Journal of Statistical Sciences, Spring and Summer, 2022
Vol. 16, No. 1, pp 239-252
DOI: 10.29252/jss.16.1.239

Comparison of Clustering High Dimensional Data by Random Projections Method and Some Common Methods of Dimensional Reduction

Nourani Pileh Roud, S., Golalizadeh, M.
Depatment of Statistics, Tarbiat Modares University.

Corresponding author: M. Golalizadeh, golalizadeh@modares.ac.ir

Received: 26 October 2021 **Revised:** 22 December 2021 **Accepted and Published Online:** 2 January 2022.

Introduction

The clustering of the high dimensional data is usually encountered with problems such as the curse of dimensionality. To overcome such obstacles, dimensionality reduction methods are often used. This view is typically referred to by two approaches; variable selection and variable extraction. Recently, researchers proposed a way that is claimed to lose less information in clustering high-dimensional data than other techniques. Among them, that presented by Anderlucci et al. (2021) under the title of Random Projections is very popular. The RP method is based on creating random projections, selecting a small subset, and then performing clustering tasks. Comparison and superiority of this method with conventional approaches of dimensionality reduction, using four critical criteria of clustering including adjusted Rand index, Jaccard index, Fowlkes-Malo index and the accuracy index is performed on three gene expression datasets in this article.

Material and Methods

One of the variable selection methods is the variable selection approach for clustering based on the Gaussian model. On the other hand, the principal components analysis method is one of the most popular methods for extracting variables. Another practical, new and exciting approach to performing dimensionality reduction is the Random Projections method. Using a group random projections, Andrelucci et al. (۲۰۲۱) proposed clustering algorithm to cluster the high-dimensional data. This algorithm obtains the final output

through Gaussian mixture model clustering applied to the optimal subset of random projections. Then, the original high-dimensional data is mapped onto the reduced spaces. Finally, model selection criteria are calculated for them and observations are clustered using optimal projections.

Results and Discussion

In this paper, the proposed methods by Anderlucci et al. (2021) are described and compared on three gene expression datasets, including leukaemia, lymphoma, and prostate cancers. Based on the gained results, using the introduced criteria, both competing methods have lower values than the random projections method and therefore have weaker performance. The final result is that the random projections method performs better for the three mentioned datasets. It should be noted that the purpose of the current study was only to compare the performance of clustering based on the three mentioned approaches and some different clustering criteria. So, other analytical aspects related to the random projection were not considered. Further exploration of these methods will be followed in our future research.

Conclusion

Clustering of high-dimensional data faces different statistical challenges, and various methods exist to overcome the related problems. One of these practical tools is reducing the data dimension. This article examined the random projection from both theoretical and practical aspects. Also, its performance was evaluated on three real data sets and compared with other standard methods, and its superiority was shown based on several conventional indicators of clustering measures. To conduct future research, one can address the probabilistic aspects of the random projections approach by considering proper statistical inference methods.

Keywords: High dimensional data, Model-based clustering, Dimension reduction methods, Random Projections.

Mathematics Subject Classification (2010): 62H25, 62H30.

مجله علوم آماری، بهار و تابستان ۱۴۰۱
جلد ۱۶، شماره ۱، ص ۲۳۹ – ۲۵۲
DOI: 10.29252/jss.16.1.239
مقاله پژوهشی

مقایسه عملکرد خوشه‌بندی داده‌های بُعد بالا توسط روش تصویرهای تصادفی و برخی روش‌های مرسوم کاهش بُعد

صدیقه نورانی پیله‌رود، موسی گلعلی‌زاده
گروه آمار، دانشگاه تربیت مدرس

نویسنده مسئول : موسی گلعلی‌زاده، golalizadeh@modares.ac.ir

تاریخ دریافت: ۱۴۰۰/۰۸/۴ تاریخ بازنگری: ۱۴۰۰/۱۰/۱ تاریخ پذیرش و انتشار: ۱۴۰۰/۱۰/۱۲

چکیده: امروزه مشاهدات اندازه‌گیری شده در بسیاری از حوزه‌های علمی، مثل علوم زیستی اغلب بُعد بالا هستند، به این معنی که تعداد متغیرها از تعداد نمونه بیشتر است. یکی از مشکلاتی که در خوشه‌بندی مدل‌مبنای اینگونه داده‌ها رخ می‌دهد برآورد تعداد زیادی پارامتر است. برای رفع چنین مشکلی، ابتدا باید بُعد داده‌ها را قبل از خوشه‌بندی کاهش داد که این امر می‌تواند از طریق روش‌های کاهش بُعد انجام شود. یک رویکرد اخیر در این زمینه که مورد توجه فراوان قرار گرفته روش تصویرهای تصادفی است. در مقاله حاضر این روش از هر دو منظر نظری و کاربردی مورد بررسی قرار گرفته و برتری آن در مقایسه با برخی رویکردهای مرسوم کاهش بُعد مانند تحلیل مولفه‌های اصلی و روش انتخاب متغیر در تحلیل سه مجموعه داده واقعی نشان داده شده است.

واژه‌های کلیدی: داده‌های بُعد بالا، خوشه‌بندی مدل‌مبنا، روش‌های کاهش بُعد، تصویرهای تصادفی.
کد موضوع‌بندی ریاضی (۲۰۱۰): 62H25، 62H30.

۱ مقدمه

خوشه‌بندی به عنوان یکی از مهم‌ترین مولفه‌های تحلیل داده آماری، نقش بسزایی در اخذ اطلاعات مفید از مجموعه داده‌ها دارد. همانطور که ژو و تیان (۲۰۱۵) اشاره کردند هر الگوریتم خوشه‌بندی به دلیل پیچیدگی اطلاعات، نقاط قوت و ضعف خاص خود را دارد. با این حال، کماکان خوشه‌بندی هم به دلیل تسهیل تحلیل داده‌ها و هم به خاطر

ارائه نمایش تصویری مناسب از دسته‌بندی مشاهدات از محبوبیت بالایی برخوردار است. به عنوان مثال، نمونه‌ای از کاربرد آن در آمارشکل در نبیل و گل‌علی‌زاده (۱۳۹۳) نشان داده شده است. با پیشرفت تکنولوژی محاسباتی، در بسیاری از حوزه‌های علمی مشاهده می‌شود که اکثر داده‌های جمع‌آوری شده اخیر از گونه بُعد بالا[۱] هستند. اطلاق چنین مفهومی به داده‌ها در واقع به وجود تعداد بیشتر متغیر (p) در مقایسه با تعداد نمونه (n) برمی‌گردد. انجام خوشه‌بندی داده‌های بُعد بالا از هر دو جنبه نظری و کاربردی مشکلاتی را به همراه دارد (بوقژن و همکاران، ۲۰۰۷). یکی از شاخص‌ترین این مشکلات پدیده مشقت‌بُعدچندی[۲] است که توسط بلمن (۱۹۵۷) معرفی شده است. از جمله راهکارهای مقابله با چنین مشکلاتی استفاده از روش‌های کاهش بُعد[۳] قبل از انجام خوشه‌بندی است. برای اجرای این امر، روش‌های انتخاب متغیر[۴] و استخراج متغیر[۵] از محبوبیت بالایی برخوردارند (جولیفه، ۱۹۸۶). می‌توان حدس زد که روش‌های کاهش بُعد نیز مشکلاتی مانند از دست دادن برخی اطلاعات مهم را به همراه دارند اما به دلیل کارایی مناسب آن‌ها کماکان مورد توجه محققین است.

اخیرا محققین روشی را پیشنهاد دادند که ادعا می‌شود در مقایسه با سایر رویکردها اطلاعات کمتری را در خوشه‌بندی داده‌های بُعد بالا از دست می‌دهد و در نتیجه نسبت به سایر روش‌های کاهش بُعد از اهمیت بیشتری برخوردار است. روش مورد اشاره که تحت عنوان تصویرهای تصادفی[۶] (RP) توسط آندرلوچی و همکاران (۲۰۲۱) مطرح شد مبتنی بر ایجاد تصویرهای تصادفی، انتخاب زیرمجموعه بسیار کوچک و سپس انجام خوشه‌بندی است. ایده اصلی این روش توسط کانینگز و سامورث (۲۰۱۷) مطرح شد که کارایی مناسب آن را در مسئله رده‌بندی راهنماییده به نمایش گذاشتند.

مقاله حاضر به تشریح کامل نظری و کاربردی روش تصویرهای تصادفی می‌پردازد. چون رویکرد اصلی این روش براساس خوشه‌بندی مدل‌مبنا است، برای درک بهتر روش مورد اشاره از یک مدل آمیخته گاوسی[۷] (GMM) بهره برده می‌شود. لذا، نحوه تعامل این رویکرد با ساختار نظری GMM نیز تشریح می‌شود. علاوه بر آن، مقایسه و برتری روش RP با رویکردهای مرسوم کاهش بُعد، با استفاده از چهار معیار مهم خوشه‌بندی با عنوان‌های شاخص رَند تعدیل شده[۸] (ARI)، شاخص جاکارد[۹] (JI)، شاخص فاولکس مالو[۱۰] (FMI) و شاخص دقت[۱۱] انجام می‌شود. لازم به اشاره است که این شاخص‌ها معیاری بر اساس هم‌خوانی خوشه‌بندی صورت گرفته توسط مدل مورد مطالعه و گروه‌بندی واقعی داده‌ها هستند و عددی بین ۱− تا ۱ را اختیار می‌کنند. همانطور که انتظار می‌رود مقادیر

[1] High Dimension
[2] Curse of Dimensionality
[3] Dimension Reduction
[4] Variable Selection
[5] Variable Extraction
[6] Random Projections
[7] Gaussian Mixture Model
[8] Adjusted Rand Index
[9] Jaccard Index
[10] Fowlkes Mallows Index
[11] Accuracy

کم این معیارها نمایشگر عملکرد ضعیف خوشه‌بندی و مقادیر بزرگ آن‌ها شاهدی از خوشه‌بندی خیلی خوب است (ژو و تیان، ۲۰۱۵).

با توجه به مطالب مطرح شده، ادامه مقاله حاضر به این ترتیب تدوین شده است که در بخش بعد، مطالب کلی در خصوص برخی از مشکلات مرتبط با خوشه‌بندی داده‌های بُعد بالا و راه‌های مقابله با آن ارائه می‌شود. در ادامه درباره روش‌های کاهش بُعد و به ویژه رویکردهایی مانند روش‌های انتخاب و استخراج متغیر بحث و نقاط ضعف و قوت آنان نیز مطرح خواهد شد. سپس، مطالب نظری مرتبط با روش RP ارائه می‌شود. در بخش تحلیل مثال واقعی، نحوه پیاده‌سازی مطالب مطرح شده در این مقاله بر روی سه مجموعه داده بیان ژنی شامل سرطان‌های لوسمی۱۲، لنفوم و پروستات تشریح و مقایسه آن با روش‌های مرسوم دیگر انجام می‌شود. در نهایت، نتیجه‌گیری و پیشنهادات آتی در راستای موضوع مقاله ارائه می‌شود.

۲ مشکلات خوشه‌بندی داده‌های بُعد بالا

به اذعان بیشتر محققینی که در حوزه تحلیل داده‌های بُعد بالا فعالیت دارند تحلیل آماری اینگونه داده‌ها دارای مشکلات زیادی است. می‌توان گفت که به طور کلی، اکثر روش‌های آماری مرسوم، -بدون انجام هیچ تعدیلی- عملکرد ضعیفی را در فضاهای بُعد بالا از خود نشان می‌دهند. بنابه فرن و برودلی (۲۰۰۳) ابعاد بالا باعث ایجاد دو چالش در پیاده سازی الگوریتم‌های یادگیری ناراهنماییده[1] از جمله تحلیل خوشه‌ای می‌شوند. منظور از یادگیری ناراهنماییده تحلیلی است که در آن داده‌ها بدون دخالت متغیرهای پاسخ برچسب‌گذاری می‌شوند. چالش اول این است که وجود برخی از متغیرها که تاثیری در امر خوشه‌بندی ندارند ممکن است اجرای الگوریتم خوشه‌بندی را دچار سردرگمی کند. چالش دوم به پراکندگی داده‌ها در ابعاد بالا یا به زبانی دیگر پیچیدگی ساختار داده‌ها برمی‌گردد. بنابراین، یافتن هر ساختاری که متناسب با تغییرات داده‌ها باشد برای این چنین الگوریتم‌هایی سخت و دشوار است. برای غلبه بر اینگونه مشکلات در تحلیل داده‌های بُعد بالا و بویژه امر خوشه‌بندی رویکردهای مختلفی پیشنهاد شده است. روش‌های کاهش بُعد قبل از مرحله خوشه‌بندی مورد توصیه فراوان قرار گرفته است. از بین روش‌های موجود، روش‌های استخراج متغیر یا روش‌های انتخاب متغیر از محبوبیت بالایی بین محققین برخوردار است.

۳ روش‌های کاهش بُعد

بنا به کانینگهام (۲۰۰۸)، استفاده از روش‌های کاهش بُعد به عنوان یک مرحله قبل از پردازش داده‌ها برای ساده‌سازی مدل‌بندی داده‌های بُعد بالا ضروری است. مطالعه منابع علمی بی‌شمار نشان می‌دهد که اصطلاح کاهش بُعد را می‌توان به دو رویکرد متفاوت زیر تقسیم‌بندی کرد:

الف : روش انتخاب متغیر که در آن با انتخاب برخی متغیرها می‌توان مشاهدات مربوط به بُعد بالا را در یک

[1] Unsupervised Learning

زیرفضای پایین‌تر (ابعاد پایین‌تر) قرار داد. رافتری و دین (۲۰۰۶)، موگیس و همکاران (۲۰۰۹) و فوپ و مورفی (۲۰۱۸) منابع مفیدی برای مطالعه این روش هستند. یکی از روش‌های انتخاب متغیر، رویکرد انتخاب متغیر برای خوشه‌بندی مبتنی بر مدل گاوسی است. این رویکرد که با کلمه اختصاری $Clustvarsel$ در تحلیل‌های آماری مبتنی بر کتابخانه‌های کامپیوتری در نرم‌افزار R مشخص می‌شود، یکی از روش‌های بسیار کارا برای بررسی متغیرهای تاثیرگذار است. با این حال، باید توجه داشت که این روش از نظر بار محاسباتی بسیار سنگین و اجرای آن به مدت زمان طولانی نیاز دارد.

ب : روش استخراج متغیر که در آن متغیرهای اولیه در تحلیل را به طریق مناسبی با هم ترکیب می‌شوند تا بیشترین اطلاعات ممکن راجع به داده‌های اولیه را در خود حفظ کنند. روش تحلیل مولفه‌های اصلی[1] (PCA) یکی از محبوب‌ترین روش‌ها برای استخراج متغیر است که اولین بار توسط پیرسون (۱۹۰۱) معرفی شد. این رویکرد با استفاده از تبدیل (ترکیب خطی یا غیرخطی) مجموعه داده‌های اولیه به مجموعه جدیدی از متغیرها عمل کاهش بُعد را انجام داده و همزمان تغییرات موجود در داده‌ها را نیز به خوبی توصیف می‌کند. به عبارت دیگر می‌توان گفت که نتیجه رویکرد PCA تعداد کمی متغیرهای جدید است که می‌توانند جایگزین مناسبی برای تعداد زیادی متغیر اولیه باشند (اوریت و هاترن، ۲۰۱۱).

یکی دیگر از رویکردهای موثر، جدید و مورد توجه برای انجام کاهش ابعاد روش RP است. این روش، یک روش هندسی ساده برای کاهش بُعد مجموعه‌ای از نقاط در فضای اقلیدسی است که تقریبا فاصله‌های دو به دو نقاط در فضای تصویر شده را حفظ می‌کند و در مقایسه با سایر روش‌ها دارای ساختار ساده و خطای محاسباتی به مراتب کمتر است (بینگهام و مانیلا، ۲۰۰۱). این روش از تصویر کردن داده‌های اولیه بُعد بالا به ابعاد پایین‌تر با استفاده از یک ماتریس تصادفی که دارای ستون‌های متعامد با طولی واحد هستند، بهره می‌برد. برای جزئیات بیشتر به آندرلوچی و همکاران (۲۰۲۱) مراجعه شود. روش RP علاوه بر آنکه بُعد داده‌های بُعد بالا را به طرز چشمگیری کاهش می‌دهد می‌تواند اطلاعات موجود در داده‌های اولیه را نیز تقریبا به طور کامل حفظ کند. داس‌گوپتا (۲۰۰۰) نشان داد که چنین نتیجه‌ای توسط لم جانسون و لیندنستراوس (۱۹۸۴) تضمین می‌شود. یکی از کارکردهای مستقیم این لم در استفاده از روش RP آن است که اگر دو چگالی در فضای بُعد بالا از هم فاصله کافی داشته باشند، انتظار می‌رود در فضای کاهش یافته d بُعدی تقریبا همان فاصله حفظ شود. این نتیجه به همراه ایده کانینگز و ساموث (۲۰۱۷) در تحلیل رده‌بندی راهنماییده، کمک می‌کند تا مجموعه‌ای از تصویرهای تصادفی مستقل با بُعد کم تولید کرده و برای هر حالت یک GMM را به کار برد.

[1] Principle Component Analysis

۴ خوشه‌بندی به روش تصویرهای تصادفی

یکی از عمده‌ترین ضعف‌های تصویرهای تصادفی ناپایداری[1] است، به معنای آن که تصویرهای تصادفی متفاوت ممکن است نتایج خوشه‌بندی کاملا متنوعی از داده‌های اولیه را به همراه داشته باشند. برای مقابله با این چالش و همچنین برای انتخاب دقیق تصویرهای تصادفی از داده‌های اولیه بهتر است از یک GMM استفاده شود. طبیعی است که برخی از این تصویرها می‌توانند یک ساختار خوشه‌ای کاملا مشخصی را از داده‌های اولیه در فضای کاهش یافته نشان دهند. بنابه کانینگز و سامورث (۲۰۱۷) به منظور جلوگیری از ضعف‌های عمده 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: $T = [Y, Y_c] = [XA, X\bar{A}]$ که در آن $Y \in \mathbb{R}^{n\times d}$ 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 T هستند به خوبی قابل تبیین باشند. آنگاه کل مجموعه داده‌های کاهش یافته توسط یک GMM با تابع چگالی

$$f_Y(y|z) = \sum_{k=1}^{G} \pi_k \phi_k(y|\theta_{Y_k}, z), \qquad (۱)$$

مدل‌بندی می‌شوند که در آن ϕ_k نماد توزیع نرمال چندمتغیره با بردار پارامتر $\theta_{Y_k} = (\mu_{Y_k}, \Sigma_{Y_k})$ و G تعداد خوشه مورد نیاز است. شرط بیان شده و افرازبندی مربوطه بیانگر آن هستند که اطلاعات مربوط به عضویت نمونه‌ها در خوشه مشخص مثل Z توسط Y_c خیلی شبیه به آن چیزی است که در بردار Y وجود دارد. بنابراین، Y_c و Z به شرط Y مستقل از هم هستند. به دلیل آنکه بردار T دارای توزیع نرمال چندمتغیره است، تابع چگالی شرطی Y_c به شرط Y، یعنی $f(y_c|y)$؛ نیز از توزیع نرمال چندمتغیره پیروی می‌کند. پس داریم: $f(y_c|y) = \phi_{Y_c}(y_c|\theta_{Y_c|Y}, y)$. یعنی Y_c به شرط Y دارای توزیع نرمال چندمتغیره با بردار پارامتر $\theta_{Y_c|Y} = (\mu_{Y_c|Y}, \Sigma_{Y_c|Y})$ است. پس می‌توان نوشت:

$$\begin{aligned} \phi(y_c; \mu_{Y_c|Y}, \Sigma_{Y_c|Y}) &= (2\pi)^{-(\frac{p-d}{2})} |\Sigma_{Y_c|Y}|^{-(\frac{1}{2})} \\ &\times \exp\Big\{ -\frac{1}{2}(y_c - \mu_{Y_c|Y})' \Sigma_{Y_c|Y}^{-1} (y_c - \mu_{Y_c|Y}) \Big\}. \end{aligned} \qquad (۲)$$

[1] Unstable

لازم به اشاره است که توزیع T از حاصل ضرب چگالی حاشیه‌ای Y، یعنی $f(y|z)$، و چگالی شرطی $Y_c|Y$، یعنی $f(y_c|y)$ به صورت

$$f(t|z) = \Big[\sum_{k=۱}^{G} \pi_k \phi_k(y|\theta_{Y_k}, z)\Big] \phi_{Y_c}(y_c|\theta_{Y_c|Y}, y) = \sum_{k=۱}^{G} \pi_k \phi_k(t|\theta_{T_k}, z),$$

به دست می‌آید، که در آن $\theta_{T_k} = (\mu_{T_k}, \Sigma_{T_k})$. تابع لگاریتم درستنمایی تابع چگالی اخیر به صورت

$$\log L(z|T) = \sum_{i=۱}^{n} \log \Big[f(T_i|z)\Big] = \sum_{i=۱}^{n} \log \Big[\sum_{k=۱}^{G} \pi_k \phi_k(T_i|\theta_{T_{ik}}, z)\Big]. \qquad (۳)$$

است. اگر هدف انتخاب تصویرهای مناسب برای رویکرد RP باشد محاسبه منفی معیار اطلاع بیزی[1] (BIC) به کمک رابطه (۳) از طریق تساوی

$$\begin{aligned} BIC_{GM}(T) &= ۲\log L(z|t) - q_T \log(n), \\ &= ۲\Big[\log L(z|y) + \log L(y|y_c)\Big] - q_T \log(n), \end{aligned}$$

صورت می‌گیرد، که در آن q_T تعداد پارامترهای برآورد شدهٔ مدل کلی است. حال به دلیل آنکه T افرازبندی شده است، تعداد پارامترهای برآورد شده مدل کلی شامل مجموع تعداد پارامترهای هر دو مدل مرتبط با Y و Y_c است. اگر تعداد پارامترهای مدل‌های مرتبط با T، Y و Y_c به ترتیب با نمادهای q_T، q_Y و q_{Y_c} مشخص شوند، آنگاه: $q_T = p + \frac{p(p+۱)}{۲}$، $q_{Y_c} = (p-d)(d+۱) + \frac{(p-d)[(p-d)+۱]}{۲}$ و $q_Y = d + \frac{d(d+۱)}{۲}$. واضح است که q_T از حاصل جمع q_Y و q_{Y_c} است. در نتیجه می‌توان نوشت:

$$\begin{aligned} BIC_{GM}(T) &= ۲\log L(z|y) - q_Y \log(n) + ۲\log L(y|y_c) - q_{Y_c} \log(n), \\ &= ۲\log[f(y)] - q_Y \log(n) + ۲\log[f(y_c|y)] - q_{Y_c} \log(n), \\ &= BIC_{GMM}(Y) + BIC_{reg}(Y_c|Y), \end{aligned}$$

همانطور که ملاحظه می‌شود اولین BIC در آخرین تساوی مربوط به یک GMM برای داده‌های d بُعدی است. از طرفی دیگر، توزیع شرطی $Y_c|Y$ نرمال بوده و میانگین این توزیع را، معمولا، رگرسیون Y_c برحسب Y می‌نامند. لذا، می‌توان دومین BIC را مرتبط با یک مدل رگرسیون خطی از $p-d$ ستون‌های آخر ماتریس T دانست که d تای اول آن مدنظر تحلیل است. توجه شود که در خوشه‌بندی داده‌های بُعد بالا، بُعد Y معمولا خیلی کوچکتر از بُعد

[1] Bayesian Information Criterion

Y_c است. از اینرو، بُعد Y تاثیر کمتری در BIC_{reg} دارد.

به منظور اعمال انعطاف پذیری بیشتر رویکرد اخیر، فرض می‌شود $\Sigma_{Y_c|Y}$ یک شکل کلی دارد، یعنی هیچ محدودیتی بر روی تمامی عناصر ماتریس Σ جز مثبت بودن عناصر روی قطر اعمال نمی‌شود. توجه شود که شرط مثبت معین بودن Σ کماکان ضروری است. تعداد پارامترهای ماتریس $\Sigma_{Y_c|Y}$ برابر $\frac{(p-d)[(p-d)+1]}{2}$ است. اما هنگامی که بُعد فضای داده‌های اولیه (p) در مقایسه با بُعد فضای کاهش یافته (d) زیاد باشد، یک شکل ساده‌تری برای $\Sigma_{Y_c|Y}$ اختیار می‌شود. به عنوان مثال، می‌توان نوشت: $\Sigma_{Y_c|Y} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_{p-d}^2)$. به دلیل آنکه فقط عناصر روی قطر اصلی ماتریس Σ مدنظر بوده و سایر عناصر آن صفر است، تعداد پارامترهای ماتریس کواریانس برابر $p-d$ خواهد شد. در نتیجه تعداد پارامترهای آزاد برای مدل رگرسیون خطی به عبارت $q_{Y_c} = (p-d)(d+1) + (p-d)$ کاهش می‌یابد.

برای افرازبندی داده‌های اولیه X به G خوشه، الگوریتمی تحت عنوان الگوریتم خوشه‌بندی تصویرهای تصادفی گروهی با نام اختصاری `RPEClust` در کتابخانه `RPEClust` توسط آندرلوچی و همکاران (۲۰۱۹) پیشنهاد شد. این الگوریتم، از طریق نتایج حاصل از خوشه‌بندی GMM اعمال شده در زیرمجموعهٔ بهینه از تصویرهای تصادفی، افراز نهایی را به دست می‌آورد. اولین گام این روش دربرگیرنده محاسبه تعدادی ماتریس‌های تصادفی است. پس از ایجاد این تصویرها، داده‌های اولیه بُعد بالا باید روی فضاهای کاهش یافته نشانیده[1] شوند. سرانجام، مقادیر BIC برای آن‌ها محاسبه شده و مشاهدات باید با استفاده از تعدادی تصویرهای بهینه دسته‌بندی شوند. نحوه محاسبه تعداد تصویرهای تولید شده (B)، بهینه (B^*) برای رسیدن به خوشه مطلوب و بُعد مورد نیاز برای فضای تصویر شده (d) دارای پیچیدگی‌های نظری خاصی است اما آندرلوچی و همکاران (۲۰۲۱) مقادیر $B = 1000$، $B^* = 100$ و $d = O(10 \log G)$ را به عنوان مقادیر بهینه توصیه کردند.

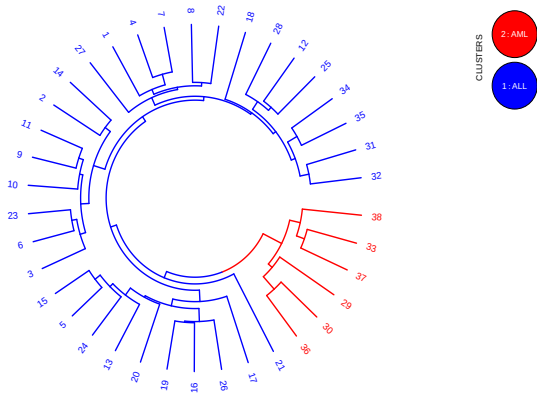
۵ تحلیل مثال‌های واقعی

در این مقاله سه مجموعه داده بیان ژنی با رویکردهای متفاوت شامل، داده‌های لوسمی۱۲ جمع‌آوری شده توسط گلوپ و همکاران (۱۹۹۹) موجود در کتابخانه `multtest`، مجموعه داده‌های لنفوم تهیه شده توسط چونگ و کلس (۲۰۱۰) موجود در کتابخانه `spls` و مجموعه داده‌های پروستات معرفی شده توسط سینگ و همکاران (۲۰۰۲) موجود در کتابخانه `flare` از نرم افزار R مورد استفاده قرار می‌گیرند.

داده‌های لوسمی۱۲ شامل مقادیر بیان ۳۰۵۷ ژن از ۳۸ بیمار لوسمی۱۲ (شایع‌ترین نوع سرطان خونی) است که ۲۷ بیمار از نوع سرطان لوسمی لنفوبلاستیک حاد(ALL)[2] و ۱۱ بیمار لوسمی میلوئید(AML)[3] رنج می‌برند. از مجموعه داده‌های لنفوم (لنفوم، در واقع، سرطان سیستم لنفاوی است که بخشی از شبکه مبارزه با میکروب دفاعی

[1] Embedded
[2] Acute Lymphatic Leukemia
[3] Acute Myeloid Leukemia

بدن را تشکیل می‌دهد)، ۴۲ نفر مبتلا به سرطان نوع لنفوم بزرگ سیال سلول B [۱] $(DLBCL)$، ۹ نفر نوع لنفوم فولیکولار [۲] (FL) و ۱۱ نفر نوع لوسمی لنفوسیتی مزمن [۳] (CLL) هستند. داده‌های پروستات (غدهٔ مهمی در سیستم تولید مثل مردان) شامل ۵۲ تومور پروستات و ۵۰ نمونه طبیعی هستند. لازم به ذکر است که برای کاهش فرآیند محاسبات از تومور پروستات ۲۷ تومور و از نمونه طبیعی ۲۳ نمونه انتخاب شده است. توجه شود که تعداد خوشه برای سه مجموعه داده مورد مطالعه در این مقاله معلوم بوده، لذا از آن در طی فرآیند خوشه‌بندی استفاده می‌شود. بنا به توصیف داده‌ها، تعداد خوشه‌های مجموعه داده‌های لوسمی۱۲ و پروستات برابر ۲ و مجموعه داده لنفوم برابر ۳ است. انتظار می‌رود با مطالعه بر روی مقادیر ژن سه مجموعه داده ذکر شده، از طریق علم آمار و به ویژه تحلیل خوشه‌ای به نتایج مناسبی رسید. برای اخذ اطلاعات اولیه از داده‌ها، روش خوشه‌بندی سلسله‌مراتبی برای هر سه مجموعه داده مورد استفاده قرار گرفت. به عنوان مثال، با اعمال این روش، نمودار درختواره مربوط به مجموعه داده لوسمی۱۲ در شکل ۱ به دست آمد که وجود دو نوع خوشه (گروه)، همان‌گونه که انتظار می‌رفت، را می‌توان به وضوح مشاهده کرد. بنا به رویکرد تحلیل داده‌های بُعد بالا، بررسی نقش متغیرهایی که منجر به چنین خوشه‌بندی تفکیکی شده‌اند نیز حائز اهمیت است که در ادامه به بررسی این موضوع پرداخته می‌شود.



شکل ۱. نمودار درختواره مربوط به مجموعه داده لوسمی۱۲.

چون داده‌های مورد مطالعه در این مقاله همگی از نوع بُعد بالا هستند، تحلیل آنها از طریق روش‌های مرسوم تحلیل چندمتغیره ممکن نیست. 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. برای اجرای رویکرد

[1] Diffuse Large B-cell Lymphoma
[2] Follicular Lymphoma
[3] Chronic Lymphocytic Leukemia

نوین RP از الگوریتم $RPEClu$ با $B = ۱۰۰۰$، $B^* = ۱۰۰$ و $d = \{۸, ۱۲\}$ استفاده شد.

جدول ۰۱. معیارهای روش‌های کاهش بُعد در خوشه‌بندی داده‌ها

مجموعه داده	روش	معیارها			
		ARI	JI	FMI	$Accuracy$
لوسمی۱۲	PCA	−۰/۱۰	۰/۴۴	۰/۶۱	۰/۴۹
	$Clustvarsel$	−۰/۰۲	۰/۳۵	۰/۵۲	۰/۴۹
	RP	۰/۸۹	۰/۹۱	۰/۹۵	۰/۹۵
لنفوم	PCA	۰/۴۹	۰/۶۶	۰/۸۰	۰/۷۵
	$Clustvarsel$	۰/۴۹	۰/۵۵	۰/۷۲	۰/۷۵
	RP	۱	۱	۱	۱
پروستات	PCA	۰/۰۴	۰/۴۰	۰/۵۷	۰/۵۲
	$Clustvarsel$	۰/۱۱	۰/۴۱	۰/۵۸	۰/۵۵
	RP	۰/۲۶	۰/۴۶	۰/۶۳	۰/۶۳

برای نشان دادن برتری رویکرد $RPEClu$ براساس معیارهای معرفی شده، آن را با چند روش دیگر مقایسه کرده و نشان داده می‌شود که این روش برای خوشه‌بندی براساس چنین معیارهایی مناسب‌تر از روش‌های رقیب است. برای این منظور، از روش‌های $Clustvarsel$ و PCA استفاده شد. توجه شود که برای بهره‌برداری از روش PCA، مولفه‌های تبدیل یافته‌ای انتخاب می‌شوند که با هم حداقل ۹۰ درصد تغییرات بین داده‌ها را توصیف کنند. اجرای PCA نشان داد که در تحلیل داده‌های لوسمی۱۲، لنفوم و پروستات به ترتیب ۱۵، ۳۶ و ۸ مؤلفهٔ اصلی اول ما را به این هدف می‌رسانند. در ادامه نشان داده می‌شود که اگرچه روش PCA توانایی خوبی برای کاهش بُعد داده‌ها دارد اما ممکن است نتایج خوشه‌بندی مناسبی را به دنبال نداشته باشد. برای مقایسه روش PCA و $Clustvarsel$ با روش RP، خوشه‌بندی مدل‌مبنا براساس امتیازهای حاصل از اجرای این سه روش بر روی سه مجموعه داده انجام شد.

همان‌طورکه در جدول ۱ ملاحظه می‌شود هر دو روش رقیب دارای معیارهای ARI، JI، FMI و $Accuracy$ بسیار کمی در مقایسه با روش RP هستند و لذا عملکرد ضعیف‌تری دارند. باید اشاره شود که هدف از مطالعه انجام شده فقط مقایسه عملکرد خوشه‌بندی براساس سه رویکرد مورد اشاره و با لحاظ چند معیار متفاوت خوشه‌بندی بود و به جنبه‌های تحلیلی دیگر مرتبط با موضوع توجهی نشد. کنکاش بیشتر مطلب حول این روش‌ها در تحقیقات آتی دنبال خواهد شد.

بحث و نتیجه‌گیری

خوشه‌بندی داده‌های بُعد بالا با چالش‌های متفاوت آماری روبرو است و روش‌های متنوعی برای غلبه بر مشکلات مرتبط با آن وجود دارد. یکی از این رویکردهای مناسب، کاهش بُعد داده‌ها است که در این مقاله رویکرد تصویرهای تصادفی از هر دو جنبه نظری و کاربردی مورد بررسی قرار گرفت. همچنین عملکرد آن بر روی سه مجموعه داده واقعی

و مقایسه آن با روش‌های دیگر انجام و برتری آن بر اساس چند شاخص مرسوم ارزیابی خوشه‌بندی نشان داده شد. به منظور انجام تحقیقات آتی می‌توان با در نظر گرفتن روش‌های استنباطی مناسب به جنبه‌های احتمالاتی رویکرد تصویرهای تصادفی پرداخت. به علاوه، انتخاب مدل‌های آماری دیگر به جای مدل آمیخته گاوسی نیز می‌تواند مورد توجه قرار گیرد.

تقدیر و تشکر

نویسندگان از نکات ارزنده داوران، سردبیر و ویراستار محترم مجله که باعث بهبود هرچه بیشتر مقاله شده‌اند کمال تشکر و قدردانی را دارند.

مراجع

نبیل، م. و گل‌علی‌زاده، م. (۱۳۹۳)، نحوه خوشه‌بندی آماری داده‌های شکل، مجله علوم آماری، ۸، ۲۴۳-۲۲۳.

Anderlucci, L., Fortunato, F. and Montanari, A. (2021), High-dimensional Clustering via Random Projections, *Journal of Classification*, DOI: 10.1007/s00357-021-09403-7.

Anderlucci, L., Fortunato, F. and Montanari, A. (2019), *RPEClust: Random Projection Ensemble Clustering Algorithm*, R Package Version 0.1.0. Available at https://cran.microsoft.com/snapshot/2020-04-22/web/packages/RPEClust/index.html.

Bellman, R. (1957), *Dynamic Programming*, Princeton University Press, Los Angeles.

Bingham, E. and Mannila, H. (2001), Random Projection in Dimensionality Reduction: Applications to Image and Text Data, In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 245-250.

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

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.

Chung, D. and Keles, S. (2010), Sparse Partial Least Squares Classification for High-dimensional Data. *Statistical Applications in Genetics and Molecular Biology*, **9**, 1–30.

Cunningham, P. (2008), Dimension Reduction, In *Machine Learning Techniques for Multimedia*, Springer, Berlin, 91-112.

Dasgupta, S. (2000), Experiments with Random Projection, In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, 143–151.

Everitt, B. and Hothorn, T. (2011), *An Introduction to Applied Multivariate Analysis With R*, Springer, Washington.

Fern, X. Z. and Brodley, C. E. (2003), Random Projection for High-dimensional Data Clustering, A Cluster Ensemble Approach, In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, 186–193.

Fop, M. and Murphy, T. B. (2018), Variable Selection Methods for Model-based Clustering, *Statistics Surveys*, **12**, 18–65.

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.

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

Jolliffe, I.T. (1986), *Principal Component Analysis*, Springer, New York.

Maugis, C., Celeux, G. and Martin-Magniette, M. L. (2009), Variable Selection for Clustering with Gaussian Mixture Models, *Biometrics*, **65**, 701–709.

Pearson, K. (1901), LIII. On Lines and Planes of Closest Fit to Systems of Points in Space, *The London, Edinburgh, Dublin Philosophical Magazine and Journal of Science*, **2**, 559-572.

Raftery, A. E. and Dean, N. (2006), Variable Selection for Model-based Clustering, *Journal of the American Statistical Association*, **101**, 168-178.

Singh D, Febbo P, Ross K, Jackson D, Manola J, Ladd C, Tamayo P, Renshaw A, DAmico A, Richie J, Lander E, Loda M, Kantoff P, Golub T, and Sellers W (2002), Gene expression correlates of clinical prostate cancer behavior, *Cancer Cell*, **1**, 203–209.

Xu, D. and Tian, Y. (2015), A Comprehensive Survey of Clustering Algorithms, *Annals of Data Science*, **2**, 165–193.