بهبود کارایی تکمیل داده­های مفقود شده در مطالعه­ی هم­خوانی سراسر ژنوم

زهره کریمی1\* ،‌ زهرا نریمانی۲

1. دانشکده فنی و مهندسی، دانشگاه دامغان، دامغان، ایران
2. دانشکده علوم رایانه و فن آوری اطلاعات، دانشگاه تحصیلات تکمیلی علوم پایه زنجان، زنجان ، ایران

z.karimi@du.ac.ir

چکیده: مطالعه­ی هم­خوانی سراسر ژنوم، از حوزه­های مهم در بیوانفورماتیک در راستای کشف ارتباط بین بیماری­ها و نوع خاصی از تنوع ژنتیکی است. از چالش­های مطالعه­ی هم­خوانی سراسر ژنوم، وجود ژنوتیپ­های گمشده است. روش تکمیل ماتریس با مرتبه­ی پایین از روش­های مشهور و موفق در این کاربرد است. هدف این روش، کمینه­سازی مرتبه­ی ماتریس داده است. از آن­جا که مساله­ی بهینه­سازی مربوطه محدب نیست از تقریب­هایی استفاده می­شود که کارایی لازم را ندارد. این روش­ها در مواجهه با زمانی که تعداد داده­های گمشده زیاد باشد موفق نبوده و نیز به اندازه­ی کافی مقیاس­پذیر نیستند. در این مقاله روش حداقل مربعات بازموزون تکراری که دو چالش بیان شده را در نظر گرفته است، جهت مطالعه­ی هم­خوانی سراسر ژنوم اعمال شده است. این روش روی مجموعه داده­ی استاندارد مورد ارزیابی قرار گرفته و بر اساس ملاک­ خطای فروبینیوس نسبی مورد ارزیابی قرار گرفته است. نتایج بدست آمده کارایی روش پیشنهای در مقایسه با روش­های مشابه را تایید می­کند.

**کلید واژه‌ها:** مطالعه­ی هم­خوانی سراسر ژنوم، تکمیل ماتریس رتبه پایین، جایگذاری داده­های گمشده، روش حداقل مربعات بازموزون تکراری

1- مقدمه

مطالعه­ی هم‌خوانی سراسر ژنوم (GWAS) [[1]](#footnote-1)در دانش ژنتیک، بررسی سراسری ژنوم بر روی مجموعه‌ای از تنوّع‌های ژنتیکی در افراد مختلف است. هدف این بررسی رسیدن به رابطه‌های مشترک، و گونه‌ای هم‌خوانی، همبستگی و وابستگی و ویژگی مشترک در بین دارندگان آن تفاوت ژنتیکی‌ است. این مطالعات معمولاً روی بررسی ارتباط بین وجود چندریختی‌های تک-نوکلئوتیدی (SNP) و ویژگی‌هایی مانند بیماری‌های رایج انسانی، مثل آلزایمر، متمرکز است اما داده‌های به‌دست آمده می‌تواند برای هر ارگانیسم زنده­ی دیگری نیز مورد استفاده قرار گیرد. در مطالعات GWAS، مجموعه داده­ی­ مورد مطالعه عبارت است از اطلاعات مربوط به SNPهای موجود روی دی‌ان‌اِیِ یک جمعیت شرکت‌کننده، که متشکل از دو گروه کنترل و دارای یک فنوتیپ خاص (برای مثال افراد سالم و افراد دارای بیماری) است. با توسعه­ی روش­های با بازدهی بالا[[2]](#footnote-2)، داده­های ژنتیکی زیادی توسط محققان تولید شده و تعداد این داده­ها در حال افزایش است. تعامل با این مجموعه داده­ها منوط به درنظرگرفتن چالش­های آن­ها شامل این موارد است: (1) تعداد زیاد داده­ها، (2) ژنوتیپ­های ناکامل و (3) پنل­های مرجع متعدد و مجموعه داده­های مطالعه­ی متنوع از پلت فورم­های متفاوت با مجموعه­های متفاوت از SNP ها. با توجه به این چالش­ها، به منظور بهبود روش­های موجود، مساله­ی جایگذاری[[3]](#footnote-3) ژنوتیپ (در مورد داده­های گمشده) و حتی توسعه­ی مدل­های ریاضی جدید مورد توجه محققان قرار گرفته است.

 تاکنون روش­های گوناگونی برای مساله­ی جایگذاری ژنوتیپ پیشنهاد شده است. این روش­ها در دو دسته­ی مبتنی بر مدل و تکمیل ماتریس با مرتبه­ی (رتبه) پایین[[4]](#footnote-4) قرار می­گیرند. روش­های مبتنی بر مدل به پنل­های مرجع بستگی دارند. روش تکمیل داده­ها با استفاده از تکمیل ماتریس با مرتبه­ی پایین، محدودیت­های روش­های مبتنی بر مدل را ندارد. تمرکز این مقاله روی جایگذاری داده­های گمشده در GWAS با بهره­گیری از رویکرد تکمیل ماتریس از رتبه پایین است. مساله­ی بهینه­سازی بدست آمده در این روش، یک مساله­ی محدب نیست، لذا راه­های گوناگونی جهت تعامل با این چالش تاکنون پیشنهاد شده است؛ از جمله کمینه کردن نرم nuclear به جای کمینه کردن رتبه ماتریس. حل مساله­ی بهینه­سازی با نرم nuclear به جای مرتبه­ی ماتریس، از نظر محاسباتی کارایی کافی را نداشته و زمانی که تعداد مشاهده­ها کم باشد مناسب نیست. در این مقاله روش حداقل مربعات بازموزون تکراری[[5]](#footnote-5) (IRLS) جهت حل مساله­ی تکمیل ژنوتیپ پیشنهاد می­شود [1]. روش پیشنهادی مقیاس پذیر بوده و برای زمانی که داده­های موجود کم باشد مناسب است.

در ادامه، ابتدا پژوهش­های مربوط به مساله­ی تکمیل ژنوتیپ در بخش 2 بررسی شده و سپس در بخش 3، روش پیشنهادی ارائه شده است. نتایج ارزیابی در بخش 4 ارائه شده و نتیجه­گیری در انتها آمده است.

2- پژوهش های مرتبط

روش­های کنونی جهت تکمیل ژنوتیپ به دو گروه روش­های مبتنی بر مدل و روش­های مبتنی بر تکمیل ماتریس دسته­بندی می­شوند. بسته­های نرم­افزاری گوناگونی جهت تکمیل ژنوتیپ در دسترس هستند [2] fastPHASE،[3, 4, 5]MaCH ، [6, 7] IMPUTE2 ، BEAGLE [8] و Mendel [9]نمونه­هایی از این نرم­افزارها هستند. تمام این بسته­ها مبتنی بر مدل­های احتمالاتی هستند. تکنیک­های آماری مدل مخفی مارکوف و الگوریتم میانگین­گیری-بیشینه­سازی[[6]](#footnote-6) بدین منظور اعمال شده­اند [6, 10]. این روش­ها از الگوهای هاپلوتیپ در پنل­های مرجع جهت پیش­بینی ژنوتیپ­های نامشخص استفاده می­کنند. روش­های مبتنی بر مدل، به میزان در دسترس بودن پنل­های مرجع وابسته هستند و بنابراین کارایی آن­ها محدود به کیفیت پنل­های مرجع انتخابی است. مدل­های آماری کنونی در تعامل با مجموعه داده­های با پنلهای مرجع چندگانه و پنل­های متنوع مورد مطالعه منعطف نیستند.

دسته­ی دوم روش­ها، روش­های مبتنی بر تکمیل ماتریس هستند. در ادبیات یادگیری ماشین، تکمیل ماتریس ابزاری کارا و موثر در کاربردهای زیادی است. هدف تکمیل ماتریس، بازیابی کل یک ماتریس است در حالی که بخشی از آن مشاهده شده است. ایده­ی این مساله، یافتن ساده­ترین ماتریس مطابق با مدخل­های مشاهده شده است و جستجو به یافتن ماتریس از رتبه پایینی با اختلاف خطای مربع کوچک نسبت به مدخل­های مشاهده شده تبدیل شده است. این روش­ها در مقایسه با روش­های مبتنی بر مدل، نیاز به پنل­های مرجع هاپلوتیپ ندارند، مرتبط بودن نمونه­ها را در نظر می­گیرند و پیاده­سازی آنها آسانتر است. Chi و همکاران یک روش تکمیل ماتریس با رتبه کم را برای GWAS پیشنهاد دادند [11]. آنها مساله­ی بهینه­سازی یافتن ماتریس با کمینه رتبه را با یافتن ماتریس با کمینه نرم nuclear جایگزین کرده و آن را با استفاده از الگوریتم نستروف[[7]](#footnote-7) حل کرده­اند [12]. نرم Nuclear برابر جمع مقادیر منفرد[[8]](#footnote-8) است. Jiang و سایرین دو روش یکی بر مبنای تکمیل ماتریس رتبه پایین[[9]](#footnote-9) (LRMC) و دیگری بر اساس عامل بندی خوشه­بندی توام ماتریس[[10]](#footnote-10) (MCCF) پیشنهاد کرده­اند [13]. در روش LRMC نیز مساله­ی نرم nuclear در نظر گرفته شده و با روش امتداد نقطه­ی ثابت [[11]](#footnote-11) (FPCA) حل شده است. حل مساله­ی کمینه کردن نرم nuclear به جای کمینه کردن مرتبه­ی ماتریس، هر چند مساله­ی بهینه­سازی را محدب ساخته و حل آن را ساده­تر می­سازد، اما با چالش­هایی مواجه است. در این مقاله، تمرکز ما روی حل مساله­ی تکمیل ژنوتیپ در GWAS با رویکردی متفاوت از روش­های موجود با هدف مرتفع نمودن چالش­های آنها است. این چالش­ها به همراه روش پیشنهادی در بخش بعد ارائه شده است.

3- روش پیشنهادی

‌ مساله­ی بهینه­سازی تکمیل ماتریس با رتبه­ی پایین به صورت رابطه­ی (1) است که برای تکمیل داده­های مفقود شده بکار می­رود.

|  |  |
| --- | --- |
| (1) |  |

 از درجه­ی r، مجموعه اندیس­ها است و مساله، بازسازی از دانش و عملگر نمونه­برداری یا تصویر[[12]](#footnote-12) است که ماتریس را به اندیس­های نگاشت می­کند.

|  |  |
| --- | --- |
| (2) |  |

 *حل* مساله­ی بهینه­سازی مذکور به دلیل غیرمحدب بودن و غیرهموار بودن مشکل است. معمولاً نرم درجه با نرم nuclear که محدب است جایگزین می­شود [11]. نرم nuclear به صورت رابطه­ی (۳) تعریف می شود:

|  |  |
| --- | --- |
| (3) |  |

این تقریب مساله با دو چالش مواجه است: (1) مساله­ی بهینه­سازی را معادل یک برنامه نیمه معین[[13]](#footnote-13) می­سازد و لذا کارایی لازم را از نظر محاسباتی برای مسائل بزرگ ندارد. (2) در صورتی که تعداد داده­های m کمی بیشتر از درجه آزادی باشد، کمینه­ی نرم nuclear بر کمینه­ی رابطه­ی (1) منطبق نیست. لذا در این مقاله یک روش مقیاس پذیر، نارایب و موثر برای حل مساله­ی بهینه­سازی (1) استفاده می شود. این روش برای زمانی که تعداد ورودی های شناخته شده کم باشد موثر است. روش IRLS بر مبنای پژوهش­های گوناگونی است که نشان داده­اند که بهینه­سازی تابع هدف لگاریتم دترمینان هموار شده برای منجر به راه­حل کمینه­سازی رتبه می­شود [14, 15]. به طور خاص، می توان نشان داد که کمینه کننده تابع هدف لگاریتم دترمینان هموار شده حداقل به اندازه­ی کمینه کننده­ی نرم nuclear محدب با حداقل کننده رتبه منطبق است [16].

 را در نظر بگیرید و تابع یک تابع هدف لگاریتم دترمینان هموار شده باشد که به صورت تعریف می­شود که در آن رابطه­ی (۴)‌ برقرار است.

|  |  |
| --- | --- |
| (4) |  |

 پیوسته­ی مشتق­پذیر است [17]. روشی که برای حل این مساله در نظر گرفته شده است، روش IRLS است. IRLS الگوریتمی تکراری است که گام اصلی در هر تکرار آن حل یک مساله­ی حداقل مربعات وزن­دار است. وزن در تکرار kم، ، به این صورت تعریف می­شود: با در نظرگرفتن و ماتریس داده با تجریه­­ی مقدار منفرد[[14]](#footnote-14) ، یعنی و که ماتریس­های متعامد هستند، عملگر خطی به صورت رابطه­ی (5) تعریف می شود:

|  |  |
| --- | --- |
| (5) |  |

 *ضرب درایه­ای[[15]](#footnote-15) در است و ماتریسی با مقادیر مثبت است که برای هر i=1,…, و j=1,…,. عملگر یک عملگر مثبت خودالحاق[[16]](#footnote-16) با مقادیر ویژه­ی مثبت است.* گام­های الگوریتم MatrixIRLS در شکل 1 آمده است. جهت مطالعه­ی جزییات بیشتر در مورد الگوریتم به [1] مراجعه کنید.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ورودی: اندیس­ها، ، مشاهدات، ، رتبه تخمین زده شده­ی ماتریس، ، K، تعداد دفعات تکرار الگوریتمخروجی: 1. مقداردهی اولیه ، و ، ( و ابعاد ماتریس ورودی هستند)
2. برای k=1 تا K گام های زیر را انجام بده:
	1. مساله­ی حداقل مربعات وزن­دار را حل کن: با استفاده از گرادیان مزدوج[[17]](#footnote-17) مساله­ی (۶) را حل کن :

|  |  |
| --- | --- |
| (6) |  |

* 1. *جمله­ی هموارسازی را بروزرسانی کن : مین مقدار منفرد[[18]](#footnote-18) را محاسبه کرده و بروزرسانی (۷) را انجام بده:*

|  |  |
| --- | --- |
| (7) |  |

* 1. *عملگر وزن را بروزرسانی کن: برای ، اولین مقادیر منفرد و ماتریس­های و که بردارهای منفرد چپ و راست هستند برای بروزرسانی بر اساس رابطه­ی (۵) عمل کن.*
 |
| شکل 1- الگوریتم MatrixIRLS |

4-ارزیابی

جهت ارزیابی، دو مجموعه داده­ی Chr22\_3841و chr22 مورد ارزیابی قرار گرفته­اند. مجموعه داده­ی chr22، 790 سطر و 112 ستون، و مجموعه داده­ی chr22\_3841، 3841 سطر و 2184 ستون دارد. این دو مجموعه داده، مجموعه داده­های استاندارد جهت ارزیابی روش­های تکمیل داده GWAS هستند [16]. محاسبه­ی مقادیر منفرد این مجموعه داده­ها موید رتبه پایین بودن این ماتریس است [13].

ملاک­ ارزیابی، خطای فروبینیوس نسبی است که از رابطه­ی (8) محاسبه می­شود.

|  |  |
| --- | --- |
| (8) | =خطای فروبینیوس نسبی |

 مقدار تخمینی در روش IRLS یا ماتریس تخمینی در سایر روش­ها است. روش پیشنهادی با روش FPCA مقایسه شده است [13]. نتایج مقایسه­ی روش پیشنهادی با روش FPCA برای مجموعه داده­های chr22 و chr22\_3841 به ترتیب در جداول 1 و 2 آمده است. همان­طور که مشخص است نتایج بدست آمده در هر دو مجموعه داده بهبود یافته است.

جدول 1- نتیجه­ی مقایسه­ی روش پیشنهادی با سایر روش ها روی مجموعه داده Chr22\_3841

|  |  |
| --- | --- |
|  | خطای فروبینیوس نسبی |
| روش پیشنهادی | 0/0394 |
| FPCA | 3157/0 |

جدول 2- نتیجه­ی مقایسه­ی روش پیشنهادی با سایر روش ها روی مجموعه داده Chr22

|  |  |
| --- | --- |
|  | خطای فروبینیوس نسبی |
| FPCA | 1890/0 |
| روش پیشنهادی | 0/067 |

5-نتیجه گیری

در این پژوهش، مساله­ی وجود داده­های گمشده در مطالعه­ی همخوانی سراسر-ژنوم مورد بررسی قرار گرفت. راهکارهای موجود برای حل این مساله دو دسته هستند؛ راهکاره مبتنی بر مدل، که نیاز به وجود پنل­های ژنتیکی مرجع برای تکمیل داده­ها دارند، و راهکارهای مبتنی بر تکمیل ماتریس. روش­های مبتنی بر تکمیل ماتریس موجود از نرم unclear برای حل مساله استفاده می­کردند. ناکارامدی این روش­ها و همچنین مشکل بیشتر بودن تعداد داده­ها از درجه آزادی ماتریس داده، باعث می­شد کمینه­ی نرم nuclear دیگر برای حل مسأله مناسب نباشد. در این مقاله، روش حداقل مربعات بازموزون تکراری جهت حل این چالش­ها پیشنهاد شده است. این روش بر مبنای تابع هدف لگاریتم دترمینان هموار شده است و نسبت به روش­های موجود مقیاس­پذیرتر بوده و در برابر وجود تعداد کمی مشاهده مقاوم است. ارزیابی انجام شده روی مجموعه داده­های معیار، نشان دهنده­ی بهبود روش پیشنهادی بر اساس ملاک خطای فروبینیوس نسبی در مقایسه با رویکردهای مشابه است.

**مراجع:**

|  |  |
| --- | --- |
| [1]  | C. Kümmerle و C. M. Verdun, “Escaping Saddle Points in Ill-Conditioned Matrix Completion with a Scalable Second Order Method,” *arXiv preprint arXiv,* p. 2009.02905, 2020.  |
| [2]  | P. Scheet و M. Stephens, “A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase,” *The American Journal of Human Genetics, 78(4),* pp. 629-644, 2006.  |
| [3]  | Y. Li, “ Mach 1.0: rapid haplotype reconstruction and missing genotype inference,” *Am J Hum Genet, 79, 2290,* 2006.  |
| [4]  | Y. Li, C. J. Willer, J. Ding, P. Scheet و G. R. Abecasis, “MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes,” *Genetic epidemiology, 34(8),* pp. 816-834, 2010.  |
| [5]  | Y. Li, C. Willer, S. Sanna و G. Abecasis, “Genotype imputation,” *Annual review of genomics and human genetics, 10,* pp. 387-406, 2009.  |
| [6]  | J. Marchini و B. Howie, “Genotype imputation for genome-wide association studies,” *Nature Reviews Genetics, 11(7),* pp. 499-511, 2010.  |
| [7]  | B. N. Howie, P. Donnelly و J. Marchini, “A flexible and accurate genotype imputation method for the next generation of genome-wide association studies,” *PLoS genetics, 5(6), e1000529,* 2009.  |
| [8]  | B. L. Browning و S. R. Browning, “A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals,” *The American Journal of Human Genetics, 84(2),* pp. 210-223, 2009.  |
| [9]  | K. L. Ayers و K. Lange, “Penalized estimation of haplotype frequencies,” *Bioinformatics, 24(14),* pp. 1596-1602, 2008.  |
| [10]  | S. R. Browning, “Missing data imputation and haplotype phase inference for genome-wide association studies,” *Human genetics, 124(5),* pp. 439-450, 2008.  |
| [11]  | E. C. Chi, H. Zhou, G. K. Chen, D. Vecchyo, D. O و K. Lange, “Genotype imputation via matrix completion,” *Genome research, 23(3),* pp. 509-518, 2013.  |
| [12]  | A. Beck و M. Teboulle, “A fast iterative shrinkage-thresholding algorithm for linear inverse problems,” *SIAM journal on imaging sciences, 2(1),* pp. 183-202, 2009.  |
| [13]  | B. Jiang, M. Shiqian, C. Jason, Q. Linbo, P. H. Matthew, B. Ian, J. Daniel, Z. Shuzhong و H. Xiuzhen, “SparRec: An effective matrix completion framework of missing data imputation for GWAS,” *Scientific reports 6, no. 1,* pp. 1-15, 2016.  |
| [14]  | M. Fazel, “Matrix rank minimization with applications,” *Ph.D. Thesis, Electrical Engineering Department, Stanford University,* 2002.  |
| [15]  | E. J. Candes, Y. C. Eldar, T. Strohmer و V. Voroninski, “Phase retrieval via matrix completion,” *SIAM review, 57(2),* pp. 225-251, 2015.  |
| [16]  | S. Foucart, “Concave Mirsky Inequality and Low-Rank Recovery,” *SIAM J. Matrix Anal. Appl., 39(1),* p. 99–103, 2018.  |
| [17]  | F. Andersson, M. Carlsson و K. M. Perfekt, “Operator-Lipschitz estimates for the singular value functional calculus,” *Proceedings of the American Mathematical Society, 144(5),* pp. 1867-1875, 2016.  |

An Improved data imputation method for genome-wide association studies

\*Zohre Karimi1, Zahra Narimani2

1 School of Engineering, Damghan University, Damghan, Iran

2 Department of Computer Science and Information Technology, Institute for Advanced Studies in Basic Sciences, Zanjan, Iran

z.karimi@du.ac.ir

*Abstract*— One of the challenges in genome-wide association studies is the existence of missing genotypes. Low-rank matrix completion method is one of the well-known and successful methods in this application. The aim of this method is to minimize the rank of the data matrix. Since the relevant optimization problem is not convex, approximations that do not have the required efficiency are used. These methods are not successful in the face of a large number of lost data and are not sufficiently scalable. In this paper, iteratively reweighted least squares method, which resolve these two challenges is applied to genome-wide association studies. This method has been evaluated on two benchmark data sets based on the relative Frobenius error. The results confirm the efficiency of the proposed method compared to similar methods.

Keywords— genome-wide association studies, low rank matrix completion, missing data imputation, iteratively reweighted least squares method.

1. Genome-wide association studies [↑](#footnote-ref-1)
2. High throughput [↑](#footnote-ref-2)
3. Imputation [↑](#footnote-ref-3)
4. Low rank Completion [↑](#footnote-ref-4)
5. Iteratively reweighted least squares [↑](#footnote-ref-5)
6. Expectation-Maximization [↑](#footnote-ref-6)
7. Nesterov [↑](#footnote-ref-7)
8. Singular values [↑](#footnote-ref-8)
9. Low rank matrix completion [↑](#footnote-ref-9)
10. Matrix Co-clustering Factorization [↑](#footnote-ref-10)
11. Fixed point continuation method [↑](#footnote-ref-11)
12. Projection [↑](#footnote-ref-12)
13. Semidefinite Programming [↑](#footnote-ref-13)
14. Singular Value decomposition [↑](#footnote-ref-14)
15. Entry wise [↑](#footnote-ref-15)
16. Self-adjoint [↑](#footnote-ref-16)
17. Conjugate gradient [↑](#footnote-ref-17)
18. Singular value [↑](#footnote-ref-18)