تخمین هاپلوتایپ با استفاده از فاکتورسازی ماتریس رتبه‌پایین در حضور داده‌های پرت

تینا تربالی[۱] سینا مجیدیان[۲] محمد حسین کهایی[۳]

۱- دانشجوی کارشناسی ارشد- دانشکده مهندسی برق- دانشگاه علم و صنعت ایران- تهران- ایران
tarbali.tina@gmail.com

۲- دانشجوی دکتری- دانشکده مهندسی برق- دانشگاه علم و صنعت ایران- تهران- ایران
s_majidian@yahoo.com

۳- دانشیار- دانشکده مهندسی برق- دانشگاه علم و صنعت ایران- تهران- ایران
kahaei@iust.ac.ir

چکیده: تخمین هاپلوتایپ بر اساس اطلاعاتDNA برای کشف بیماری‌های ژنتیکی انسان استفاده می‌شود. این مسئله در پردازش ژنومی سیگنال‌ها به صورت یک ماتریس رتبه‌پایین قابل مدل سازی است که به علت محدودیت‌های موجود در خوانش هاپلوتایپ، فقط تعداد کمی از درایه‌ها مشاهده می‌شوند. در نتیجه یک روش موثر برای بازیابی هاپلوتایپ از مشاهدات ناقص، استفاده از روش-های تکمیل ماتریس است. در این مقاله به کمک روش‌های تکمیل ماتریس، سعی در تخمین هاپلوتایپ از طریق فاکتورسازی ماتریسی شده‌است. در مراجع از روش گرادیان کاهشی برای حل مسئله استفاده شده‌است. اما در روش های قبلی داده‌های پرت نیز در محاسبات وارد می‌شود که باعث خطا در تخمین هاپلوتایپ شده است. به عبارتی در این روش‌ها به شروط موجود برای ماتریس-های هاپلوتایپ توجه نشده است و این موضوع باعث تخمین داده‌های پرت برای هاپلوتایپ شده است. در این مقاله با روش تکمیل ماتریس و با در نظر گرفتن این شروط در ماتریس هاپلوتایپ، یک تابع هزینه جدید به صورت عبارت جریمه برای تخمین هاپلوتایپ معرفی می‌کنیم. عبارت جدید اضافه شده به تابع هزینه باعث می‌شود که اثر داده‌های پرت کاهش یافته و در نتیجه دقت تخمین هاپلوتایپ افزایش می‌یابد. نتایج شبیه سازی نیاز کاهش خطای بازیابی هاپلوتایپ را تایید می‌کند.

واژه‌های کلیدی: تخمین هاپلوتایپ، ماتریس رتبه پایین، تکمیل ماتریس

نوع مقاله: پژوهشی

DOI: 10.29252/jiaeee.18.3.1007

تاریخ ارسال مقاله: ۱۳۹۸/۰۷/۲۸
تاریخ پذیرش مشروط مقاله: ۱۳۹۹/۰۷/۱۳
تاریخ پذیرش مقاله: ۱۳۹۹/۰۹/۱۷
نام نویسنده‌ی مسئول: دکتر محمد حسین کهایی
نشانی نویسنده‌ی مسئول: ایران – تهران – بزرگراه رسالت – خیابان هنگام – دانشگاه علم و صنعت ایران– دانشکده‌ی مهندسی برق

۱- مقدمه

به علت ارتباط بسیار زیاد بین انواع بیماری‌ها با مسائل ژنتیکی، استخراج اطلاعات از داده‌های ژنتیکی از جمله داده‌های پزشکی و رشتهDNA [۱]، نقش مهمی دارد. همچنین به منظور ایجاد توانایی در تشخیص الگوی بیماری‌ها، باید از داده‌های ژنتیکی در دسترس استفاده کرد. به همین علت رشته DNA به یک ابزار همه جانبه و یک عامل کلیدی برای علم پزشکی تبدیل شده‌است. رشته DNA معمولا مجموعه داده‌های بسیار زیادی را تولید می‌کند که پردازش آن‌ها چالش‌های محاسباتی پیچیده‌ای را ارائه می‌دهد. در نتیجه یافتن روش مناسب برای استفاده از این داده‌ها، بسیار مورد توجه است.

مسئله‌ی تخمین هاپلوتایپ با استفاده از مدل‌سازی ماتریسی به عنوان یکی از زمینه‌های تحقیقاتی بین رشته‌ای مطرح شده است. تخمین هاپلوتایپ به کشف ارتباط داده ژنتیکی و بیماری‌ها کمک کرده که منجر به طراحی دارو و روش‌های درمانی مبتنی بر فرد می‌گردد [۴].

برای تخمین هاپلوتایپ از روش‌هایی مانند الگوریتم انتقال پیام [۵] و SDP [۶] استفاده شده است. در این روش‌ها از گراف برای مدل‌سازی مسئله استفاده شده است. در صورت وجود نویز، امکان تخمین هاپلوتایپ به کمک الگوریتم‌های بیان شده میسر نیست. همچنین میتوان از روش‌های پردازش سیگنال [۳] دیجیتال برای تخمین هاپلوتایپ استفاده کرد. این روش‌ها در سال‌های اخیر برای تجزیه و تحلیل داده-های ژنتیکی در حال توسعه هستند. از جمله مهم‌ترین مفاهیم مورد استفاده در این روش برای داده‌های ژنتیکی میتوان به تبدیل فوریه گسسته [۴]، فیلتر دیجیتال، تبدیل موجک دیجیتال [۵] اشاره کرد [۷].

یکی از روش‌های تخمین هاپلوتایپ روش تکمیل ماتریس است [۴]. روش‌های مختلفی برای تکمیل ماتریس از جمله کمینه‌سازی متناوب،ADMiRa [۶]،SET [۷] و AP [۸] مطرح می‌شود [۸-۱۰]. یکی از روش‌های تکمیل ماتریس استفاده از فاکتورسازی ماتریسی است که ماتریس رتبه پایین به صورت حاصل‌ضرب دو ماتریس با ابعاد کمتر تجزیه و در هر مرحله با ثابت فرض شدن یکی از ماتریس‌ها مسئله نسبت به ماتریس دیگر کمینه می‌شود و محدب بودن هر دو ماتریس تجزیه شده به حل مسئله کمک شایانی می‌کند [۱۱, ۱۲].

یکی از روش‌های تخمین هاپلوتایپ، حل مسئله به کمک فاکتورسازی ماتریسی است. استفاده از این روش در پردازش سیگانال و تخمین پارامترها بسیار متداول است [۱۳]. اما در الگوریتم‌های پیاد شده به کمک این روش امکان حضور داده‌های پرت در جواب مسئله وجود دارد. این امر به این علت است که ویژگی‌های موجود در درایه‌های ماتریس تجزیه شده هاپلوتایپ در مراحل حل مسئله بی اثر بوده است.

در این مقاله به کمک روش‌های تکمیل ماتریس، سعی در تخمین هاپلوتایپ از طریق فاکتورسازی ماتریسی شده است. با توجه به اینکه داده‌های ژنتیکی علاوه بر رتبه پایین بودن ماتریس مشاهدات، دارای ویژگی‌هایی در ساختار ماتریس‌ها است، تابع هزینه جدید به صورت عبارت جریمه برای مسئله بیان می‌شود. تخمین هاپلوتایپ به کمک الگوریتم گرادیان کاهشی قابل حل است. عبارت جدید اضافه شده به مسئله، به تخمین هاپلوتایپ با دقت بیشتر کمک کرده و بازیابی را بهبود بخشیده است. زیرا عبارت تنظیم اضافه شده به تابع هزینه باعث کنترل درایه‌های ماتریس و عدم انتخاب داده پرت می‌شود. نتایج حاصل از شبیه‌سازی موید این مطلب است.

۲- بیان مسئله

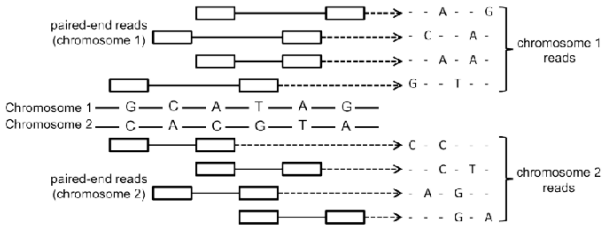
۲-۱- معرفی مفاهیم زیستی

نوکلئیک اسید یکی از اجزای تشکیل‌دهنده‌ی هسته‌ی سلول انسان است که بر دو نوع ریبونوکلئیک اسید [۹]و دئوکسی‌ریبونوکلئیک اسید [۱۰] یافت می‌شود. DNA رشته‌ای از نوکلئیک اسیدها است که شامل بسیاری از اجزای کوچکتر مرتبط با نوکلئوتید می‌باشد. هر نوکلئوتید شامل یکی از چهار اسیدهای آمینه ممکن یعنی آدنین [۱۱] ، تیامین [۱۲] ، سیتوزین [۱۳] و گوانین [۱۴] است. DNA دارای ساختار دو رشته‌ای است که در جهت مخالف هم اجرا می‌شوند. به طور مثال در انسان تعداد کروموزوم‌ها [۱۵] ۴۶ عدد هستند که شامل دو سری یکسان ۲۳ تایی هستند که هر کدام از آن‌ها از یکی از والدین به ارث رسیده‌است. کروموزوم جفت را کروموزم‌های همسان [۱۶] می‌گویند. مجموع DNA در هر سری از این ۲۳ کروموزوم شامل حدود ۳میلیارد جفت باز می‌باشد. جانداری که دو نسخه‌ی یک DNA را داشته باشد اصطلاحا دیپلوئید [۱۷] نامیده می‌شود. DNA وظیفه‌ی انتقال و ذخیره اطلاعات ژنتیکی را بر عهده دارد. به عبارتی اطلاعات بنیادی درباره‌ی نحوه‌ی زندگی یک موجود زنده در DNA قرار دارد. شایع‌ترین تغییرات بین دوکروموزوم در یک جفت همولوگ عبارتند از چند ریختی تک نوکلئوتیدی [۱۸] که اسنیپ خوانده می‌شود. یک اسنیپ جایگاهی از ژن است که در بین افراد جامعه، به کمک نوکلئوتیدهای مختلفی مشاهده می‌شود. در واقع به جایگاه ژن هنگام مقایسه دو رشته، اسنیپ گفته می‌شود و برای راحتی می‌توان فقط محل اسنیپ را نمایش داد. از طرفی رشته‌ی اسنیپ بر روی هر یک از جفت کروموزوم‌ها هاپلوتایپ نامیده می‌شود. بنابراین، تغییرات بین دو کروموزوم در یک جفت توسط هاپلوتایپ نشان داده می‌شود [۱۴, ۱۵].

برای به دست آوردن دنباله‌ی هاپلوتایپ، می‌توان از روش‌های توالی-یابی استفاده کرد. اولین روش توالی‌یابی توسط ماکسام [۱۹] و گیلبرت [۲۰] انجام شد [۱۶]. در ادامه فردریک سنگر [۲۱] روشی مبتنی بر خاصیت فسفر سانس ارائه کرد [۱۷]. هزینه‌ی بالاو دقت کم این روش‌ها باعث شد که از روش‌های دیگری از جمله روش توالی شاتگان [۲۲] استفاده شود. در این روش ژنوم به طور تصادفی در نقاط مختلف بریده می‌شود که به توالی هرکدام از این تکه‌ها، خوانش [۲۳] می‌گویند. به بیان دیگر در روش شاتگان از هم‌پوشانی [۲۴]قطعات مختلف خوانش، رشته DNA

حاصل شده است. به علت انتخاب تصادفی این قطعات، بهتر است که هر نقطه از ژنوم توسط تعداد معینی از قطعات هم‌پوشانی شود تا احتمال خطا در این نتیجه‌گیری کم شود [۱۸].

روش توالی‌یابی شاتگان قادر به خوانش رشته DNA از طریق جفت-انتها[۲۵] است. این روش مجموعه‌ای از خوانش‌ها را تولید می‌کند که در آن هر خوانش به مجموعه‌ای از اطلاعات جزئی در مورد کروموزوم که از آن تولید شده است، مربوط می‌شود. در مواردی که تخمین هاپلوتایپ مورد نظر است این هم‌پوشانی و مقایسه به تفاوت جایگاه بین کروموزوم‌ها می‌پردازد. در این روش بدون دانستن محتوای نوکلئوتید و فقط با شناختن طول، نمونه‌برداری صورت می‌گیرد. به عبارت دیگر یک خوانش با طول مشخص در دسترس است که ابتدا و انتهای آن مشخص است اما میان آن مشخص نیست [۴, ۱۹]. شکل (۱) نشان‌دهنده‌ی خوانش جفت‌انتها برای بازیابی هاپلوتایپ است.



شکل (۱): چگونگی خوانش جفت کروموزوم‌ها [۴]

در شکل (۱) برای هر کروموزوم چهار خوانش انجام شده‌است و در هر خوانش دو جایگاه مشاهده شده و نوکلئوتید مربوط به آن جایگاه ارائه شده است و بقیه جایگاه‌ها با '-' نمایش داده شده‌اند. منظور از '-' آن است که خوانش، اطلاعاتی درباره‌ی آن جایگاه نوکلئوتید ارائه نکرده است. برای توضیح بیشتر شکل می‌توان توالی کروموزوم ۱ را در نظر گرفت. کروموزوم ۱ شامل نوکلئوتیدهای GCATAG است که در خوانش اول نوکلئوتیدهای دو جایگاه ارائه شده است که به ترتیب A و Gرا نمایش می‌دهد و دیگر جایگاه‌های آن با '-' نمایش داده شده است که نشان از عدم خوانش کروموزوم است. به همین ترتیب خوانش دوم نیز دو جایگاه کروموزوم ۱ را نشان داده شده است که به ترتیب نوکلئوتید C و A ارائه شده است. همین روند برای هر دو کروموزوم و تمام خوانش‌ها انجام شده است و شکل (۱) نشان از اطلاعات این خوانش‌ها دارد.

۲-۲- مدل ماتریس و ساختار مسئله

برای استفاده از روش‌های محاسبات مهندسی و پردازشی نیاز است که رشته DNA به عدد تبدیل شود [۱, ۱۷]. یعنی به هریک از نوکلئوتید اسیدهای آدنین، سیتوزین، گوانین و سیتوزین یک عدد نسبت داده - شود. انتخاب مناسب این اعداد باعث بهبود عملکرد روش‌های حل مسئله می‌شود. روش‌هایی از جمله واس[۲۶] یا دودویی [۷] وEIIP[۲۷]برای این نمادگذاری مناسب است. یکی دیگر از روش‌های عددی برای بیان دنباله موردنظر، روش خوانش جفت‌انتها است. به کمک شکل (۱) می-توان ماتریسی $\mathbb{R}^{m\times n}$ به نام ماتریس خوانش ایجاد کرد که n نشان دهنده طول هاپلوتایپ و m مربوط به تعداد قطعات تقسیم شده برای خوانش است.ردیفiام **R**، r_i اساسا اطلاعات مربوط به هاپلوتایپ ارائه شده توسط iامین خوانش را جمع آوری می‌کند. به علت اینکه در انسان‌ها و دیگرترکیبات دوگانه، مکان‌های اسنیپ، دو قطبی هستند، به این معنی که در هر موقعیت هاپلوتایپ تنها دو از چهار نوکلئوتیدA، C، GوT ممکن است رخ دهد. بنابراین می‌توان نوکلئوتیدها را در موقعیت اسنیپ با استفاده از نمادهای باینری نمادگذاری کرد [۴]. این نمادگذاری با توجه به شماره‌ی جایگاه اسنیپ و همچنین فراوانی یا نادر بودن آن نوکلئوتید در اسنیپ تعریف می‌شود. جهت سهولت بررسی، درایه‌هایی از r_i که فاقد اطلاعات اسنیپ هستند، صفر درنظرگرفته می‌شوند. به کمک این نمادگذاری، ماتریس**R** شامل ورودی‌های مبنای سه تایی است. به طور خاص، درایه‌ی (i, j) (jماتریس **R**، اطلاعات مربوط به موقعیتjامین اسنیپ توسطiامین خوانش را ارائه می‌دهد. اگرiامین ورودی موقعیتjامین اسنیپ را پوشش ندهد، درایه‌ی(i, j) از ماتریس **R**، $\mathbf{R}_{ij}=0$ را نتیجه می‌دهد. در (۱) ماتریسی از خوانش قطعه‌ای اسنیپ ارائه شده است [۴].

$$\mathbf{R}=\begin{bmatrix} 0 & 1 & 0 & 0 & -1 & 0 \\ -1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 & -1 \\ 1 & 0 & 0 & -1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & -1 & 0 \end{bmatrix} \qquad (۱)$$

همانطور که بیان شد در ماتریس حاصل از خوانش جفت‌انتها تمام اطلاعات در دسترس نیست. به همین علت به کمک روش‌های بازیابی تکمیل ماتریس رتبه‌پایین سعی در تخمین هاپلوتایپ شده است.

در مسئله ماتریس **R**، ماتریس مشاهدات نویزی و ماتریسی رتبه‌پایین است و **M** ماتریس هدف است که ماتریس بازیابی شده از **R** است. هر سطر از ماتریس **R** مربوط به یک نمونه از هاپلوتایپ است. تجزیه‌ی ماتریسی به صورت $\mathbf{M}=\mathbf{U}\mathbf{V}^T$ صورت می‌گیرد. تخمین هاپلوتایپ از تخمین ماتریس **M**حاصل می‌شود

$$\mathbf{M}=\mathbf{U}\mathbf{V}^T \qquad (۲)$$

V و **U** به ترتیب ماتریس‌های $m\times k, n\times k$ و k نشان‌دهنده‌ی k-پلوئیدی بودن جاندار است. ستونjام از **V**، v_j،دنباله‌ای از jامین هاپلوتایپ است. iامین ردیف از **U**، u_i، نشان می‌دهد که این خوانش مربوط به کدام یک از جفت کروموزوم پدری و مادری است. توجه شود که هر سطر از ماتریس**M**، m_i، یک رشته کامل هاپلوتایپ است. به عنوان مثال، $m_i \in \mathbf{H}$. از آنجایی که $k\prec m, k\prec n$ است طبق رابطه‌ی $rank(\mathbf{M})\le\min\{rank(\mathbf{U}), rank(\mathbf{V})\}$، رتبه‌ی ماتریس **M** حداکثر برابر با **k**خواهد بود. به همین علت این مدل، رتبه‌پایین است [۴]. شایان ذکر است که مدل فاکتورسازی ماتریسی پیش از این در مسائل

مختلفی از جمله تخمین جهت منابع استفاده شده است [۲, ۳, ۷, ۲۰].

ماتریس $\mathbf{M}$ به کمک اپراتور معرفی شده زیر قابل توصیف است .

$$P_{\Omega}(\mathbf{M})=\begin{cases}\mathbf{M}_{ij}\,,if\ (i,j)\in\Omega\\ o\,,otherwise\end{cases} \qquad (۳)$$

در صورتی‌که مسئله دارای خطا نباشد رابطه‌ی $\mathbf{R}=P_{\Omega}(\mathbf{M})$ برقرار است و درایه‌های $\mathbf{R}$ به کمک شرط $(i,j)\in\Omega$ قابل بیان است. مجموعه‌ی Ω خوانش‌های مربوط به کروموزوم‌های مورد بررسی را نشان می‌دهد. در حالت کلی خوانش رشته DNA دارای خطاست در نتیجه $P_{\Omega}(R)\neq P_{\Omega}(M)$. به عنوان مثال ورودی $(i,j)\in\Omega$ در $\mathbf{R}$، $\mathbf{R}_{ij}$، به صورت زیر تعیین می‌شود:

$$\mathbf{R}=\begin{cases}\mathbf{M}_{ij} & With\ \text{Pr}obability:\ 1-p\\ -\mathbf{M}_{ij} & With\ \text{Pr}obability\quad p\end{cases} \qquad (۴)$$

در رابطه (۴)، p میزان خطا دنباله است. وجود خطا درایه‌ها را منفی می‌کند. زیرا اعمال نویز با کاهش یا افزایش مقدار در درایه‌ها بی معنی است.

هر درایه از سطر $\mathbf{R}$ نمونه‌ای از هاپلوتایپ است. به عنوان مثال، $\mathbf{R}_{ij}=-1$ به این معنی است که iامین ورودی موقعیت jامین اسنیپ را پوشش می‌دهد و اطلاعاتی را که توسط 1- کدگذاری شده است را فراهم می‌کند. با این حال، اینکه کدام هاپلوتایپ kتوسط iامین ورودی نمونه‌برداری شده‌است، مجهول است. به طور کلی، ماتریس $\mathbf{R}$ را می‌توان به صورت نمونه‌گیری از ماتریس $\mathbf{M}$ با مقداری خطا استخراج کرد [۲۱].

۳- روش پیشنهادی

هدف در این مقاله بهبود خطای تخمین هاپلوتایپ به کمک ویژگی‌های موجود در ساختار مسئله است. در ابتدا ویژگی‌های ماتریس‌های ژنتیکی مطرح می‌شود، سپس تابع هزینه جدید ارائه شده‌است.

ماتریس بیان شده $\mathbf{U}$ در بخش قبل دارای ابعاد $m\times k$ است که با توجه به انتخاب K دارای ویژگی‌های منحصر به فردی خواهد بود. در صورتیکه k=1 انتخاب شود، ماتریس $\mathbf{U}$ برداری با درایه‌های ±1 خواهد بود. درایه‌ی ۱+، نشان‌دهنده خوانش مربوط به دنباله‌ی هاپلوتایپ $\mathbf{h}_1$ است. همچنین درایه‌ی ۱- به معنی خوانش از هاپلوتایپ $\mathbf{h}_2$ است. در صورتیکه k=2 انتخاب شود، ماتریس $\mathbf{U}$ دارای درایه‌های $\{0,1\}$ است که با در نظر گرفتن قواعد خوانش ژنتیکی، در هر سطر یک درایه با مقدار ۱ و دیگر درایه‌های آن صفر است. ماتریس $\mathbf{V}$ دارای ابعاد $n\times k$ است و مقادیر درایه‌های آن ±1 است.

۳-۱- مسئله بهینه‌سازی پیشنهادی

به کمک اطلاعاتی که از درایه‌های ماتریس‌های $\mathbf{U}$ و $\mathbf{V}$ عنوان شد، در صورتیکه $k=1$ انتخاب شود، ماتریس‌های $\mathbf{U}\in\mathbb{R}^{m\times k}$ و $\mathbf{V}\in\mathbb{R}^{n\times k}$ دارای درایه‌های ±1 است که می‌توان بدون یافتن اطلاعات دقیق از هر درایه‌ی ماتریس، مقدار $\|\mathbf{U}\|_F^2$ و $\|\mathbf{V}\|_F^2$ را به کمک نرم فروبنیوس به صورت رابطه (۵) و (۶) به دست آورد.

$$\|\mathbf{U}\|_F^2=\sum_{i,j}\mathbf{U}_{ij}^2=m \qquad (۵)$$

$$\|\mathbf{V}\|_F^2=\sum_{i,j}\mathbf{V}_{ij}^2=n \qquad (۶)$$

همچنین در صورت انتخاب $k=2$، ماتریس $\mathbf{U}$ که دارای یک درایه‌ی یک در هر سطر است و دیگر درایه‌های آن صفر است رابطه $\|\mathbf{U}\|_F^2=m$ برقرار است. اما ماتریس $\mathbf{V}$ که دارای درایه‌های ±1 است، رابطه (۷) جایگزین رابطه (۶) می‌شود.

$$\|\mathbf{V}\|_F^2=\sum_{i,j}\mathbf{V}_{ij}^2=2n \qquad (۷)$$

در ادامه از روابط (۵) و (۶) استفاده شده است. از طرفی به کمک فاکتورسازی ماتریسی در مسئله، یکی از شروط مسئله برای بازیابی ماتریس‌ها $(\mathbf{M})_{ij}=\left(\mathbf{U}\mathbf{V}^{\mathrm{T}}\right)_{ij}=\pm1$ است. هدف مسئله تخمین $\mathbf{M}$ است. برای محاسبه‌ی آن با تجزیه‌ی ماتریسی به دنبال $\mathbf{U}$و $\mathbf{V}$هستیم که شرط $\left(\mathbf{U}\mathbf{V}^{\mathrm{T}}\right)_{ij}=\pm1$ را اقناع کند. البته نه هر نوع ماتریسی که دراین رابطه صدق کند. به طور مثال، اگر درایه‌های ماتریس $\mathbf{U}_{ij}=1/1000$ و $\mathbf{V}_{ij}=1000$ باشد، به کمک ضرب ماتریسی، $(\mathbf{M})_{ij}=\pm1$ حاصل می‌شود. اما با وجود اینکه جواب حاصل دارای شرط لازم برای درایه‌های ماتریس $\mathbf{M}$ است، مناسب داده‌های ژنتیکی مسئله نیست. زیرا درایه‌های ماتریس $\mathbf{U}$و $\mathbf{V}$ نباید مقادیر بزرگی داشته باشند و این مقادیر با اطلاعاتی که در فرض مسئله وجود دارد هم‌خوانی ندارد. می‌توان نتیجه گرفت که مسئله مورد نظر علاوه بر شرط لازم، مناسب بودن میزان بزرگی مقادیر درایه‌های ماتریس $\mathbf{U}$ و $\mathbf{V}$به عنوان شرط کافی، نیز نیاز دارد.

در نتیجه بر اساس رابطه‌های (۵) و (۶)، مسئله‌ی بهینه‌سازی جدیدی را پیشنهاد می‌کنیم.

$$f(\mathbf{U},\mathbf{V})=\left\|P_{\Omega}\left(\mathbf{R}-\mathbf{U}\mathbf{V}^{T}\right)\right\|_F^2+\lambda_1\left(\|\mathbf{U}\|_F^2-m\right)+\lambda_2\left(\|\mathbf{V}\|_F^2-n\right) \qquad (۸)$$

که $\mathbf{U}\in\mathbb{R}^{m\times k}$ و $\mathbf{V}\in\mathbb{R}^{n\times k}$ و ماتریس رتبه پایین مشاهده شده $\mathbf{R}\in\mathbb{R}^{m\times n}$، n طول هاپلوتایپ است. در مسئله پیشنهادی پارامترهای اضافه شده، باعث لحاظ شدن شرط کافی برای مقادیر ماتریس‌های $\mathbf{U}$و $\mathbf{V}$ هنگام کمینه‌سازی تابع هزینه، شده است.

الگوریتم گرادیان کاهشی به کمک تابع پیشنهادی بیان می‌شود.

در رابطه‌ی (۸) هدف یافتن $\mathbf{U}$ و $\mathbf{V}$ است که تابع $f(\mathbf{U},\mathbf{V})$ را کمینه کند. محاسبه $\mathbf{U}$و $\mathbf{V}$به ازای مقادیر $t=0,1,2,...$ به صورت رابطه زیر بیان می‌شود:

$$\mathbf{V}_{t+1}=\mathbf{V}_t-\alpha\nabla f(\mathbf{V}_t) \qquad (۹)$$

$$\mathbf{U}_{t+1}=\mathbf{U}_t-\beta\nabla f\left(\mathbf{U}_t\right) \quad (۱۰)$$

به کمک تابع جدید بیان شده، مشتقات جزئی تابع هزینه $f\left(\mathbf{U},\mathbf{V}\right)$، $\nabla f\left(\mathbf{U}_t,\mathbf{V}_t\right)$ و $\nabla f\left(\mathbf{U}_t,\mathbf{V}_{t+1}\right)$ به صورت زیر تعریف می‌شود:

$$\nabla f\left(\mathbf{V_t}\right)=-2\left(P_\Omega\left(\mathbf{R}-\mathbf{U}_t\mathbf{V}_t^T\right)\right)^T\mathbf{U}_t+\lambda_2\mathbf{V}_t \quad (۱۱)$$

$$\nabla f\left(\mathbf{U}_t\right)=-2P_\Omega\left(\mathbf{R}-\mathbf{U}_t\mathbf{V}_{t+1}^T\right)\mathbf{V}_{t+1}+\lambda_1\mathbf{U}_t \quad (۱۲)$$

حال به کمک رابطه‌های بیان شده (۱۱) و (۱۲) و تابع جدید مطرح شده (۷)، الگوریتم گرادیان کاهشی [۴] به کمک شرایط موجود در مسئله پیشنهادی، جدول (۱) بازنویسی می‌شود.

جدول (۱): الگوریتم گرادیان کاهشی با تابع پیشنهادی جدید

گام اول	مقدار دهی اولیه $\mathbf{U}_0$ و $\mathbf{V}_0$ به کمک روش تجزیه مقدار تکین
گام دوم	$\nabla f\left(\mathbf{V_t}\right)=-2\left(P_\Omega\left(\mathbf{R}-\mathbf{U}_t\mathbf{V}_t^T\right)\right)^T\mathbf{U}_t+\lambda_2\mathbf{V}_t$ $\mathbf{V}_{t+1}=\mathbf{V}_t-\alpha\nabla f\left(\mathbf{V_t}\right)$ $\nabla f\left(\mathbf{U}_t\right)=-2P_\Omega\left(\mathbf{R}-\mathbf{U}_t\mathbf{V}_{t+1}^T\right)\mathbf{V}_{t+1}+\lambda_1\mathbf{U}_t$ $\mathbf{U}_{t+1}=\mathbf{U}_t-\beta\nabla f\left(\mathbf{U}_t\right)$
گام سوم	گرد کردن $\mathbf{V}_{tLast}$ به ±1

۳-۲- همگرایی مسئله

هدف ما در این بخش محاسبه‌ی ضریب آلفا موجود در رابطه (۹) است تا به کمک آن‌ها الگوریتم گرادیان کاهشی جدول (۱) حل شود.

تابع هزینه مسئله و گرادیان ماتریس $\mathbf{V}$ که قبلا در رابطه‌ی (۸) و (۱۱) بیان شد

در رابطه‌ی (۱۱) ماتریس $\mathbf{U}$ ثابت فرض شده و از تابع هدف نسبت به ماتریس$\mathbf{V}$ مشتق گرفته می‌شود.

رابطه (۸) برای ماتریس‌های $\left(\mathbf{U}_t,\mathbf{V}_{t+1}\right)$ و $\left(\mathbf{U}_t,\mathbf{V}_t\right)$ به صورت زیر بازنویسی می‌شود:

$$f\left(\mathbf{U}_t,\mathbf{V}_{t+1}\right)=\left\|P_\Omega\left(\mathbf{R}-\mathbf{U}_t\mathbf{V}_{t+1}^T\right)\right\|_F^2+\lambda_2\left\|\mathbf{V}_{t+1}\right\|_F^2+\lambda_1\left\|\mathbf{U}_t\right\|_F^2 \quad (۱۳)$$

$$f\left(\mathbf{U}_t,\mathbf{V}_t\right)=\left\|P_\Omega\left(\mathbf{R}-\mathbf{U}_t\mathbf{V}_t^T\right)\right\|_F^2+\lambda_2\left\|\mathbf{V}_t\right\|_F^2+\lambda_1\left\|\mathbf{U}_t\right\|_F^2 \quad (۱۴)$$

به کمک رابطه (۹) برای ماتریس تکرار شونده $\mathbf{V}$به ساده‌سازی رابطه‌های (۱۳) و (۱۴) می‌پردازیم.

هدف محاسبه $f\left(\mathbf{U}_t,\mathbf{V}_{t+1}\right)-f\left(\mathbf{U}_t,\mathbf{V}_t\right)$ است که به کمک فرض همگرایی موجود برای مسئله‌ی بهینه‌سازی با تابع هزینه (۸) بتوان مفدار آلفا را به دست آورد. در نتیجه ابتدارابطه $\left\|P_\Omega\left(\mathbf{R}-\mathbf{U}_t\mathbf{V}_{t+1}^T\right)\right\|_F^2-\left\|P_\Omega\left(\mathbf{R}-\mathbf{U}_t\mathbf{V}_t^T\right)\right\|_F^2$ را بسط می‌دهیم. برای جلوگیری از پیچیدگی محاسبات، تفاضل $\left\|P_\Omega\left(\mathbf{R}-\mathbf{U}_t\mathbf{V}_{t+1}^T\right)\right\|_F^2-\left\|P_\Omega\left(\mathbf{R}-\mathbf{U}_t\mathbf{V}_t^T\right)\right\|_F^2$ را$\mathbf{Q}$نام‌گذاری می‌کنیم.

$$Q=\left\|P_\Omega\left(\mathbf{R}-\mathbf{U}_t\left(\mathbf{V}_t-\alpha\nabla f\left(\mathbf{V}_t\right)\right)^T\right)\right\|_F^2-\left\|P_\Omega\left(\mathbf{R}-\mathbf{U}_t\mathbf{V}_t^T\right)\right\|_F^2 \quad (۱۵)$$

رابطه (۱۵) به کمک رابطه (۹) حاصل شده که به صورت ساده‌تر نیز قابل بیان است:

$$Q=\left\|P_\Omega\left(\mathbf{R}-\mathbf{U}_t\mathbf{V}_t^T+\alpha\mathbf{U}_t\nabla f\left(\mathbf{V}_t\right)^T\right)\right\|_F^2-\left\|P_\Omega\left(\mathbf{R}-\mathbf{U}_t\mathbf{V}_t^T\right)\right\|_F^2 \quad (۱۶)$$

به کمک تعریف نرم فروبنیوس، نرم به صورت سیگما بیان می‌شود و محدوده‌ی سیگما به کمک تعریف اپراتور $P_\Omega(.)$ که در رابطه (۳) تعریف شد، به دست می‌آید. در نتیجه رابطه (۱۶) به صورت زیر حاصل می‌شود:

$$\begin{aligned}Q&=2\alpha\sum_{(i,j)\epsilon\Omega}\left(\mathbf{R}-\mathbf{U}_t\mathbf{V}_t^T\right)_{ij}\left(\mathbf{U}_t\nabla f\left(\mathbf{V}_t\right)^T\right)_{ij}\\&+\alpha^2\sum_{(i,j)\epsilon\Omega}\left(\left(\mathbf{U}_t\nabla f\left(\mathbf{V_t}\right)^T\right)_{ij}\right)^2\end{aligned} \quad (۱۷)$$

بخش دوم تفاضل (۱۳) و (۱۴) که شامل جملات با ضریب λ است، در زیر بیان شده است:

$$\lambda_2\left\|\mathbf{V}_{t+1}\right\|_F^2-\lambda_2\left\|\mathbf{V}_t\right\|_F^2=\lambda_2\left\|\mathbf{V}_t-\alpha\nabla f\left(\mathbf{V}_t\right)\right\|_F^2-\lambda_2\left\|\mathbf{V}_t\right\|_F^2 \quad (۱۸)$$

رابطه (۱۸) به کمک تعریف نرم فروبنیوس به صورت زیر حاصل می‌شود:

$$\lambda_2\left\|\mathbf{V}_{t+1}\right\|_F^2-\lambda_2\left\|\mathbf{V}_t\right\|_F^2=\lambda_2\sum_{i,j}\left[\left(\mathbf{V}_t-\alpha\nabla f\left(\mathbf{V}_t\right)\right)_{ij}\right]^2-\lambda_2\left\|\mathbf{V}_t\right\|_F^2 \quad (۱۹)$$

با طرف راست رابطه (۱۹) می‌توان نوشت:

$$\lambda_2\left\|\mathbf{V}_{t+1}\right\|_F^2-\lambda_2\left\|\mathbf{V}_t\right\|_F^2=\lambda_2\alpha^2\left\|\nabla f\left(\mathbf{V_t}\right)\right\|_F^2-2\alpha\lambda_2\sum_{i,j}\left(\mathbf{V}_t\right)_{ij}\left(\nabla f\left(\mathbf{V}_t\right)\right)_{ij} \quad (۲۰)$$

به کمک روابط ساده شده‌ی (۱۷) و (۲۰)، رابطه $f\left(\mathbf{U}_t,\mathbf{V}_{t+1}\right)-f\left(\mathbf{U}_t,\mathbf{V}_t\right)$ به صورت زیر بیان می‌شود:

$$\begin{aligned}f\left(\mathbf{U}_t,\mathbf{V}_{t+1}\right)-f\left(\mathbf{U}_t,\mathbf{V}_t\right)&=\alpha^2\left\|P_\Omega\left(\mathbf{U}_t\nabla f\left(\mathbf{V}_t\right)^T\right)\right\|_F^2+\lambda_2\alpha^2\left\|\nabla f\left(\mathbf{V}_t\right)\right\|_F^2+\\&2\alpha\sum_{(i,l)\in\Omega}\left(\mathbf{R}-\mathbf{U}_t\mathbf{V}_t^T\right)_{ij}\left(\mathbf{U}_t\nabla f\left(\mathbf{V}_t\right)^T\right)_{ij}-\lambda_2\alpha^2\left\|\nabla f\left(\mathbf{V}_t\right)\right\|_F^2\end{aligned} \quad (۲۱)$$

با توجه به اینکه همگرایی مسئله مورد نظر است از نامساوی زیر کمک گرفته می‌شود:

$$f\left(\mathbf{U}_t,\mathbf{V}_{t+1}\right)-f\left(\mathbf{U}_t,\mathbf{V}_t\right)\le0 \quad (۲۲)$$

در صورتی‌که رابطه (۲۱) برابر با صفر قرار گیرد، ریشه‌های معادله به صورت $\left\{0,\frac{\alpha_1}{\alpha_2}\right\}$ حاصل می‌شود که $\frac{\alpha_1}{\alpha_2}$ به صورت زیر تعریف می‌شود:

$$\begin{aligned}\alpha_1&=2[\lambda_2\Sigma\left(\mathbf{V}_t\right)_{ij}\left(\nabla f\left(\mathbf{V}_t\right)\right)_{ij}\\&-\sum_{(i,j)\epsilon\Omega}\left(\mathbf{R}-\mathbf{U}_t\mathbf{V}_t^T\right)_{ij}\left(\mathbf{U}_t\nabla f\left(\mathbf{V}_t\right)\right)_{ij}\end{aligned} \quad (۲۳)$$

$$\alpha_2 = \lambda_2 \left\| \nabla f\left(\mathbf{V}_t\right) \right\|_F^2 + \left\| P_\Omega\left(\mathbf{U}_t \nabla f\left(\mathbf{V}_t\right)^T\right) \right\|_F^2 \qquad (۲۴)$$

به کمک ریشه‌های بـه دسـت آمـده و نامسـاوی (۲۲) کـه مربـوط بـه همگرایی مسئله است، می‌توان نتیجه گرفت که اگر آلفا بـین دو ریشـه انتخاب شود، در نامساوی (۲۲) صدق می‌کند.

در نتیجه با توجه به بازه‌هـای انتخـابی بـرای مقـدار ثابـت $C \in (0,1)$، ضریب آلفا به صورت زیر به دست می‌آید:

$$\alpha = C \frac{\alpha_{numerator}}{\alpha_{Deno\,minator}} \qquad (۲۵)$$

$$\alpha_{numerator} = 2\left[\lambda_2 \Sigma(\mathbf{V}_t)_{ij} \left(\nabla f\left(\mathbf{V_t}\right)\right)_{ij} - \sum_{(i,j)\in\Omega} \left(\mathbf{R} - \mathbf{U}_t \mathbf{V}_t^T\right)_{ij} \left(\mathbf{U}_t \nabla f(\mathbf{V_t})\right)_{ij}\right] \qquad (۲۶)$$

$$\alpha_{Deno\,minator} = \lambda_2 \left\| \nabla f\left(\mathbf{V}_t\right) \right\|_F^2 + \left\| P_\Omega\left(\mathbf{U}_t \nabla f\left(\mathbf{V}_t\right)^T\right) \right\|_F^2 \qquad (۲۷)$$

آلفای به دسـت آمـده در رابطـه (۲۵) بـرای جایگـذاری در رابطـه (۹) استفاده می‌شود تا به کمک الگوریتم موجود در جدول (۱) بـه تخمـین هاپلوتایپ کمک کند. محاسبه ضریب بتا هم مشابه ضریب آلفـا انجـام خواهد گرفت و رابطه نهایی به صورت زیر بیان می‌شود :

$$\beta = C \frac{2\beta_{numerator}}{\beta_{Deno\,minator}} \qquad (۲۸)$$

$$\alpha_{numerator} = 2\Big[\lambda_1 \Sigma(\mathbf{U}_t)_{ij} \left(\nabla f\left(\mathbf{U}_t, \mathbf{V_{t+1}}\right)\right)_{ij} - \sum_{(i,j)\in\Omega} \left(\mathbf{R} - \mathbf{U}_t \mathbf{V}_{t+1}^T\right)_{ij} \left(\nabla f\left(\mathbf{U}_t, \mathbf{V_{t+1}}\right)\mathbf{V}_{t+1}^T\right)_{ij}\Big] \qquad (۲۹)$$

$$\beta_{Deno\,minator} = \lambda_1 \sum \left(\nabla f\left(\mathbf{U}_t, \mathbf{V}_{t+1}\right)_{ij}\right)^2 + \sum_{(i,j)\in\Omega} \left(\nabla f\left(\mathbf{U}_t, \mathbf{V_{t+1}}\right)\mathbf{V}_{t+1}^T\right)_{ij} \qquad (۳۰)$$

مقدار ثابت $C \in (0,1)$ برای شبیه سازی‌ها اعمال می‌شود.

۴- نتایج شبیه‌سازی

در این بخش در ابتدا معیـار ارزیـابی نتـایج ارائـه شـده اسـت. روشـی عملکرد بهتری دارد که معیارها خطـای کمتـری را در بازیـابی داده‌هـا نتیجه دهد. در این مقالـه از معیـار خطـای تخمـین و کمینـه خطـای تصحیح استفاده می‌شود.

در ابتدا فاصله همینگ[۲۸] به صورت زیر تعریف می‌گردد:

$$\mathrm{d}(a,b) = \begin{cases} 1 & a \neq b \\ 0 & o.w \end{cases} \qquad (۳۱)$$

کمینه خطای تصحیح[۲۹]که معیـار عملکـرد روش پیشـنهادی اسـت بـه صورت زیر تعریف می‌شود:

$$MEC = min\left(D\left(\mathbf{h},\hat{\mathbf{h}}\right), D\left(\mathbf{h},-\hat{\mathbf{h}}\right)\right) \qquad (۳۲)$$

در رابطه (۳۲)، $\mathbf{h}$ هاپلوتایپ واقعی و $\hat{\mathbf{h}}$ نسخه‌ی بازیابی شده هاپلوتایپ است. تساوی $D\left(\mathbf{h},\hat{\mathbf{h}}\right) = \sum_{l=1}^{n} d\left(\mathbf{h_l},\hat{\mathbf{h}}_l\right)$ در رابطه (۳۲) برقرار است که n طـول هاپلوتایپ است.

خطای تخمین به صورت زیر بیان می‌شود:

$$MEC_r = \frac{MEC}{n} \qquad (۳۳)$$

بـرای ارائـه نتـایج شـبیه سـازی تـابع هزینـه پیشـنهادی در تخمـین هاپلوتایپ بررسی و نتـایج آن بـا روش مرجـع [۴] مقایسـه مـی‌شـود. روش‌هـای گرادیـان کاهشـی بـین روش‌هـای موجـود بـرای تخمـین هاپلوتایـــپ از جملـــه HapCompass، SPH، FAST و Belief Propagation بهترین جواب را نتیجه می‌دهد. بـه همـین علـت روش گرادیان کاهشی به عنوان روش مرجـع انتخـاب شـده اسـت [۱۹, ۲۲, ۲۳].

نتایج به دست آمده کاهش خطای تخمین هاپلوتایپ را نشان می‌دهـد. در هر مرحله اجرا، داده‌هایی شبیه سازی‌ها برای داده‌های با سـه نـرخ نویز $\{0.1,0.2,0.3\}$ و سـه نـرخ ورودی $\{0.1,0.2,0.3\}$ و طـول هاپلوتایـپ $\{50,100\}$ هاپلوتایپ اجرا شد که در تمام موارد تابع پیشنهادی نسـبت به مقاله مرجع کاهش معیار MEC_r به دست آمد. در ادامه بخشـی از این نتایج بیان می‌شود:

جدول (۲): مقایسه عملکرد تابع پیشنهادی با مرجع در نرخ ورودی ۰/۲

پیشنهادی $\lambda = 10$		پیشنهادی $\lambda = 2$		مرجع[۴]	
درصد بهبود نسبت به مرجع	MEC_r	درصد بهبود نسبت به مرجع	MEC_r	MEC_r	نرخ نویز
۸	۰,۰۰۲۱	۲۰	۰,۰۰۱۹	۰,۰۰۲۴	۰,۱
۰,۹۹	۰,۰۳۰	۲,۱۳	۰,۰۲۹۶	۰,۰۳۰۳	۰,۲
۴,۸۷	۰,۲۰۹۰	۴,۳۶	۰,۲۱۰	۰,۲۱۹۷	۰,۳

جدول (۳): مقایسه عملکرد تابع پیشنهادی با مرجع در نرخ نویز ۰/۲

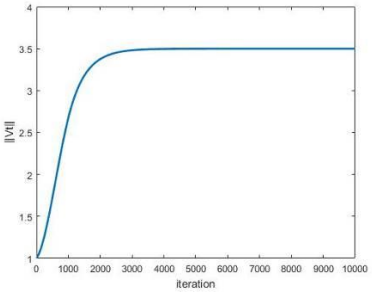
پیشنهادی $\lambda = 10$		پیشنهادی $\lambda = 2$		مرجع[۴]	
درصد بهبود نسبت به مرجع	MEC_r	درصد بهبود نسبت به مرجع	MEC_r	MEC_r	نرخ ورودی
۵,۴۳	۰,۱۶۰	۱۱,۹۳	۰,۱۴۹	۰,۱۶۹۲	۰,۱
۲,۸۱	۰,۰۲۰۷	۷	۰,۰۱۹۸	۰,۰۲۱۳	۰,۲
۲,۱۰	۰,۰۰۹۳	۹,۴۷	۰,۰۰۸۶	۰,۰۰۹۵	۰,۳

در مراجع روش‌های مختلفی برای انتخاب ضریب λ بیان می‌شود که در آن‌ها دلیل و اثبات ریاضی برای انتخاب این مقدار وجود ندارد. انتخاب λ می‌توان به کمک نتایج حاصل از شبیه‌سازی انجام گیرد و یا به کمک اثبات‌های دیگر موجود در مسئله، شرایطی برای مقدار دهی در نظر گرفت. ما نیز در این مقاله به کمک نتایج حاصل از شبیه سازی مقدارهای ۲و ۱۰ برای ضریب λ انتخاب کردیم [۲۰, ۲۴-۲۷].

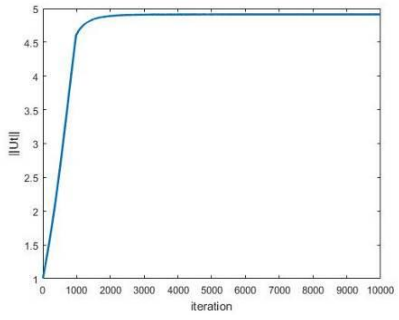
در این شبیه‌سازی m=n=50 انتخاب شده است و تابع برای دو مقدار λ پیشنهادی اجرا و با مقاله مرجع مقایسه می‌شود. نویز اعمال شده بر اساس تعریف نویز در مسئله‌ی هاپلوتایپ به صورت رابطه (۳) می‌باشد و نرخ نویز نسبت درایه‌های نویزی در **R** به کل درایه‌های دارای مقدار در **R** است. نرخ ورودی از نسبت تعداد درایه‌های دارای مقدار در **R** نسبت به کل درایه‌های آن تعریف می‌شود. همانطور که در جدول‌های (۲) و (۳) مشخص است به دلیل وجود ترم‌های مرتبط با ضرایب λ_1 و λ_1 کاهش خطای تخمین هاپلوتایپ و در نتیجه تخمین هاپلوتایپ با دقت بیشتری انتظار می‌رفت. بنابراین این نتایج صحت تابع پیشنهادی (۸) را تایید می‌کند. زیرا معیار عملکرد رابطه (۳۳) در جدول‌های (۲) و (۳) کاهش میزان MEC_r را نشان می‌دهد. یعنی وجود ضرایب λ_1 و $\lambda 2$ باعث شده است که برای تخمین هاپلوتایپ از انتخاب داده‌های پرت و نامناسب برای ماتریس‌های **U**و **V**جلوگیری شود.

با توجه به برتری روش بیان شده نسبت به گرادیان کاهشی و یکسان بودن نوع داده‌ها در شبیه سازی‌ها میتوان نتیجه گرفت که روش پیشنهادی از دیگر الگوریتم‌های مورد استفاده برای تخمین هاپلوتایپ عنوان شده در ابتدا این بخش هم برتری دارد.

در روش پیشنهادی همگرایی مسئله هم مورد ارزیابی قرار گرفته است که نتایج حاصل در شکل‌های (۱) و (۲) قابل مشاهده است. همگرایی مسئله برای طول هاپلوتایپ ۵۰، نرخ ورودی ۰/۳، نرخ نویز ۰/۱ و ضریب λ ۱۰ اجرا شده است. محور افقی تعداد تکرارها و محور عمودی مربوط به نرم **V** و **U** است. همگرایی مسئله در مثال بیان شده قابل مشاهده است.



شکل (۱): همگرایی نرم V برای نرخ نویز ۰/۱ و نرخ ورودی ۰/۳



شکل (۲): همگرایی نرم U برای نرخ نویز ۰/۱ و نرخ ورودی ۰/۳

۵- نتیجه‌گیری

الگوریتم گرادیان کاهشی یکی از روش‌های تخمین هاپلوتایپ به کمک تجزیه‌ی ماتریس است. در روش‌های تخمین گزارش شده، از اطلاعات درایه‌های ماتریس استفاده نشده‌است. به همین علت در این مقاله با گذاشتن شرط بر روی میزان بزرگی مقادیر درایه‌های ماتریس **U**و **V**، تابع هزینه جدید پیشنهاد شد. تابع پیشنهادی مانع از تخمین ماتریس‌های پرت و نامناسب **U**و **V** شد که باعث تخمین دقیق‌تر هاپلوتایپ شد. نتایج حاصل از شبیه‌سازی، نشان‌دهنده‌ی کاهش خطای تخمین در تابع هزینه پیشنهادی نسبت به روش‌های پیشین است. این نتایج در داده‌های متفاوت ماتریس خوانش برقرار است که برتری تابع پیشنهادی را نشان می‌دهد.

مراجع

[۱] فرازکیش راضیه. مدل‌سازی قابلیت اطمینان در نانوربات‌های زیستی. مجله مهندسی برق و الکترونیک ایران. ۱۳۹۹; ۱۷ (۳) :۱۱-۱۶.

[۲] آتشبار محمود، کهائی محمدحسین. جهت یابی چند گوینده با استفاده از روش .WCSSDOA مجله مهندسی برق و الکترونیک ایران. ۱۳۹۵; ۱۳ (۲) :۶۱-۷۴

[۳] مجیدیان سینا، حدادی فرزان. تخمین جهت منابع با استفاده از زیرفضای خنری-رائو. مجله مهندسی برق و الکترونیک ایران. ۱۳۹۶; ۱۴ (۲) :۳۷-۴۷

[4] C. Cai, S. Sanghavi, and H. Vikalo, "Structured low-rank matrix factorization for haplotype assembly," IEEE Journal of Selected Topics in Signal Processing, vol. 10, pp. 647-657, 2016.

[5] Z. Puljiz and H. Vikalo, "A message passing algorithm for haplotype assembly," in 2013 Asilomar Conference on Signals, Systems and Computers, 2013, pp. 1726-1729.

[6] S. Das and H. Vikalo, "SDhaP: haplotype assembly for diploids and polyploids via semi-definite programming," BMC genomics, vol. 16, p. 260, 2015.

[7] T. Inbamalar and R. Sivakumar, "Study of DNA sequence analysis using DSP techniques," J. Autom. Control Eng., vol. 1, 2013.

[8] X. Jiang, Z. Zhong, X. Liu, and H. C. So, "Robust matrix completion via alternating projection," IEEE Signal Processing Letters, vol. 24, pp. 579-583, 2017.
[9] H. C. So and W.-J. Zeng, "Outlier-Robust Matrix Completion via lp-Minimization," 2018.
[10] X. P. Li, L. Huang, H. C. So, and B. Zhao, "A survey on matrix completion: perspective of signal processing," arXiv preprint arXiv:1901.10885, 2019.
[11] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," in Proceedings of the forty-fifth annual ACM symposium on Theory of computing, 2013, pp. 665-674.
[12] J. Yu, G. Zhou, C. Li, Q. Zhao, and S. Xie, "Low Tensor-Ring Rank Completion by Parallel Matrix Factorization," IEEE transactions on neural networks and learning systems, 2020.
[13] Q. Wang, X. He, X. Jiang, and X. Li, "Robust Bi-stochastic Graph Regularized Matrix Factorization for Data Clustering," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.
[14] S. Barik and H. Vikalo, "Matrix Completion and Performance Guarantees for Single Individual Haplotyping," IEEE Transactions on Signal Processing, vol. 67, pp. 4782-4794, 2019.
[15] R. Rizzi, V. Bafna, S. Istrail, and G. Lancia, "Practical algorithms and fixed-parameter tractability for the single individual SNP haplotyping problem," in International Workshop on Algorithms in Bioinformatics, 2002, pp. 29-43.
[16] A. M. Maxam and W. Gilbert, "A new method for sequencing DNA," Proceedings of the National Academy of Sciences, vol. 74, pp. 560-564, 1977.
[17] N. M. Haan and S. J. Godsill, "Bayesian models for DNA sequencing," in 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002, pp. IV-4020-IV-4023.
[18] A. S. Motahari, G. Bresler ,and N. David, "Information theory of DNA shotgun sequencing," IEEE Transactions on Information Theory, vol. 59, pp. 6273-6289, 2013.
[19] D. Aguiar and S. Istrail, "HapCompass: a fast cycle basis algorithm for accurate haplotype assembly of sequence data," Journal of Computational Biology, vol. 19, pp. 577-590, 2012.
[20] Z. Zhou, X. Li, J. Wright, E. Candes, and Y. Ma, "Stable principal component pursuit," in 2010 IEEE international symposium on information theory, 2010, pp. 1518-1522.
[21] H. Si, H. Vikalo, and S. Vishwanath, "Information-theoretic analysis of haplotype assembly," IEEE Transactions on Information Theory, vol. 63, pp. 3468-3479, 2017.
[22] F. Geraci, "A comparison of several algorithms for the single individual SNP haplotyping reconstruction problem," Bioinformatics, vol. 26, pp. 2217-2225, 2010.
[23] Z. Puljiz and H. Vikalo, "Decoding genetic variations: Communications-inspired haplotype assembly," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 13, pp. 518-530.۲۰۱۵ ,
[24] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," Journal of the ACM (JACM), vol. 58, pp. 1-37, 2011.
[25] M. Mardani, G. Mateos, and G. B. Giannakis, "Decentralized sparsity-regularized rank minimization: Algorithms and applications," IEEE Transactions on Signal Processing, vol. 61, pp. 5374-5388, 2013.
[26] F. Shang, J. Cheng, Y. Liu, Z.-Q. Luo, and Z. Lin, "Bilinear factor matrix norm minimization for robust PCA: Algorithms and applications," IEEE transactions on pattern analysis and machine intelligence, vol. 40, pp. 2066-2080, 2017.
[27] R. Cabral, F. De la Torre, J. P. Costeira, and A. Bernardino, "Unifying nuclear norm and bilinear factorization approaches for low-rank matrix decomposition," in Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 2488-2495.

زیرنویس‌ها

[1]Deoxyribonucleic Acid
[2]Semi-Definite Programming
[3]Digital Signal Processing (DSP)
[4]Discrete Fourier Transform (DFT)
[5]Discrete Wavelet Transform (DWT)
[6]Atomic Decomposition for Minimum Rank Approximation
[7]Subspace Evolution and Transfer
[8]Alternating Projection
[9]Ribonucleic Acid (RNA)
[10]Deoxyribonucleic Acid(DNA)
[11]Adenin(A)
[12]Thyamine(T)
[13]Cytosine(C)
[14]Guanine(G)
[15]Chromosomes
[16]Homologous
[17]Diploid
[18]Single Nucleotide Polymorphisms (SNP)
[19] AllanMaxam
[20] WalterGilbert
[21] Frederick Sanger
[22] Shotgun Sequencing
[23]Read
[24]Overlap
[25]Paired-end
[26]Voss
[27]Electron–Ion Interaction Potential
[28]Hammming distance
[29]Minimum Error Correction (MEC)