کشف و دسته‌بندی ژن‌های موثر براساس روش تصویر تصادفی

صدیقه نورانی پیله‌رود*، موسی گلعلی‌زاده

علوم ریاضی، تربیت مدرس، تهران، ایران
sedigheh.noorani@modares.ac.ir

چکیده:

تحقیق درباره خوشه‌بندی داده‌های بیان ژنی از دیرباز مورد توجه بسیاری از محققین بوده است. داده‌های مرتبط با ژن‌ها معمولا از نوع بُعد بالا هستند به این معنی که تعداد متغیرها از تعداد مشاهدات بیشتر است. خوشه‌بندی چنین داده‌هایی به روش‌های مرسوم چالش‌های خاصی را به همراه دارد. برای مقابله با چنین مشکلاتی یک رویکرد جدید که توجه محققین بیشماری را به خود جلب کرده، روش تصویرهای تصادفی است. این رویکرد از یک مدل آمیخته گاوسی بهره می‌برد، ولی برای انتخاب دقیق این تصویرها از الگوریتم خاصی استفاده می‌شود. در مقاله حاضر عملکرد این روش از هر دو منظر نظری و کاربردی مورد بررسی قرار گرفته است و برتری آن بر روی تحلیل داده‌های واقعی مربوط به گروه‌بندی بیماران داده‌های لوسمی ۱۲ نشان داده شده است.

کلید واژه‌ها: داده‌های بُعد بالا، تصویر تصادفی، مدل آمیخته گاوسی، داده‌های بیان ژنی.

۱. مقدمه

با پیشرفت تکنولوژی محاسباتی، در بسیاری از حوزه‌های علمی مشاهده می‌شود که اکثر داده‌های جمع‌آوری شده اخیر از گونه بُعد بالا[1] هستند. اطلاق چنین مفهومی به داده‌ها در واقع به وجود تعداد بیشتر متغیر (p) در مقایسه با تعداد نمونه (n) برمی‌گردد [۳]. خوشه‌بندی چنین داده‌هایی مشکلات تحلیلی و محاسباتی خاصی را به همراه دارد. یکی از مورد توجه‌ترین این مشکلات پدیده مشقت‌بُعدچندی[2] است که توسط بلمن معرفی شده است [۲].

اخیرا، در دنیای تحلیل داده‌های بُعد بالا رویکردی تحت عنوان تصویر تصادفی[3] (RP) پیشنهاد شد که ادعا می‌شود در مقایسه با سایر روش‌های کاهش بُعد اطلاعات کمتری را از دست می‌دهد [۱]. این روش مبتنی بر ایجاد تصویرهای تصادفی از داده‌های اولیه، انتخاب زیرمجموعه بسیار کوچک ولی بهینه از تصویرهای حاصل و سپس انجام خوشه‌بندی بر روی تصویرهای بهینه است. لازم به اشاره است که ایده اصلی این روش قبلا در زمینه‌ی رده‌بندی راهنماییده مطرح و عملکرد موفقیت‌آمیز آن نشان داده شد [۴].

[1] High Dimension
[2] Curse of Dimensionality
[3] Random Projection

از نقطه نظر ریاضی، روش RP از تصویر کردن داده‌های اولیه بُعد بالا به ابعاد پایین‌تر با استفاده از یک ماتریس تصادفی که دارای ستون‌های متعامد[۴] با طول واحد هستند، بهره می‌برد. ایده اصلی روش RP در قالب یک لم آمده است [۸]. بنا به این لم، هر n نقطه p بُعدی می‌تواند به وسیله‌ی یک ماتریس تصادفی $A \in \mathbb{R}^{p\times d}$ با ستون‌های متعامد به صورت خطی روی مختصات $d = O(\frac{\log n}{\varepsilon^۲})$ که $d << p$، تصویر شود. در این حالت، فاصله‌های دو به دو نقاط برابر $۱ \pm \varepsilon$ است. به بیان دقیق‌تر، بعد از تصویر داده‌ها نامساوی زیر بین هر دو نقطه p بُعدی برقرار است:

$$(۱-\varepsilon)\|x_i - x_j\|_۲ \leq \|A^T x_i - A^T x_j\|_۲ \leq (۱+\varepsilon)\|x_i - x_j\|_۲$$

که x_i به ازای $i = ۱, \ldots, n$ برداری p بُعدی و $\|.\|_۲$ نشان‌دهنده نرم $L_۲$ است.

یکی از کارکردهای مستقیم این لم در استفاده از روش RP در چارچوب خوشه‌بندی مدل‌مبنای داده‌های بُعد بالا این است که اگر دو چگالی در فضای بُعد بالا از هم فاصله کافی داشته باشند، انتظار می‌رود در فضای کاهش یافته d بُعدی همان فاصله حفظ شود. چنین نتیجه‌ای به همراه ایده مطرح شده در [۴] برای رده‌بندی راهنماییده، به محقق کمک می‌کند تا مجموعه‌ای از تصویرهای تصادفی مستقل با بُعد کم تولید کرده و برای هر یک از آن‌ها یک مدل آمیخته گاوسی[۵] (GMM) را به کار برد. لازم به اشاره است که برای مقایسه مدل‌ها استفاده از یک معیار آماری ضروری است که در مقاله حاضر از معیار اطلاع بیزی[۶] (BIC) استفاده شده است. روش محاسبه این معیار به شرح زیر است:

$$BIC = -۲\log(L) + k\log(n), \tag{۱}$$

که L تابع درستنمایی مدل، n اندازه نمونه و k تعداد پارامترهای برآورده شده توسط مدل است.

بیماری لوسمی ۱۲(سرطان خون) یکی از شایع‌ترین نوع بیماری سرطان است. این بیماری به صورت غیرعادی مقدار بسیار زیادی سلول خونی تولید می‌کند. این سلول‌های خونی عملکرد صحیحی نسبت به سلول‌های نرمال نداشته و باعث می‌شوند تولید سلول‌های سفید خون متوقف شده و توانایی فرد را در مقابله با بیماری‌ها به حداقل برسانند. بیماری لوسمی مغز استخوان به دو نوع حاد و مزمن تقسیم می‌شود که در این مطالعه دو نوع لوسمی لنفوبلاستیک حاد[۷] (ALL) و لوسمی میلوئید[۸] (AML) مورد مطالعه قرار خواهند گرفت. AML شایع‌ترین شکل لوسمی حاد است که ۳ تا ۴ برابر بیشتر در بزرگسالان نسبت به ALL شایع است، با این حال، بروز AML کمتر از انواع دیگر تومورها است. برای مطالعات بیشتر در مورد این دو نوع لوسمی می‌توان به منابع [۱۰] و [۹] مراجعه کرد.

پزشکان درصدند تا با مطالعه بر روی بیان‌های ژن بیماران بتوانند با استفاده از رفتارهای ژن‌های مشابه، راهکارهایی برای مطالعه و درمان این نوع سرطان خون بیابند. انتظار می‌رود با مطالعه بر روی مقادیر بیان ژن بیماران لوسمی ۱۲ از طریق علم آمار به روابط بین ژن‌ها پی برد.

با توجه به مطالب مطرح شده، ادامه مقاله حاضر به این ترتیب تدوین شده است که در بخش بعد، مطالب نظری مرتبط با روش تصویرهای تصادفی ارائه می‌شود. در بخش تحلیل مثال واقعی، نحوه پیاده‌سازی مطالب مطرح شده در این مقاله بر روی داده‌های بیان ژن مربوط به بیماری لوسمی ۱۲ تشریح می‌شود. در نهایت، نتیجه‌گیری و پیشنهادات آتی در راستای موضوع مقاله ارائه می‌شود.

[4] Orthogonal
[5] Gaussian Mixture Model
[6] Bayesian Information Criterion
[7] Acute Lymphatic Leukemia
[8] Acute Myeloid Leukemia

۲. ساختار ریاضی روش تصویر تصادفی

بنابه [۴] به منظور جلوگیری از ضعف‌های عدیده مربوط به فضاهای بُعد بالا، یک راه‌حل ساده برای رویارویی با مشکلات محاسباتی مرتبط با بُعد بالا افرازبندی بردار متغیر پیشگو است. از نقطه نظر ریاضی، می‌توان نوشت:

$$Y^* = [Y, Y_c] = [XA, X\bar{A}], \tag{۲}$$

که $Y \in \mathbb{R}^{n\times d}$ ماتریس داده‌های کاهش یافته موثر، $Y_c \in \mathbb{R}^{n\times(p-d)}$ ماتریس داده‌های کاهش یافته غیرموثر، $X \in \mathbb{R}^{n\times p}$ ماتریس داده‌های اولیه بُعد بالا، $A \in \mathbb{R}^{p\times d}$ ماتریس تصویر تصادفی و $\bar{A} \in \mathbb{R}^{p\times(p-d)}$ ماتریس متمم متعامد A است.

ایده اصلی چنین ترفندی این است که خوشه‌بندی مدل‌مبنا بر اساس داده‌های کاهش یافته یعنی $Y = XA$ انجام شود به شرط آن که ساختار خوشه‌ها توسط ماتریس‌های بلوکی d بُعدی که تقریب مناسبی برای Y^* هستند به خوبی قابل تبیین باشند. برای جزئیات بیشتر در این خصوص می‌توان به [۱] مراجعه کرد.

همانطور که در مقدمه بیان شد یکی از روش‌های موثر و جدید برای انجام خوشه‌بندی مدل‌مبنای داده‌های بُعد بالا استفاده از ابزار تصویرهای تصادفی است. ایده کلی آن که در برگیرنده تصویر کردن مرحله به مرحله، افرازبندی مناسب داده‌ها و خوشه‌بندی در هر گام بود، توصیف شد.

در این بخش، الگوریتمی با عنوان الگوریتم خوشه‌بندی تصویر تصادفی گروهی برای افرازبندی داده‌های اولیه X به گروه‌های G که توسط [۱] پیشنهاد شد، تشریح می‌شود که نحوه پیاده‌سازی خوشه‌بندی از طریق تصویرهای تصادفی را در عمل ممکن می‌کند. مراحل انجام الگوریتم به صورت زیر است:

- بر اساس یک اندازه خاص، B تصویر تصادفی مستقل d بُعدی تولید کنید. فرض کنید به ازای $b = ۱, \ldots, B$ آن‌ها را A_b بنامیم.
- مدل آمیخته گاوسی برای G مولفه را بر روی داده‌های تصویر شده یعنی $Y = XA_b$ برازش دهید.
- افراز داده‌های القا شده را ثابت نگه دارید.
- رگرسیون خطی Y_c بر روی Y- Y_c نقش متغیر پاسخ و Y نقش متغیر پیشگو را داشته باشد-را برازش دهید که $Y_c = X\bar{A}_b$ و $\bar{A}_b$ ماتریس متمم متعامد A_b است.
- معیار BIC برای مدل آمیخته گاوسی و رگرسیون خطی معرفی شده را به دست آوردید و آن دو را با هم جمع کنید.
- از میان تمامی تصویرها، تصویرهای تصادفی B^* که بالاترین مقادیر BIC را دارند، انتخاب کنید.
- با جمع‌بندی نتایج، بردار عضویت خوشه برای بهترین تصویرهای B^* را به دست آورید.
- داده‌های اولیه X را با توجه به عضویت توافقی مرحله‌ی قبل افرازبندی کنید.

به منظور جلوگیری از انتخاب خوشه‌های بسیار مشابه یا نادرست، محققین راه‌حل‌های متنوعی را برای این مشکل پیشنهاد کردند. به عنوان مثال، بر اساس نتایج عددی و تجربیات شبیه‌سازی، مقادیر $B = ۱۰۰۰$ و $B^* = ۱۰۰$ به عنوان مقادیر مناسب در [۱] پیشنهاد شد. لازم به ذکر است که B تعداد تصویرهای تولید شده و B^* تعداد تصویرهای بهینه است.

نشان داده شد که توزیع‌های مختلف با ابعاد بالا هنگامی که به طور تصادفی بر روی یک فضای بُعد پایین تصویر می‌شوند، بیشتر گاوسی به نظر می‌رسند [۶]. به‌علاوه، در [۵] ثابت شد که داده‌های حاصل از آمیزه دلخواه از G توزیع گاوسی می‌توانند به طور تصادفی در یک زیر فضایی از بُعد $O(\log G)$ نشانیده شوند، در حالی که ساختار گروه‌بندی داده‌ها تقریبا به طور کامل حفظ می‌شود. علاوه بر این، او نشان داد که اگر $d < \log G$،

آن‌گاه عملکرد تصویر کردن مطلوب نخواهد بود. دلیل این امر آن است که در صورت اتفاق چنین حالتی می‌توان نتیجه گرفت بُعد تصویرشده مستقل از بُعد اولیه داده‌ها است. به زبان ساده، یعنی اینکه، d به n و p بستگی ندارد. برای اخذ تصمیم نهایی در این مورد به نتایج تجربی [۱] استناد می‌کنیم که انتخاب $d = O(۱۰ \log G)$ را گزینه بسیار مناسبی برای اجرای روش RP اعلام کردند.

در بخش بعد در قالب تحلیل مثال واقعی خوشه‌بندی از طریق تصویر تصادفی و به ویژه عملکرد روش تصویرهای تصادفی گروهی و الگوریتم معرفی شده مورد ارزیابی قرار می‌گیرد.

۳. تحلیل مثال واقعی

در این مقاله داده‌های بیان ژن جمع‌آوری شده توسط [۷] مورد استفاده قرار گرفته است. این داده در بسته `multtest` در R موجود است. این داده شامل مقادیر بیان ۳۰۵۷ ژن از ۳۸ بیمار لوسمی ۱۲ است که ۲۷ بیمار دارای لوسمی ALL و ۱۱ بیمار لوسمی AML هستند.

به منظور کشف روابط بین ژن‌های بیماران، محاسبه ماتریس فواصل بین بیماران می‌تواند مفید فایده باشد. ماتریس فاصله برای ۳۸ بیمار مورد مطالعه در مثال واقعی محاسبه شد. طبق انتظار عناصر روی قطر اصلی صفر هستند و عناصر خارج قطر اگر مقدار بزرگی باشند به این معناست که عملکرد ژن های داده لوسمی دو بیمار مشابه نیستند و اگر مقدار فاصله آن‌ها کوچک باشد به این معناست که عملکرد ژن های دو بیمار مشابه هستند. طبق بررسی‌های انجام شده در بین تمامی مقادیر، کوچک‌ترین فاصله مربوط به سطر ۱۵ و ستون ۵ برابر ۲۷٫۷۶۳ است و بدان معنا است که ۱۵ و ۵امین بیمار دارای ژن‌های مشابه‌ای نسبت به هم در مقایسه با سایر بیماران هستند و احتمال آنکه در یک گروه از دو گروه لوسمی‌ها قرار بگیرند بیشتر است. همچنین بزرگترین فاصله مربوط به سطر ۳۳ و ستون ۲۱ و برابر ۶۲٫۶۱۸ است. لذا، ۳۳ و ۲۱امین بیمار دارای ژن‌های متفاوتی نسبت به هم در مقایسه با سایر بیماران هستند و احتمال آنکه در یک گروه قرار بگیرند بسیار کم است. بنابراین می‌توان از فاصله‌ی مورد نظر به این موضوع پی برد که دو بیمار مبتلا به یک نوع لوسمی از دو نوع لوسمی مربوطه هستند یا خیر. اما واضح است که کنار هم قرار دادن تمامی زوج فاصله و تصمیم‌گیری درباره خوشه‌بندی بیماران بر اساس فواصل دوبه‌دو ممکن نخواهد بود.

توجه شود که هدف اصلی از تحلیل داده‌های لوسمی این بود که با استفاده از اطلاعات مربوط به میزان بیان ژن و بدون بهره‌برداری از اطلاعات مربوط به نوع لوسمی هر یک از بیماران، افراد مورد مطالعه گروه‌بندی شوند. واضح است که روش‌های مرسوم تحلیل چند متغیره در برخورد با این داده‌ها به مشکل برخورد خواهند کرد. یکی از دلایل اصلی این امر آن است که بُعد داده‌ها (p) بیشتر از تعداد نمونه (n) است. به عبارتی علمی‌تر، داده‌های مورد مطالعه در این مقاله از نوع بُعد بالا هستند. از طرفی دیگر، کار کردن با تعداد ۳۰۵۷ متغیر عملا بسیار سخت و زمان‌بر است، در نتیجه برای یافتن نتایج قابل ملاحظه بهتر است یا از بین تمامی متغیرها آنهایی که از اهمیت بالایی برخوردار هستند را انتخاب کرد یا اینکه بُعد داده‌ها را به طریقی کاهش داد.

مقاله حاضر به دنبال روشی بوده است که کمترین اطلاعات مفید را از دست دهد و رویکرد RP دارای این قابلیت است. در این قسمت از آن برای کاهش بُعد داده‌های لوسمی و سپس تحلیل‌های آماری مرتبط استفاده می‌شود. برای اجرای رویکرد تصویرهای تصادفی از الگوریتم $RPEClu$ با $B = ۱۰۰۰$، $B^* = ۱۰۰$ و $d = ۸$ استفاده شد. به علاوه، بنا به اطلاعات قبلی راجع به داده‌های این مقاله، تعداد گروه‌ها (خوشه‌ها) ثابت و برابر ۲ در نظر گرفته شد. با اعمال این محدویت‌ها و با اجرای الگوریتم، معیار ARI برابر با ۱ به دست آمد. این معیار، عددی بین ۱− تا ۱ را اختیار می‌کند. در حالتی که ARI برابر با ۱ باشد، مطابقت کامل بین برچسب‌های واقعی و خوشه‌ای وجود دارد و در مقابل اگر مقدار این

شاخص برابر با ۱− باشد نشانگر برچسب‌گذاری تصادفی در حین خوشه‌بندی است. خلاصه‌ای از نتایج حاصل از اجرای این الگوریتم در جدول ۱ مشاهده می‌شود. میانگین $BIC.GMM$ نشان دهنده میانگین BIC تصویرهای تولید شده مرتبط با مدل‌های آمیخته گاوسی، $BIC.reg$ نشان

جدول ۱: شرح خلاصه‌ای از نتایج الگوریتم $RPEClu$ برای خوشه‌بندی داده‌های لوسمی.

معیارهای خوشه‌بندی	مقدار
ARI	۰٫۸۹
میانگینBIC	−۱۷۱۲۲۷۸۲
میانگین $BIC.GMM$	−۵۹۱٫۴۷۸۴
میانگین $BIC.reg$	−۱۷۱۲۲۱۹۰

دهنده میانگین BIC تصویرهای تولید شده متناسب با مدل‌های رگرسیون خطی و میانگین BIC نشان دهنده میانگین BIC تصویرهای تولیده مرتبط با مدل کلی و اولیه است. مقدار ARI به دست آمده نشانگر کیفیت روش خوشه‌بندی مورد مطالعه است. مقدار آن تاییدی بر این است که خوشه‌بندی روش RP برای داده‌های لوسمی منجر به نتایج رضایت‌بخشی شده است.

نتیجه‌گیری

روش‌های متفاوتی برای غلبه بر مشکلات خوشه‌بندی داده‌های بُعد بالا به‌ویژه داده‌های مربوط به تحلیل ژن‌ها وجود دارد. یکی از رویکردهای اخیر برای مقابله با چنین مشکلی، روش RP است که در این مقاله از هر دو جنبه نظری و کاربردی مورد بررسی قرار گرفت. در این روش برای انجام خوشه‌بندی از الگوریتم خاصی استفاده می‌شود. همچنین عملکرد این الگوریتم در یک مثال واقعی بر اساس شاخص ارزیابی خوشه‌بندی نشان داده شد.

مراجع

[1] Anderlucci, L., Fortunato, F. and Montanari, A. (2019), High-dimensional Clustering via Random Projections, Preprint ArXiv, 1909.10832.

[2] Bellman, R. (1957),Dynamic Programming, Princeton University Press, Los Angeles.

[3] Bouveyron, C., Girard, S. and Schmid, C. (2007), High-dimensional Data Clustering, Computational Statistics and Data Analysis,52, 502-519.

[4] Cannings, T. I. and Samworth, R. J. (2017), Random Projection Ensemble Classification, Journal of the Royal Statistical Society, Series B (Statistical Methodology),79, 959–1035.

[5] Dasgupta, S. (2000), Experiments with Random Projection, In Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence, UAI '00, pages 143–151, San Francisco.

[6] Diaconis, P. and Freedman, D. (1984), Asymptotics of Graphical Projection Pursuit. The Annals of Statistics, 701, 793-815.

[7] Golub, T. R. Slonim, D. K. Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., and Bloomfield, C. D. (1999), Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, Science,286, 531-537.

[8] Johnson, W. B. and Lindenstrauss, J. (1984), Extensions of Lipschitz Mappings in to a Hilbert Space, Contemporary Mathematics, 26, 189–206.

[9] Khwaja, A., Bjorkholm, M., Gale, R. E., Levine, R. L., Jordan, C. T., Ehninger, G., Linch, D. C. and others, (2016), Acute myeloid leukaemia, Nature reviews Disease primers, 2, .1-22

[10] Wang, T., Hamann, W. and Hartge, R (1983), Structural aspects of a placenta from a case of maternal acute lymphatic leukaemia, Placenta, 4, 185-195.

Detection and Classification of Effective Genes Based on Random Projection Method

Nourani Pileh Roud, S., Golalizadeh, M.

Depatment of Statistics, Tarbiat Modares University, Tehran, Iran.

sedigheh.noorani@modares.ac.ir

***Abstract*— Research on gene expression clustering has long been of interest to many researchers. Gene-related data are usually high-dimensional, meaning that the number of variables exceeds the number of observations. Clustering such data in conventional ways poses particular challenges. To deal with such problems, a new approach that has attracted the attention of countless researchers is the random projection method. This approach uses a mixed Gaussian model, but a special algorithm is used to accurately select these projections. In the present paper, the performance of this method has been studied from both theoretical and practical perspectives and its superiority over the analysis of real data related to the grouping of leukemia 12 patients has been shown.**

Keywords*—*High-dimensional Data, Random Projection, Gaussian Mixed Model, Gene Expression Data.