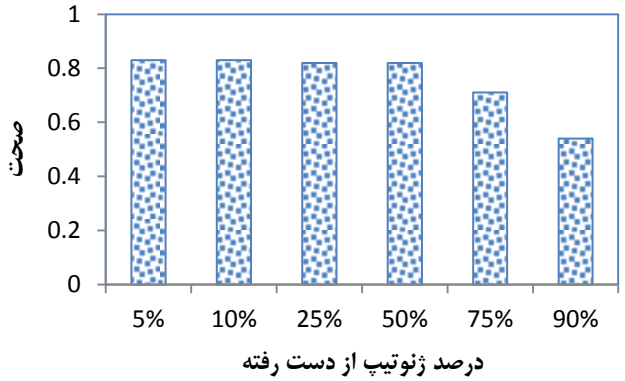
	درصد ژنوتیپ از دست رفته					
	٪۵	٪۱۰	٪۲۵	٪۵۰	٪۷۵	٪۹۰
همبستگی	۰/۹۹۸	۰/۹۷۴	۰/۹۴۱	۰/۹۳۶	۰/۸۵۵	۰/۷۲۴

همان‌گونه که مشاهده می‌شود خصوصاً در سناریوهای ۷۵ و ۹۰ درصد ژنوتیپ از دست رفته صحت بازیابی در مقایسه با شکل ۱ افزایش یافته است. افزایش در صحت بازیابی ژنوتیپ در نتیجه افزایش تعداد افراد جمعیت که در مطالعه حاضر مشاهده شد با تحقیقات دیگر تطابق است. برای مثال پی و همکاران (۱۸) در یک مطالعه شبیه‌سازی و با استفاده از

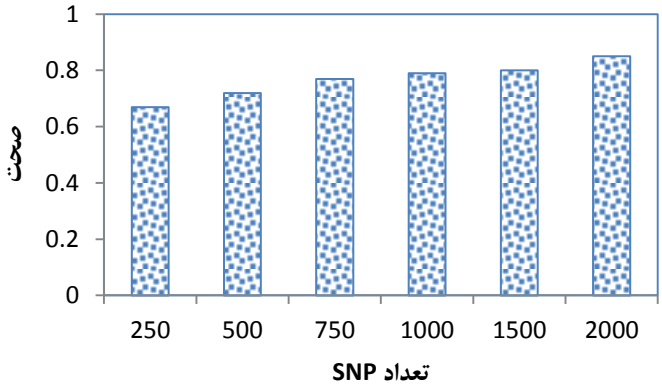
1- Hidden Markov Model

روش‌های بازیابی ژنوتیپ مبتنی بر خوشه‌بندی هاپلوتایپی، زنجیره مخفی مارکوف و مدل مخفی مارکوف گزارش کردند که در سناریوهای مختلف از تراکم نشانگری، با افزایش اندازه جمعیت از ۵۰ به ۴۵۰ صحت بازیابی ژنوتیپ به میزان ۵ درصد افزایش یافت. روشیارا و شولتز (۱۹) و روشیارا و همکاران (۲۰) با استفاده از شبیه‌سازی و همچنین اطلاعات مربوط به تراشه‌های SNP در انسان تأثیر اندازه جمعیت بر صحت بازیابی ژنوتیپ را بررسی کرده و گزارش نمودند که در عمده روش‌های بازیابی ژنوتیپ، با افزایش اندازه جمعیت از ۴۰ به ۲۵۰۰ نفر، صحت بازیابی ژنوتیپ افزایش یافت. حیدری تبار و همکاران (۱۰) نیز افزایش در تعداد افراد را به عنوان یک راهکار مؤثر برای افزایش صحت بازیابی ژنوتیپ پیشنهاد دادند خصوصاً زمانی که افراد جمعیت با تراشه‌های SNP با تراکم پایین تعیین ژنوتیپ شده باشند. با افزایش اندازه جمعیت تعداد افراد دارای ژنوتیپ معلوم برای جایگاه مورد نظر افزایش می‌یابد و در نتیجه آن احتمال اینکه ژنوتیپ از دست رفته به درستی بازیابی شود افزایش خواهد یافت چرا که قطعیت در تصمیم نهایی در مورد ژنوتیپ صحیح برای جایگاه افزایش می‌یابد.

همانطور که در شکل ۳ مشاهده می‌شود با افزایش تعداد نشانگر توانایی بازیابی ژنوتیپ توسط روش KNN افزایش یافت. در این حالت همبستگی بین ماتریس ژنوتیپی اولیه با ماتریس ژنوتیپی بازیابی شده نیز افزایش یافت (جدول ۳) به نحوی که در سناریو ۲۵۰ SNP این همبستگی ۰/۹۳۶ بود اما با افزایش تعداد SNP همبستگی به ۰/۹۸۳ افزایش پیدا کرد که نشاندهنده افزایش تعداد ژنوتیپ به درستی بازیابی شده است. پِی و همکاران (۱۸) نشان دادند که با افزایش تراکم نشانگری از یک نشانگر به ازاء هر ۱۰kbبه یک نشانگر به ازاء هر ۳kb، صحت بازیابی از ۷۲ درصد به ۸۳ درصد افزایش می‌یابد. وِریکن و همکاران (۲۹) صحت بازیابی ژنوتیپ زیرمجموعه‌های با تعداد SNP مختلف را مورد مطالعه قرار دادند و گزارش کردند که هرچه تعداد SNP در نمونه بیشتر باشد، صحت بازیابی افزایش خواهد یافت و درصد SNPهایی که به درستی بازیابی می‌شوند افزایش می‌یابد. در مطالعه دیگر، شروتن و همکاران (۲۳) در گونه گاو نشان دادند که افزایش تعداد نشانگر به ازاء هر سانتی مورگان منجر به افزایش صحت بازیابی ژنوتیپ خواهد شد. همچنین، غفوری کسبی و گودرز تله جردی (۸) نشان دادند در یک درصد ثابت از ژنوتیپ‌های از دست رفته، با افزایش تعداد نشانگر صحت بازیابی ژنوتیپ افزایش می‌یابد به نحوی که با افزایش تعداد نشانگر از ۵۰۰ به ۳۰۰۰ نشانگر، صحت بازیابی ژنوتیپ ۱۰ درصد افزایش یافت. اگر چه با افزایش تعداد افراد و یا تعداد نشانگر صحت بازیابی ژنوتیپ افزایش می‌یابد اما این مساله به قیمت افزایش زمان محاسباتی به دست می‌آید. همانطور که در جدول ۴ مشاهده می‌شود در زمانی که ماتریس ژنوتیپی شامل ۲۵۰ فرد و ۲۵۰ SNP است محاسبات در ۲۶ ثانیه انجام شد. اما وقتی ابعاد ماتریس ژنوتیپی به ۱۰۰۰ فرد و ۱۰۰۰ SNP افزایش یافت زمان انجام محاسبات ۴۰ دقیقه بود. به نظر می‌رسد با افزایش ابعاد ماتریس ژنوتیپی مدت زمان انجام محاسبات به صورت تصاعدی افزایش می‌یابد. اگر ماتریس ژنوتیپی مورد بررسی بزرگ باشد مدت زمان انجام محاسبات می‌تواند چندین ساعت یا چند روز باشد که این مساله می‌تواند بازدهی کلی روش KNN را تحت تاثیر قرار دهد بویژه زمانی که نتیجه بازیابی ژنوتیپ باید هرچه سریعتر در دسترس محقیقین قرار گیرد. البته استفاده از کامپیوترهای با توانایی محاسباتی بالا مانند سرورهای تحقیقاتی می‌تواند این مشکل را حل کند.



شکل ۲- تاثیر افزایش تعداد افراد از ۲۵۰ به ۱۰۰۰ فرد بر صحت بازیابی ژنوتیپ در درصدهای مختلف از ژنوتیپ‌های از دست رفته
Figure 2. The effect of increasing the number of individuals from 250 to 1000 on accuracy of genotype imputation in in different percent of missing genotypes



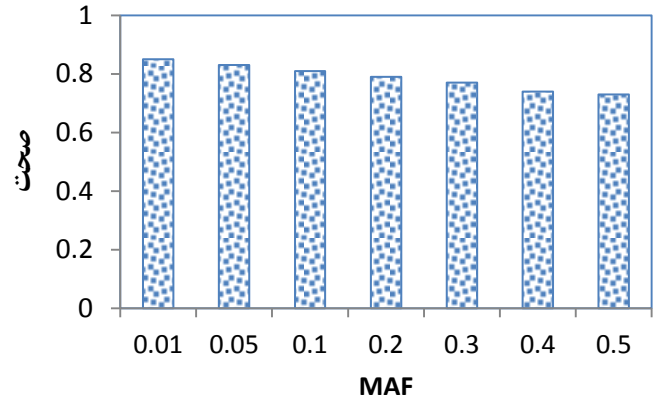
شکل۳- تأثیر تعداد نشانگر بر صحت بازیابی ژنوتیپ (جمعیت شامل ۱۰۰۰ فرد)
Figure 3.The effect of number of marker on accuracy of imputation (population includes 1000 individuals)

جدول ۳- همبستگی بین ماتریس ژنوتیپی اولیه و ماتریس ژنوتیپی بازیابی شده (تعداد فرد برابر ۱۰۰۰ و نرخ خذف ۵۰٪)
Table 3. Correlation between original and imputed genotypic matrix (population includes 1000 individuals and missing rate 50%)

	تعداد SNP					
	۲۵۰	۵۰۰	۷۵۰	۱۰۰۰	۱۵۰۰	۲۰۰۰
همبستگی	۰/۹۳۶	۰/۹۴۲	۰/۹۵۱	۰/۹۵۸	۰/۹۷۱	۰/۹۸۳

جدول۴- زمان انجام محاسبات در روش KNN در سناریوهای مختلف از تعداد فرد و تعداد SNP (نرخ حذف برابر ۵۰٪)
Table 4. Computing time for KNN in different scenarios of number of individuals and number of SNP (missing rate eq to 50%)

ابعاد ماتریس ژنوتیپی	۲۵۰×۲۵۰	۵۰۰×۵۰۰	۷۵۰×۷۵۰	۱۰۰۰×۱۰۰۰
زمان	۲۶ ثانیه	۳ دقیقه	۱۵ دقیقه	۴۰ دقیقه

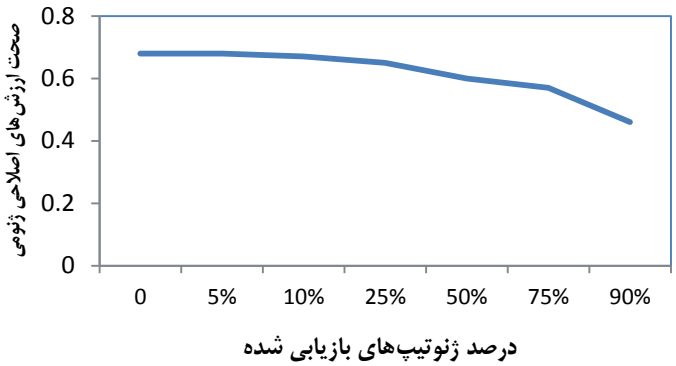


شکل ۴- تأثیر سطوح مختلف MAF بر صحت بازیابی ژنوتیپ
Figure 4.The effect of MAF on accuracy of imputation

در شکل ۴ تأثیر سطوح مختلف MAF بر صحت بازیابی ژنوتیپ را نشان می‌دهد (نسبت حذف ژنوتیپ‌ها در همه سطوح MAF ثابت و به میزان ۵۰٪ بود). با افزایش MAF از ۰/۰۱ به ۰/۵ میزان صحت بازیابی ژنوتیپ از ۰/۸۵ به ۰/۷۳ کاهش یافت که وجود یک رابطه معکوس بین میزان MAF و صحت بازیابی ژنوتیپ را نشان می‌دهد. لین و همکاران (۱۲) و وِنگ و همکاران (۳۲) سطوح مختلف MAF را بر صحت بازیابی ژنوتیپ بررسی کرده و گزارش کردند که افزایش MAF به بیشتر از ۵٪ کاهش صحت بازیابی ژنوتیپ را در پی خواهد داشت. با افزایش MAF درصد ژنوتیپ هتروزیگوت برای جایگاه‌های مختلف افزایش می‌یابد. در حالت اخیر معمولاً در اکثر روش‌های بازیابی ژنوتیپ صحت بازیابی به دلیل کاهش عدم قطعیت تصمیم نهایی در مورد نوع ژنوتیپ برای جایگاه‌ها کاهش می‌یابد (۸).

تأثیر درصدهای مختلف از ژنوتیپ‌های بازیابی شده بر صحت پیش‌بینی ارزش‌های اصلاحی ژنومی در شکل ۵ نشان داده شده است. همانطور که مشاهده می‌شود در حالت از دست رفتن اطلاعات ژنوتیپی ۵٪ از SNPها و بازیابی آنها،

صحت ارزش‌های اصلاحی ژنومی برآورد شده با زمانی که ارزش‌های اصلاحی ژنومی با استفاده از اطلاعات معلوم همه SNPها پیش‌بینی می‌شود برابری می‌کند. در حالت بازیابی اطلاعات ۲۵٪ از ژنوتیپ‌ها و پیش‌بینی ارزش‌های اصلاحی ژنومی با استفاده از آنها نیز صحت پیش‌بینی با صحت پیش‌بینی ارزش‌های اصلاحی ژنومی با استفاده از اطلاعات معلوم تقریباً برابر است. با افزایش درصد ژنوتیپ‌های بازیابی شده به بیش از ۲۵ درصد، صحت پیش‌بینی نیز کاهش می‌یابد به طوری که در حالتی که ۹۰٪ اطلاعات ژنوتیپی بازیابی شدند، حداکثر کاهش در صحت پیش‌بینی ارزش‌های اصلاحی ژنومی دیده شد. در حالت اخیر ۳۲٪ کاهش در صحت پیش‌بینی ارزش‌های اصلاحی ژنومی مشاهده شد. نتایج مشابه توسط پی و همکاران (۱۸) و هیکی و همکاران (۱۱) گزارش شده است. علت این مساله این است که با افزایش درصد ژنوتیپ‌های از دست رفته میزان خطا در بازیابی ژنوتیپ افزایش می‌یابد و این مساله در نهایت منجر به کاهش صحت ارزیابی ژنومی می‌شود.



شکل ۵- تأثیر درصدهای مختلف از ژنوتیپ‌های بازیابی شده بر صحت پیش‌بینی ارزش‌های اصلاحی ژنومی
Figure 5. The effect of different percent of imputed genotypes on the accuracy of predicted genomic breeding values

نتیجه‌گیری کلی

عواملی مانند تعداد افراد در جمعیت، تعداد نشانگر و فراوانی آلل نادر صحت بازیابی ژنوتیپ را تحت تاثیر قرار دادند. مدت زمان بازیابی ژنوتیپ نیز با افزایش ابعاد ماتریس ژنوتیپی به صورت تصاعدی افزایش یافت. با افزایش درصد ژنوتیپ‌های بازیابی شده، صحت پیش‌بینی ارزش‌های اصلاحی ژنومی کاهش یافت. با توجه به نتایج تحقیق حاضر و گزارشات قبلی در مورد روش‌های هم خانواده KNN، به نظر می‌رسد به دلیل صحت بالاتر، استفاده از روش جنگل تصادفی بر روش KNN ارجحیت داشته باشد.

منابع

1. Ahmadi, Z. 2020. Comparison of efficiency of rrBLUP-method6 approach with some common methods in genomic evaluation of livestock. MSc thesis, Bu-Ali Sina University, Hamedan, Iran.
2. Berry, D.P. and J.F. Kearney. 2011. Imputation of genotypes from low- to high-density genotyping platforms and implications for genomic selection. Animal, 5: 1162-1169.
3. Bidder, O.R., H.A. Campbell, A. Gómez-Laich, P. Urgé, J. Walker, Y. Cai and R.P Wilson. 2014. Love thy neighbour: automatic animal behavioral classification of acceleration data using the k-nearest neighbour algorithm. PLoS ONE, 9: 1-7.
4. Cleveland, M.A. and J.M. Hickey. 2013. Practical implementation of cost-effective genomic selection in commercial pig breeding using imputation. Journal of Animal Science, 91: 3583-3592.
5. De los Campos, G. and P. Perez Rodriguez. 2018. Bayesian Generalized Linear Regression. Available at: https://cran.r-project.org/web/packages/BGLR/index.html.
6. Elshire, R.J., J.C. Glaubitz, Q. Sun, J.A. Poland and K. Kawamoto. 2011. A robust, simple genotyping-bysequencing (GBS) approach for high diversity species. PLOS ONE, 6: e19379.
7. Eskelson, B.N.I., H. Temesgen, V. Lemay, T.M. Barrett, N.L. Crookston and A.T. Hudak. 2009. The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases. Scandinavian Journal of Forest Research, 24: 235-246.
8. Ghafouri-Kesbi, F. and A. Goudarztalejerdi. 2018. Studying the performance of intelligent singular value decomposition algorithm (svd) in imputation of missing genotypes in different scenarios of number of marker, population size and minor allele frequency. Iranian Journal of Animal Science Research, 4: 553-560.
9. Ghafouri-Kesbi, F., F. Vanaei, P. Zamani and A. Ahmadi. 2021. Evaluating the performance of Random Forest in imputation of missing genotypes. The first Symposium of Research Highlights in Animal Breeding. Urmia University, Urmia, Iran.

10. Heidaritabar, M., M.P.L. Calus, A. Vereijken, A. Martien, M. Groenen and J.W.M. Bastiaansen. 2015. Accuracy of imputation using the most common sires as reference population in layer chickens. BMC Genetics, 16: 101.
11. Hickey, J.M., J. Crossa, R. Babu and G. de los Campos. 2012. Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. Crop Science, 52: 654-663.
12. Lin, P., S.M. Hartz, Z. Zhang, S.F. Saccone and J. Wang. 2010. A new statistic to evaluate imputation reliability. PLoS ONE, 5: e9697.
13. Liu, C.H., C.F, Tsa, K.L. Sue and M.W. Huang. 2020. The feature selection effect on missing value imputation of medical datasets. Applied Science, 10: 2344.
14. Mahmoud, H.A., E. Hdadad, F.A. Mousa and A. Hassanien. 2015. Cattle classifications system using fuzzy k- nearest neighbor classifier. International Conference on Informatics, Electronics and Vision (ICIEV), Fukuoka, Japan.
15. Meuwissen T.H., B.J. Hayes and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics, 157: 1819-29.
16. Mohammadi, Y. and M. Sattaei-Mokhtari. 2018. Accuracy of genomic breeding values in small genotyped populations-A simulation study. Research on Animal Production, 9: 123-128 (In Persian).
17. Mohammadi, Y. and J. Ahmadpanah. 2021. Effect of reference population size and imputation methods on the accuracy of imputation in pure and mixed populations. Research on Animal Production, 11: 109-114 (In Persian).
18. Pei, Y.F., J. Li, L. Zhang, C.J. Papasian and H.W. Deng. 2008. Analyses and comparison of accuracy of different genotype imputation methods. PLoS ONE, 3: e3551.
19. Roshyara, N.B. and M. Scholz. 2015. Impact of genetic similarity on imputation accuracy. BMC Genetics,16: 90.
20. Roshyara, N.R., K. Horn, H. Kirsten, P. Ahner and M. Scholz. 2016. Comparing performance of modern genotype imputation methods in different ethnicities. Scientific Reports, 6: 34386.
21. Salau, J., J.H. Haas, W. Junge and G. Thaller. 2020. Determination of body parts in holstein friesian cows comparing neural networks and k nearest neighbour classification. Animal, 29: 50.
22. Schwender, H. 2012. Imputing missing genotypes with weighted *k* nearest neighbors. Journal of Toxicology and Environmental Health, 75: 438-446.
23. Schrooten, C., R. Dassonneville, V. Ducrocq, R.F. Brøndum, M.S. Lund and J. Chen. 2014. Error rate for imputation from the Illumina BovineSNP50 chip to the Illumina BovineHD chip. Genetic Selection Evolution, 46: 10.
24. Shen, W., F. Cheng, Y. Zhang, X. Wei, Q. Fu and Y. Zhang. 2019. Automatic recognition of ingestive-related behaviors of dairy cows based on triaxial acceleration. Information Processing in Agriculture, 7: 427-443.
25. Su, G., R.F. Brøndum., P. Ma, B. Guldbrandtsen, G.P. Aamand and M.S. Lund. 2012. Comparison of genomic predictions using medium-density (~54,000) and high-density (~777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations. Journal of Dairy Science, 95: 4657-4665.
26. Technow, F. 2013. hypred: Simulation of genomic data in applied genetics. Available at: http://cran.rproject.org/web/packages/hypred/index.html.
27. Troyanskaya, O., M. Canto, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein and R.B. Altman. 2001. Missing value estimation methods for DNA microarrays. Bioinformatics, 17: 520-525.
28. VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. Journal of Dairy Science, 91: 4414-4423.
29. Vereijken, A.L.J., G.A.A. Albers and J. Visscher. 2010. Imputation of SNP genotypes in chicken using a reference panel with phased haplotypes. 10th World Conference of Genetics Applied on Livestock Production, 407, Germany.
30. Wang, Y., Z. Cai, P. Stothard, S. Moore, R. Goebel, L. Wang and L. Guohui. 2012. Fast accurate missing SNP genotype local imputation. BMC Research Notes, 5: 404.
31. Wang, J., M. Bell, X. Liu and G. Liu. 2020. Machine-learning techniques can enhance dairy cow estrus detection using location and acceleration data. Animals, 10: 1160.
32. Weng Z., Z. Zhang, X. Ding, W. Fu, P. Ma, C. Wang and Q. Zhang. 2012. Application of imputation methods to genomic selection in Chinese Holstein cattle. Journal of Animal Science and Biotechnology, 3: 6.
33. Wu, D., Q. Wu, X. Yin, B. Jiang, H. Wang, D. He and H. Song. 2020. Lameness detection of dairy cows based on the YOLOv3 deep learning algorithm and a relative step size characteristic vector. Biosystems Engineering, 189: 150-163.

Imputation of Missing Genotypes with Intelegent K-Nearest Neighbore Algorithm

Fatemeh Vanaei[1], Farhad Ghafouri-Kesbi[2], Ahmad Ahmadi[3] and Pouya Zamani[4]

1- M.Sc. Student, Department of Animal Science, Faculty of Agriculture, Bu-Ali Sina University, Hamedan, Iran
2- Assistant Professor, Department of Animal Science, Faculty of Agriculture, Bu-Ali Sina University, Hamedan, Iran, (Corresponding author: farhad_ghy@yahoo.com)
3- Assistant Professor, Department of Animal Science, Faculty of Agriculture, Bu-Ali Sina University, Hamedan, Iran
4- Associate Professor and, Department of Animal Science, Faculty of Agriculture, Bu-Ali Sina University, Hamedan, Iran

Received: 5 May, 2021 Accepted: 7 December, 2021

Extended Abstract

Introduction and Objective: Genotype imputation in genomic selection schemes has been considered by researchers in recent years because it can reduce the costs of genomic selection without having a negative impact on the accuracy of genomic selection. In the genotype imputation process, markers that their genotypic information has been missed for any reason are imputed using various statistical methods.

Material and Methods: To constructe genotypic matrix, a one morgan genome including one chromosome for 250 and 1000 individuals was simulated on which in different scenarios 250, 500, 750, 1000, 1500 and 2000 single necleotide polymorphismes (SNP) was distributed. In order to create genomic matrix including missing genotypes, genotypic information of respectively, 5%, 10%, 25%, 50%, 75% and 90% of SNPs was masked and then imputed with KNN. The percent of genotypes correctly imputed (the ratio of genotypes correctly imputed to total masked genotypes) as well as the correlation between primary genotypic matrix (no missing genotype) and imputed genotypic matrix were used as imputation accuracy.

Results: In the population including 250 individuals, the accuracy of imputation in the scenarios of 5%, 10%, 25%, 50%, 75% and 90% missing genotypes, were 0.82, 0.82, 0.80, 0.76, 0.62 and 0.40, respectively, but by increasing the size of the population to 1000 individuals, the imputation accuracies as 0.83, 0.83, 0.82, 0.82, 0.71 and 0.54 were obtained which in the scenarios of 75% and 90% of missing genotypes the increase in imputation accuracy was noticable. The correlation between the primary genotype matrix and the imputed genotypic matrix also decreased with increasing percentage of missing genotypes. In a fixed population size, by increasing the number of SNP from 250 to 2000, imputation accuracy increased from 0.67 to 0.84. In addition, an inverse relationship was observed between MAF and imputation accuracy in a way that by increasing MAF from 0.01 to 0.5, imputation accuracy decreased by 15%. Computation time increased following increase in dimension of genotypic matrix. Bu increasing the percent of missing genotypes, the accuracy of predicted genomic breeding values decreased. In the scenarios of 5 and 10% of missing genotypes, no change in accuracy was observed, but in the scenarios of 75 and 90% of the missing genotypes, the accuracy of prediction of breeding values decreased by 16 and 32%, respectively.

Conclusion: In general, imputation accuracy of KNN was acceptable in such a way that up to 50% of missing genotypes, KNN imputed missing genotypes with 80% accuracy and therefore one could recommend this algorithm for genomic selection schems.

Keywords: Genotype imputation, K-nearest neighbor, Minor allel frequency, Single nucleotide polymorphism

[DOR: 20.1001.1.22518622.1401.13.35.9.5]
[DOI: 10.52547/rap.13.35.130]