Computational Intelligence in Electrical Engineering
Vol. 16, No. 4, 2026
pp. 29-42
Research Paper

Clustering with Apollonius kernel Based on Laplacian Matrix and k-means

Shahin Pourbahrami
[1] 1Assistant Professor, Department of Computer Engineering, National University of Skills, Tehran, Iran

Abstract:

In this study, a novel clustering algorithm based on the Apollonius kernel is presented. By utilizing the graph Laplacian matrix and the k-means algorithm, this method outperforms traditional approaches in separating clusters. In the first step, the data are normalized to standardize the feature scales. Then, by computing the Euclidean distance between points and applying the Apollonius kernel, the pairwise similarity matrix is constructed. Using the k-nearest neighbor graph, only the connections of the nearest neighbors are retained. Next, the graph Laplacian matrix is obtained as the difference between the degree matrix and the similarity matrix. This matrix represents the relationships between data points. By performing eigenvalue decomposition on the Laplacian matrix, the eigenvectors corresponding to the smallest eigenvalues are selected. The data are then projected onto a reduced space where the clusters are more clearly distinguishable. Then, the k-means algorithm is applied to the reduced data to determine the clusters. Finally, accuracy and other standard clustering validation metrics are used to evaluate the quality of the clustering. Experimental results show that this algorithm performs better than traditional methods, especially on datasets with nonlinear structures. Furthermore, the use of the Apollonius kernel in constructing the similarity graph improves the accuracy of identifying local structures in the data.

Keywords: Clustering, Apollonius kernel, Laplacian Matrix, k-Means, Similarity Matrix, Eigenvalues.

https://doi.org/10.22108/isee.2026.145741.1748

مقاله پژوهشی

خوشه‌بندی با هستهٔ آپولونیوس بر اساس ماتریس لاپلاسین و k میانگین

شهین پوربهرامی *

استادیار گروه مهندسی کامپیوتر، دانشگاه ملی مهارت، تهران، ایران

shpourbahrami@tvu.ac.ir

چکیده: در این پژوهش، یک الگوریتم خوشه‌بندی نوین مبتنی بر هستهٔ آپولونیوس ارائه شده¬ است که با بهره‌گیری از ماتریس لاپلاسین گراف و الگوریتم k میانگین، عملکردی بهتر در تفکیک خوشه¬ها نسبت به روش‌های سنتی دارد. در مرحلهٔ نخست، داده‌ها نرمال‌سازی می‌شوند تا مقیاس ویژگی‌ها یکسان شود. سپس، با محاسبهٔ فاصلهٔ اقلیدسی بین نقاط و اعمال هستهٔ آپولونیوس، ماتریس شباهت بین نمونه‌ها محاسبه می‌شود. با استفاده از گراف kنزدیک‌ترین همسایگان، فقط اتصالات همسایگان نزدیک حفظ می‌شوند. در ادامه، ماتریس لاپلاسین گراف از تفاضل ماتریس درجه و ماتریس شباهت به دست می‌آید که روابط بین نقاط داده را نمایان می‌کند. با انجام تجزیهٔ مقادیر ویژه بر روی ماتریس لاپلاسین، بردارهای ویژهٔ متناظر با کوچک‌ترین مقادیر ویژه انتخاب و داده‌ها به فضای کاهش‌یافته منتقل می‌شوند که در آن، خوشه‌ها واضح‌تر قابل تفکیک هستند. سپس، الگوریتم k میانگین بر روی داده‌های کاهش‌یافته اعمال می‌شود تا خوشه‌ها تعیین شوند. در نهایت، برای ارزیابی کیفیت خوشه‌بندی، از شاخص دقت و سایر شاخص¬های معتبر استفاده می‌شود. نتایج تجربی نشان می‌دهد این الگوریتم در مقایسه با روش‌های سنتی، به ویژه در داده‌های با ساختار غیرخطی، عملکردی بهتر دارد. همچنین، استفاده از هستهٔ آپولونیوس در ساخت گراف شباهت باعث افزایش دقت در شناسایی ساختارهای محلی داده‌ها می‌شود.

واژه‌های کلیدی: خوشه‌بندی، هستهٔ آپولونیوس، ماتریس لاپلاسین، k میانگین، ماتریس شباهت، مقادیر ویژه.

۱- مقدمه

در دهه‌های اخیر، الگوریتم‌های خوشه‌بندی طیفی[۱] به عنوان ابزارهای قدرتمند برای شناسایی ساختارهای پیچیده و غیرخطی در داده‌ها مورد توجه قرار گرفته‌اند. این الگوریتم‌ها با استفاده از ماتریس لاپلاسین گراف و تجزیهٔ مقادیر ویژه، داده‌ها را به فضای ویژگی‌های جدیدی منتقل می‌کنند که در آن، خوشه‌ها واضح‌تر قابل تفکیک هستند. با این حال، محاسبهٔ مقادیر ویژه برای ماتریس‌های بزرگ ممکن است از نظر محاسباتی پرهزینه باشد.

یک الگوریتم ساده برای خوشه‌بندی طیفی ارائه شده است که از بردارهای ویژهٔ ماتریس لاپلاسین نرمال‌شدهٔ گراف شباهت داده‌ها استفاده می‌کند [۱]. این الگوریتم با بهره‌گیری از نظریهٔ اختلال ماتریس، شرایطی را تحلیل می‌کند که تحت آن می‌تواند خوشه‌بندی مؤثری انجام دهد. الگوریتم خوشه‌بندی طیفی، با وجود توانایی در شناسایی ساختارهای پیچیده در داده‌ها، دارای معایبی است. یکی از مهم‌ترین چالش‌ها پیچیدگی محاسباتی زیاد است. همچنین، خوشه‌بندی طیفی ممکن است در شناسایی خوشه‌هایی با اندازه‌ها یا چگالی‌های متفاوت دچار مشکل شود و به حضور نویز و داده‌های پرت حساس باشد که ممکن است بر کیفیت نتایج تأثیر منفی بگذارد. در نهایت، تعیین تعداد مناسب خوشه‌ها برای خوشه‌بندی طیفی دشوار است و نیاز به دانش قبلی یا آزمایش‌های متعدد دارد.

الگوریتم k میانگین[۲] استاندارد توسط جیمز مک‌کوئین در مقاله‌ای با عنوان «برخی روش‌ها برای طبقه‌بندی و تحلیل مشاهدات چندمتغیره» معرفی شد که در سال ۱۹٦۷ در مجموعه مقالات پنجمین سمپوزیوم برکلی دربارهٔ آمار و احتمال ریاضی منتشر شد [۲]. در این روش، مک‌کوئین

نشانی نویسنده مسئول: ایران، تهران، دانشگاه ملی مهارت، گروه مهندسی کامپیوتر

[۱] تاریخ ارسال مقاله: ۱٤۰٤/۰٤/۰۸

تاریخ پذیرش مقاله: ۱٤۰۵/۰۲/۰۵

نام نویسنده مسئول: **شهین پوربهرامی**

الگوریتمی را برای تقسیم داده‌ها به k خوشه ارائه داد که هدف آن کمینه‌سازی واریانس درون‌خوشه‌ای است. الگوریتم با انتخاب تصادفی مراکز اولیهٔ خوشه‌ها آغاز می‌شود و سپس، به صورت تکراری، هر داده به نزدیک‌ترین مرکز خوشه اختصاص می‌یابد و مراکز خوشه‌ها بر اساس میانگین داده‌های اختصاص‌یافته به آن‌ها به‌روز می‌شوند. این فرایند تا زمانی ادامه می‌یابد که مراکز خوشه‌ها تغییر نکنند یا تغییرات آنها به حداقل برسد. مک‌کوئین در این مقاله ویژگی‌های آماری الگوریتم و کاربردهای آن در تحلیل داده‌های چندمتغیره را بررسی کرد.

خوشه‌بندی k میانگین هستهٔ پراکنده چالش‌های خوشه‌بندی داده‌های پیچیده و غیرخطی را بررسی می‌کند. روش‌های سنتی مانند k میانگین در شناسایی ساختارهای غیرخطی محدودیت دارند؛ بنابراین، تکنیک‌های جایگزین مانند k میانگین هسته و خوشه‌بندی طیفی توسعه یافته‌اند. با این حال، حضور متغیرهای نامربوط در داده‌ها می‌تواند عملکرد این الگوریتم‌ها را کاهش دهد [۳]. برای مقابله با این مشکل، روش‌های انتخاب متغیر مانند فیلتر، پوششی و جاسازی شده معرفی شده‌اند. الگوریتم k میانگین هستهٔ یک روش انتخاب متغیر جاسازی‌شده با استفاده از فضای ضرب تانسوری و هستهٔ تحلیل واریانس عمومی برای خوشه‌بندی غیرخطی است. این رویکرد با استفاده از فضای ضرب تانسوری، تعاملات پیچیدهٔ بین ویژگی‌ها را مدل‌سازی می‌کند و با اعمال هستهٔ تحلیل واریانس، ساختارهای غیرخطی در داده‌ها را بهتر شناسایی می‌کند [۳].

الگوریتم خوشه‌بندی بر اساس آمار هاپکینز و k میانگین، محدودیت‌های الگوریتم سنتی k میانگین در خوشه‌بندی داده‌هایی با ساختارهای نامنظم را بررسی می‌کند و با استفاده از آمار هوپکینز[۳] و مفهوم جفت نزدیک‌ترین همسایه طراحی شده است. در مرحلهٔ اول، داده‌ها به زیرمجموعه‌های کروی و نسبتاً یکنواخت تقسیم می‌شوند. سپس، با تعیین روابط مجاورت بین این زیرمجموعه‌ها، زیرمجموعه‌های مجاور و به‌هم‌پیوسته تا رسیدن به تعداد خوشه‌های مدنظر k ادغام می‌شوند. آمار هوپکینز در فرایند تقسیم‌بندی داده‌ها برای ارزیابی تمایل به خوشه‌بندی استفاده می‌شود، در حالی که مفهوم جفت نزدیک‌ترین همسایه در فرایند ادغام زیرمجموعه‌ها به کار می‌رود [٤].

در این راستا، الگوریتم پیشنهادی ما با بهره‌گیری از هستهٔ آپولونیوس برای ساخت گراف شباهت، ماتریس لاپلاسین را تشکیل و سپس با استفاده از تجزیهٔ طیفی و الگوریتم k میانگین، خوشه‌بندی داده‌ها را انجام می‌دهد. این رویکرد با تمرکز بر ساختارهای محلی داده‌ها و کاهش هزینه‌های محاسباتی، می‌تواند عملکردی بهتر در شناسایی خوشه‌های پیچیده و غیرخطی ارائه دهد.

ساختار این مقاله به شرح زیر سازمان‌دهی شده است. بخش ۲ مروری بر الگوریتم‌های خوشه‌بندی مبتنی بر چگالی و هندسه دارد و نقاط قوت و محدودیت‌های آنها را برجسته می‌کند. در بخش ۳، الگوریتم پیشنهادی را معرفی می‌کنیم که هستهٔ آپولونیوس را با ماتریس لاپلاسین و k میانگین برای خوشه‌بندی داده‌ها ادغام می‌کند. بخش ٤ نتایج تجربی را ارائه و مورد بحث قرار می‌دهد و توانایی بهبودیافتهٔ روش پیشنهادی را در شناسایی ساختارهای پیچیده و غیرخطی در مقایسه با تکنیک‌های خوشه‌بندی سنتی و پیشرفته نشان می‌دهد. در نهایت، بخش ۵ نتیجه‌گیری مقاله را نشان می‌دهد.

۲- کارهای مرتبط

الگوریتم سادهٔ چندفیلتری مبتنی بر k میانگین با استفاده از معیار هم‌ترازی هسته در محیط بدون نظارت، ترکیب فیلترها برای خوشه‌بندی را بهینه‌سازی می‌کند. با فرموله‌کردن مسئله به صورت بهینه‌سازی نرم، الگوریتم قادر به حل مؤثر آن با استفاده از روش نزول گرادیان کاهش‌یافته است [۵].

روش یادگیری ساختار متقابل برای خوشه‌بندی‌های چندگانه[٤] با یادگیری هم‌زمان ساختارهای زوجی و خوشه‌ای، ماتریس تقسیم‌بندی برای خوشه‌بندی را بهبود می‌دهد. با استفاده از گراف شباهت تبعیض‌آمیز در فضای هسته و هم‌ترازی با تقسیم‌بندی‌های پایه، ساختارهای مختلف به صورت یکپارچه یاد گرفته می‌شوند [٦]. الگوریتم ابتدا یک گراف شباهت تبعیض‌آمیز در فضای هسته می‌سازد که با

استفاده از هسته‌های پایهٔ مختلف، شباهت‌های بین نمونه‌ها را به صورت دقیق‌تر مدل می‌کند. سپس، با هم‌ترازی این گراف با تقسیم‌بندی‌های پایه، ساختارهای مختلف به صورت یکپارچه یاد گرفته می‌شوند. یکی از ویژگی‌های این الگوریتم استفاده از منظم‌سازی لاپلاسین[5] است که به حفظ ساختار زوجی در ماتریس تقسیم‌بندی کمک می‌کند و در عین حال، اطلاعات ساختار خوشه‌ای را در گراف شباهت تزریق می‌کند. این روش به ویژه در کاربردهایی مانند تحلیل داده‌های چندمنظوره، پردازش تصاویر و شناسایی الگوهای پیچیده در داده‌های بزرگ، پتانسیل زیادی دارد.

در روش جدید مبتنی بر تابع آپولونیوس و خوشه‌بندی قلهٔ چگالی، کرنل جدید برای ماشین بردار پشتیبان[6] معرفی شده است که بر پایهٔ تابع آپولونیوس است. این رویکرد با هدف بهبود عملکرد طبقه‌بندی در داده‌های پیچیده و غیرخطی ارائه شده است [۷]. در این رویکرد، تابع آپولونیوس برای مدل‌سازی دقیق‌تر روابط هندسی بین نقاط داده استفاده می‌شود، در حالی که خوشه‌بندی قله‌های چگالی به شناسایی ساختارهای محلی و نقاط مرکزی در داده‌ها کمک می‌کند. این ترکیب باعث می‌شود کرنل آپولونیوس به صورت تطبیقی با ساختارهای مختلف داده‌ها سازگار شود و عملکردی بهتر در طبقه‌بندی داده‌های پیچیده نسبت به کرنل‌های سنتی مانند کرنل گاوسی ارائه دهد. در روش پیشنهادی، برای خوشه‌بندی از این کرنل برای افزایش دقت استفاده شده است.

الگوریتم خوشه‌بندی طیفی بر اساس عملکرد کرنل گاوسی بهبودیافته و جست‌وجوی آنتن سوسک با ضریب میرایی با بهبود تابع کرنل گاوسی و استفاده از الگوریتم جست‌وجوی شاخک سوسک با عامل میرایی، مشکلات مربوط به پارامترهای مقیاس و پایداری نتایج در خوشه‌بندی طیفی را حل می‌کند [۸]. در این الگوریتم، برای ساخت ماتریس شباهت، از یک تابع کرنل گاوسی بهبودیافته بهره‌برداری می‌شود که با استفاده از اطلاعات فاصلهٔ برخی از نزدیک‌ترین همسایگان، به طور تطبیقی پارامتر مقیاس را انتخاب می‌کند. این رویکرد به حل مشکل انتخاب پارامتر مقیاس در کرنل گاوسی کمک می‌کند. در مرحلهٔ خوشه‌بندی، از الگوریتم جست‌وجوی شاخک سوسک با ضریب میرایی استفاده می‌شود تا بر مشکل ناپایداری نتایج خوشه‌بندی غلبه کند.

خوشه‌بندی طیفی به نام SC-NR[7] روشی برای بهبود ماتریس شباهت است که از ترکیب فاصلهٔ اقلیدسی وزن‌دار با تابع کرنل گاوسی استفاده می‌کند. وزن‌ها بر اساس ترتیب همسایگان نزدیک تنظیم می‌شوند تا روابط واقعی بین نقاط بهتر بازتاب یابند. با این حال، تابع کرنل گاوسی با استفاده از فاصلهٔ اقلیدسی و یک پارامتر ثابت سیگما، شباهت بین نقاط داده را اندازه‌گیری می‌کند. استفاده از یک سیگما ثابت می‌تواند به نتایج نادرست در خوشه‌بندی منجر شود، زیرا فاصلهٔ اقلیدسی به‌تنهایی قادر به بازتاب ساختار واقعی داده‌ها نیست. برای مثال، ممکن است نقاطی از خوشه‌های مختلف دارای فاصلهٔ اقلیدسی یکسانی باشند که در نتیجه، تابع کرنل گاوسی شباهتی یکسان برای آنها محاسبه می‌کند و این امر ممکن است به خوشه‌بندی نادرست منجر شود [۹].

در الگوریتم «K میانگین استاندارد هستهٔ پراکنده برای انتخاب ویژگی در خوشه‌بندی غیرخطی»[8] [۳]. نویسندگان چارچوبی برای خوشه‌بندی غیرخطی همراه با انتخاب متغیر پراکنده ارائه کرده‌اند که بر پایهٔ گسترش الگوریتم K میانگین استاندارد کرنل با استفاده از کرنل ANOVA بنا شده است. این روش با تجزیهٔ کرنل به مؤلفه‌های مربوط به اثرات اصلی و تعامل متغیرها و اعمال قیدهای ترکیبی بر ضرایب وزنی، امکان حذف ویژگی‌های بی‌اهمیت را فراهم می‌کند و هم‌زمان، ساختارهای غیرخطی پیچیده در داده را مدل می‌سازد. الگوریتم SKKM از بهینه‌سازی تناوبی برای به‌روزرسانی خوشه‌ها و وزن‌های ویژگی استفاده می‌کند. این رویکرد یکی از نخستین تلاش‌ها برای تلفیق یادگیری غیرخطی و انتخاب متغیر در چارچوبی یکپارچه است.

۳- روش پیشنهادی

مرحلهٔ اول: ساخت گراف شباهت با استفاده از هستهٔ آپولونیوس

در این مرحله، ابتدا داده‌ها را نرمال‌سازی می‌کنیم تا مقیاس

ویژگی‌ها یکسان شود. سپس، با ضرب ویژگی‌ها در وزن‌های مشخص‌شده، اهمیت هر ویژگی را تنظیم می‌کنیم. در ادامه، یک گراف k نزدیک‌ترین همسایگان با استفاده از فاصلهٔ اقلیدسی بین نقاط ساخته می‌شود. برای محاسبهٔ شباهت بین نقاط [۱۰، ۱۱]. از هستهٔ آپولونیوس که بر اساس ساخت دایرهٔ آپولونیوس است [۱۲، ۱۳] استفاده می‌کنیم که به صورت فرمول (۱) تعریف می‌شود:

$$k(x,y)=\frac{1}{d(x,y)+\varepsilon} \qquad (۱)$$

که در آن، اپسیلون یک مقدار ثابت کوچک است که از تقسیم بر صفر جلوگیری می‌کند (پارامتر اپسیلون نقش فیلتر نرم‌کننده[9] را دارد تا از تقسیم بر صفر جلوگیری و حساسیت شباهت به فواصل خیلی کوچک را کنترل کند). شباهت محاسبه‌شده با ماتریس k نزدیک‌ترین همسایگان ضرب می‌شود تا فقط اتصالات همسایگان نزدیک حفظ شوند.

مرحلهٔ دوم: ساخت ماتریس لاپلاسین گراف

در این مرحله، ماتریس درجه را محاسبه می‌کنیم که هر قطر آن نشان‌دهندهٔ مجموع وزن‌های یال‌های متصل به گره مربوط است. سپس، ماتریس لاپلاسین گراف را با استفاده از فرمول (۲) محاسبه می‌کنیم:

$$L=W-D \qquad (۲)$$

که در آن، D ماتریس درجه و W ماتریس شباهت است. این ماتریس نمایانگر ساختار گراف و روابط بین نقاط داده است.

مرحلهٔ سوم: تجزیهٔ طیفی و کاهش ابعاد

در این مرحله، تجزیهٔ مقادیر ویژهٔ ماتریس لاپلاسین انجام می‌شود تا مقادیر ویژه و بردارهای ویژهٔ آن به دست آیند. سپس، بردارهای ویژهٔ متناظر با کوچک‌ترین مقادیر ویژه را انتخاب می‌کنیم و آنها را به صورت ماتریسی با ابعاد $n\times k$ (که در آن k تعداد خوشه‌ها و n تعداد نمونه‌های ذخیره‌شده است) نشان می‌دهیم. این ماتریس ویژگی‌هایی جدید را برای داده‌ها فراهم می‌کند که در فضای کاهش‌یافته قرار دارند و خوشه‌ها در آن واضح‌تر قابل تفکیک هستند.

مرحلهٔ چهارم: خوشه‌بندی با استفاده از k میانگین

در این مرحله، الگوریتم k میانگین را بر روی ماتریس ویژگی‌های به‌دست‌آمده از مرحلهٔ قبلی اعمال می‌کنیم. در این الگوریتم، ابتدا مراکز اولیهٔ خوشه‌ها به صورت تصادفی انتخاب می‌شوند. سپس، هر داده به نزدیک‌ترین مرکز خوشه اختصاص می‌یابد و مراکز خوشه‌ها بر اساس میانگین داده‌های اختصاص‌یافته به آنها به‌روز می‌شوند. این فرایند تا زمانی ادامه می‌یابد که مراکز خوشه‌ها تغییر نکنند یا تغییرات آنها به حداقل برسد [۲].

۴- نتایج

ارزیابی عملکرد روش پیشنهادی

برای ارزیابی کیفیت خوشه‌بندی از شاخص‌هایی مختلف همچون ARI، Accuracy (دقت) و NMI استفاده می‌کنیم. شاخص ARI میزان شباهت بین خوشه‌بندی به‌دست‌آمده و برچسب‌های واقعی داده‌ها را اندازه‌گیری می‌کند. مقدار ARI بین ۱- و ۱ متغیر است که مقدار ۱ نشان‌دهندهٔ تطابق کامل بین خوشه‌بندی و برچسب‌های واقعی است. معیار اطلاعات متقابل نرمال‌شده NMI بر پایهٔ نظریهٔ اطلاعات است و میزان اشتراک اطلاعات بین برچسب‌های واقعی و برچسب‌های پیش‌بینی‌شده توسط الگوریتم خوشه‌بندی را اندازه‌گیری می‌کند. NMI بین ۰ و ۱ مقدار می‌گیرد؛ مقدار ۱ نشان‌دهندهٔ تطابق کامل بین خوشه‌بندی و برچسب‌های واقعی است، در حالی که مقدار ۰ نشان‌دهندهٔ عدم تطابق کامل است. یکی از ویژگی‌های مهم NMI این است که مستقل از تعداد خوشه‌ها و برچسب‌ها عمل می‌کند و به همین دلیل، برای مقایسهٔ الگوریتم‌های مختلف خوشه‌بندی مناسب است.

معیار دقت[10] خوشه‌بندی بر اساس فرمول (۳)، نسبت تعداد نمونه‌هایی را که به‌درستی خوشه‌بندی شده‌اند به کل نمونه‌ها محاسبه می‌کند (برای محاسبهٔ معنادارِ دقت در مسائل خوشه‌بندی لازم است ابتدا یک نگاشت یک‌به‌یک بین

برچسب‌های خوشه‌ای و برچسب‌های واقعی کلاس پیدا شود). از آنجا که در خوشه‌بندی، برچسب‌های پیش‌بینی‌شده ممکن است با برچسب‌های واقعی از نظر نام‌گذاری متفاوت باشند، قبل از محاسبهٔ دقت، معمولاً یک نگاشت بین برچسب‌های پیش‌بینی‌شده و واقعی انجام می‌شود تا بیشترین تطابق حاصل شود. دقت خوشه‌بندی نیز بین ۰ و ۱ مقدار می‌گیرد؛ مقدار ۱ نشان‌دهندهٔ تطابق کامل است.

بر اساس فرمول (۳)، y_i نشان‌دهندهٔ برچسب واقعی کلاس است و c_i شمارهٔ برچسب خوشه‌ای است که از فرایند خوشه‌بندی به دست آمده است. تابع $\delta(x,y)=\begin{cases}1 & x=y\\ 0 & x\neq y\end{cases}$ یک تابع شاخص[11] است که در صورت برابری دو برچسب، مقدار ۱ و در غیر این صورت، مقدار ۰ می‌گیرد. بهترین نگاشت بین برچسب‌های خوشه و برچسب‌های واقعی کلاس با استفاده از الگوریتم Hungarian یا Munkres تعیین می‌شود تا بیشترین تطابق ممکن بین خوشه‌ها و کلاس‌های واقعی حاصل شود.

$$ACC=\frac{1}{n}\sum_{i=1}^{n}\delta\left(y_i,map\left(c_i\right)\right) \qquad (۳)$$

استفادهٔ هم‌زمان از NMI و دقت می‌تواند دیدی جامع‌تر از عملکرد الگوریتم خوشه‌بندی ارائه دهد، زیرا هر یک از این معیارها جنبه‌هایی متفاوت از کیفیت خوشه‌بندی را بررسی می‌کنند.

ارزیابی عملکرد الگوریتم پیشنهادی و سایر الگوریتم‌ها با مجموعه‌داده‌های واقعی

مجموعه‌داده‌های واقعی در جدول (۱)[12]، مانند Iris، Wine، Digits، Breast، Cancer، Diabetes و Olivetti Faces، به طور گسترده برای ارزیابی و آموزش الگوریتم‌های مختلف استفاده می‌شوند. مجموعه دادهٔ Iris شامل ۱۵۰ نمونه از سه گونهٔ گل زنبق است که هر نمونه دارای ٤ ویژگی است. مجموعه دادهٔ Wine شامل ۱۷۸ نمونه از سه نوع نوشیدنی با ۱۳ ویژگی شیمیایی است. مجموعه دادهٔ Digits شامل ۱۷۹۷ تصویر ۸×۸ از ارقام دست‌نویس ۰ تا ۹ است که هر تصویر به یک بردار ٦٤بُعدی تبدیل شده است. مجموعه دادهٔ Cancer شامل ٥٦۹ نمونه با ۳۰ ویژگی عددی برای تشخیص سرطان سینه است. مجموعه دادهٔ Diabetes شامل ۷٦۸ نمونه با ۸ ویژگی برای پیش‌بینی ابتلا به دیابت است. مجموعه دادهٔ Olivetti Faces شامل ٤۰۰ تصویر چهره از ٤۰ فرد مختلف است که هر تصویر به صورت یک بردار ٤۰۹٦بُعدی (٦٤×٦٤ پیکسل) نمایش داده می‌شود. این مجموعه‌داده‌ها به عنوان معیارهای استاندارد در پژوهش‌های علمی شناخته می‌شوند و برای مسائل مختلف مانند طبقه‌بندی، خوشه‌بندی و کاهش ابعاد داده‌ها کاربرد دارند.

جدول (۱): مجموعه‌داده‌های واقعی

مجموعه‌داده	تعداد خوشه‌ها	تعداد ابعاد	تعداد داده‌ها
Iris	۳	٤	۱٥۰
Wine	۳	۱۳	۱۷۸
Digits	۱۰	٦٤	۱۷۹۷
Breast Cancer (Wisconsin Breast Cancer Diagnostic (WBCD))	۲	۳۰	٥٦۹
Diabetes	۲	۸	۷٦۸
Olivetti Faces	٤۰	٤۰۹٦	٤۰۰

جدول (۲): ارزیابی الگوریتم پیشنهادی در مجموعه‌داده‌های واقعی

مجموعه‌داده	NMI الگوریتم پیشنهادی	دقت الگوریتم پیشنهادی	پارامتر k	نوع لاپلاسیَن	نکتهٔ کلیدی عملکرد
Iris	۰/٤٦۰	۰/٦۸۷	۱۰	L غیرنرمال‌شده	شکاف طیفی بین λ_3 و λ_4 نشانگر سه خوشه است.
Wine	۰/۸۵۳	۰/۹۵۵	۱۲	L_symنرمال‌شده متقارن	حساس به انتخاب k، عملکرد پایدار با k برابر ۱۲ است.
Digits	۰/۸۹۲	۰/۸۲۹	۱۵	L_rw نرمال‌شده تصادفی	تعبیه با d=k عملکردی بهتر از شکاف طیفی دارد.
Breast Cancer	۰/٦٤۵	۰/۹۳۵	۱۰	L غیرنرمال‌شده	خوشه‌بندی پایدار، اما حساس به نویز.
Breast Cancer (PCA)	۰/٦۵۸	۰/۹۳۸	۱۰	L غیرنرمال‌شده	نتیجه بعد از اعمال کاهش ابعاد با PCA.
Diabetes	۰/۷۹۳	۰/۳۷۹	۸	L_sym	داده با هم‌پوشانی زیاد، دقت کمتر ولی NMI زیاد.
Olivetti Faces	۰/۷۹۵	۰/۵۵۲	۱۵	L_rw	افزایش k باعث بهبود پایداری تعبیه شد.
Olivetti Faces (PCA)	۰/۸۰۹	۰/٦۱۸	۱۵	L_rw	نتیجه بعد از اعمال کاهش ابعاد با PCA
SD ± میانگین عملکرد الگوریتم پیشنهادی	۰/۲۲±۰/۷۳۹	۰/۲۱٤±۰/۷۳۳	—	—	—

جدول (۳): ارزیابی الگوریتم پیشنهادی، الگوریتم SKKM و k میانگین استاندارد در مجموعه‌داده‌های واقعی

مجموعه‌داده‌ها	NMI	دقت	ARI	ARI
	استاندارد میانگین k	استاندارد میانگین k	الگوریتم پیشنهادی	SKKM الگوریتم
Iris	۰/۷٤	۰/۷۸	۰/۷۰	۰/۷٤
Wine	۰/۳۱	۰/٦٤	۰/۸۷	۰/۸٦
Digits	۰/۵۰	۰/۳۷	۰/۸۳	۰/۸۱
Breast Cancer	۰/۷۳	۰/۸۳	۰/۷۵	۰/۷۲
Diabetes	۰/۰۰۵	۰/٦۵	-	-
Olivetti Faces	۰/۵۵	۰/۵٤	-	-
SD ± میانگین	۰/٤۷۲±۰/۲۵۵	۰/٦۳۵±۰/۱۵۸	۰/۷۸۸±۰/۰۷۱	۰/۷۸۳±۰/۰٦۱

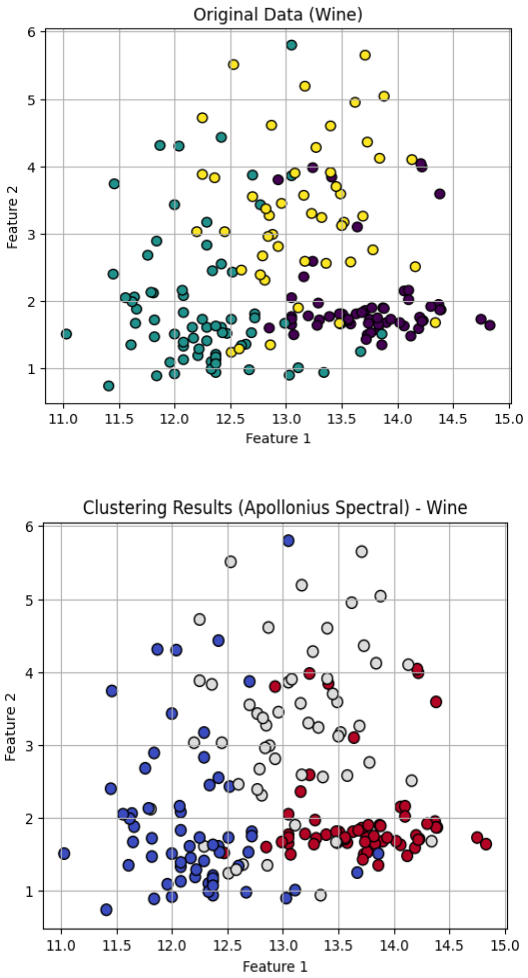
جدول (۲) نتایج عملکرد الگوریتم پیشنهادی مبتنی بر گراف آپولونیوسی را در چند مجموعه دادهٔ واقعی به همراه پارامترهای بهینه و نوع لاپلاسیَن به‌کاررفته نشان می‌دهد. همان‌گونه که در ستون‌های NMI و دقت مشاهده می‌شود، الگوریتم پیشنهادی در بیشتر مجموعه‌داده‌ها نتایجی پایدار داشته و در مقایسه با روش‌های کلاسیک، توانسته است ساختارهای درونی داده‌ها را به شکلی دقیق‌تر شناسایی کند. در مجموعه‌دادهٔ Iris، مقدار NMI برابر ۰/٤٦۰ و دقت برابر ۰/٦۸۷به دست آمده است که با استفاده از لاپلاسیَن غیرنرمال‌شده (L) حاصل شده است. بررسی طیف ویژهٔ لاپلاسیَن نشان داد شکاف قابل ملاحظه‌ای میان λ_3 و λ_4 وجود دارد که وجود سه خوشهٔ متمایز را در این داده تأیید می‌کند.

این یافته نشان می‌دهد الگوریتم پیشنهادی قادر است ساختار خوشه‌ای داده‌های کم‌بعد و نسبتاً منظم را به‌خوبی بازیابی کند. در مجموعه‌دادهٔ Wine، نمایش مقادیر حاصل NMI=۰/۸۵۳ و دقت ۰/۹۵۵ نشان‌دهندهٔ عملکرد بسیار مطلوب الگوریتم در داده‌های با ویژگی‌های پیوسته و خوشه‌های هم‌پوشان است. استفاده از لاپلاسیَن نرمال‌شدهٔ متقارن موجب پایداری نتایج شده است، زیرا این نوع لاپلاسیَن توانسته اثر تفاوت در اندازهٔ خوشه‌ها را تعدیل کند. در مقابل، در مجموعه‌دادهٔ Digits که دارای ساختار غیرخطی و چندخوشه‌ای است، الگوریتم با انتخاب لاپلاسیَن نرمال‌شدهٔ تصادفی مقدار NMI=۰/۸۹۲ و دقت ۰/۸۹۲ را به دست آورده است.

در مجموعه‌دادهٔ Breast Cancer، الگوریتم پیشنهادی با NMI=۰/٦٤٥ و دقت ۰/۹۳۵ توانسته است عملکردی پایدار ارائه دهد، هرچند به نویز حساس است. این داده نشان می‌دهد در حضور ویژگی‌های همبسته و نویزی، استفاده از لاپلاسیَن غیرنرمال‌شده می‌تواند خوشه‌بندی معناداری ایجاد کند. در مجموعه‌دادهٔ Diabetes، مقدار NMI زیاد (۰/۷۹۳) در کنار دقت کمتر (۰/۳۷۹) مشاهده می‌شود که نشان می‌دهد الگوریتم توانسته است ساختارهای نسبی خوشه‌ها را تشخیص دهد، اما به دلیل هم‌پوشانی زیاد بین داده‌ها، برچسب‌های نهایی دقت کمتری دارند. در اینجا، استفاده از لاپلاسیَن متقارن باعث بهبود پایداری نسبت به حالت غیرنرمال‌شده شده است. در نهایت، در مجموعه‌دادهٔ Olivetti Faces که از تصاویر چهره تشکیل شده، مقادیر NMI=۰/۷۹۳ و دقت ۰/۵۵۲ نشان‌دهندهٔ توانایی الگوریتم در بازنمایی روابط چهره‌هاست. استفاده از لاپلاسیَن تصادفی و افزایش k تا ۱۵ موجب بهبود پایداری تعبیه شده است.

به طور کلی، میانگین نتایج در تمام مجموعه‌داده‌ها با NMI برابر ۰/۷۳۹±۰/۲۲ و دقت برابر ۰/۷۳۳±۰/۲۱٤ نشان می‌دهد الگوریتم پیشنهادی از نظر میانگین عملکرد و پایداری دارد. تحلیل پارامترها مشخص می‌کند بازهٔ بهینه برای k حدود ۱۰ تا ۱۵ است. همچنین، نوع لاپلاسیَن باید با ساختار داده تطبیق یابد که برای داده‌های هم‌تراز و L ساده مناسب است. به طور کلی، این نتایج اثبات می‌کنند ترکیب کرنل آپولونیوس با روش تعبیهٔ طیفی می‌تولند یک چارچوب انعطاف‌پذیر و دقیق برای خوشه‌بندی داده‌های پیچیده فراهم کند.

جدول (۳) عملکرد الگوریتم پیشنهادی را در مقایسه با روش‌های پایه شامل k میانگین استاندارد بر اساس معیارهای NMI و دقت و الگوریتم SKKM بر اساس ARI روی مجموعه‌داده‌های مختلف نشان می‌دهد. نتایج NMI و دقت k میانگین استاندارد با میانگین‌های نسبتاً کم (به ترتیب ۰/٤۷۲ و ۰/٦۳۵) و انحراف معیارهای زیاد بیانگر این است که k میانگین استاندارد در برابر ساختارهای پیچیده و داده‌های با توزیع غیرخطی کارایی محدودی دارد. در مقابل، مقدار ARI الگوریتم پیشنهادی در چهار مجموعه‌دادهٔ مشترک از جمله Iris، Wine، Digits و Breast Cancer میانگین ۰/۷۸۸±۰/۰۷۱ را نشان می‌دهد که بیانگر عملکرد باثبات و دقیق در بازیابی ساختار واقعی خوشه‌هاست. این مقدار اندکی بیشتر از الگوریتم SKKM با میانگین ۰/۷۸۳±۰/۰٦۱ است که نشان می‌دهد روش پیشنهادی نه فقط در بیشتر مجموعه‌داده‌ها عملکرد مشابه یا بهتر نسبت به SKKM دارد، بلکه توانایی استخراج ویژگی‌های ساختاری عمیق‌تری از داده‌ها را نیز ارائه می‌دهد. همچنین، نبود مقدار ARI برای مجموعه‌داده‌های Diabetes و Olivetti Faces به دلیل ابعاد یا ویژگی‌های خاص آنهاست که در این آزمایش‌ها لحاظ نشده‌اند. در مجموع، جدول (۳) نشان می‌دهد الگوریتم پیشنهادی از نظر پایداری، دقت خوشه‌بندی و تولنایی تفکیک ساختارهای درهم‌تنیده، نسبت به روش‌های پایه و حتی SKKM برتری نسبی دارد.



شکل (۱): دادهٔ واقعی Wine در دو خروجی اصلی و خروجی با الگوریتم پیشنهادی

در شکل (۱)، نتایج اجرای الگوریتم بر روی داده‌ی واقعی Wine نشان داده شده است. در این شکل، توزیع اصلی داده در فضای دو ویژگی نخست نمایش یافته است و نمونه‌ها بر اساس برچسب‌های واقعی رنگ‌آمیزی شده‌اند. همچنین، خروجی خوشه‌بندی مبتنی بر گراف آپولونیوس و تعبیهٔ طیفی ارائه شده است. همان‌گونه که مشاهده می‌شود، ساختار غیرخطی داده در فضای طیفی به صورت واضح‌تر نمایان شده است و خوشه‌ها تفکیک‌پذیری بیشتری یافته‌اند که بیانگر توانایی مدل در استخراج ساختار درونی داده و بهبود جداسازی نسبت به فضای ویژگی اولیه است.

جدول (٤): ارزیابی خوشه‌بندی طیفی در مجموعه‌داده‌های واقعی

مجموعه‌داده‌ها	NMI (الگوریتم خوشه‌بندی طیفی)	دقت (الگوریتم خوشه‌بندی طیفی)
Iris	۰/۷۶	۰/۹۰
Wine	۰/٤٦	۰/۷۳
Digits	۰/۵۵	۰/۷۰
Breast Cancer	۰/۷۵	۰/۹٦
Diabetes	۰/۰۲	۰/٦۸
Olivetti Faces	۰/۵۸	۰/٦۰
میانگین ± SD	۰/۲٦٦±۰/۵۱۰	۰/۱۳۵ ±۰/۷٦۲

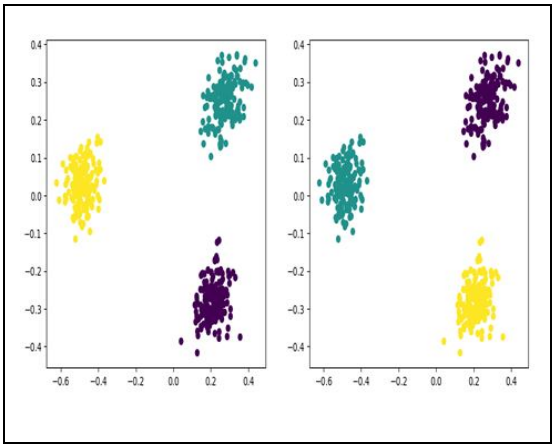
مقایسهٔ عملکرد الگوریتم پیشنهادی با الگوریتم خوشه‌بندی طیفی بر روی مجموعه‌داده‌های مختلف در جدول (۲) و (٤) نشان می‌دهد الگوریتم پیشنهادی در برخی از مجموعه‌داده‌ها عملکردی بهتر دارد. برای مثال، در مجموعه‌داده‌های Wine و Digits، الگوریتم پیشنهادی با NMI برابر ۰/۸۵۳ و ۰/۸۹۲ و دقت‌های برابر ۰/۹۵۵ و ۰/۸۹۲، نسبت به خوشه‌بندی طیفی که NMI ۰/٤۰ و ۰/۵۵ و دقت‌های ۰/۷۳ و ۰/۷۰ دارد، عملکردی بهتر نشان داده است. این بهبود می‌تواند به دلیل استفاده از تکنیک‌های پیشرفته‌تر در الگوریتم پیشنهادی باشد که ساختارهایی پیچیده‌تر را در داده‌ها شناسایی می‌کند.

از سوی دیگر، در مجموعه‌داده‌هایی مانند Iris، Breast و Cancer، الگوریتم خوشه‌بندی طیفی عملکردی بهتر نسبت به الگوریتم پیشنهادی دارد. برای مثال، در مجموعه‌دادهٔ Iris، خوشه‌بندی طیفی با NMI=۰/۷٦ و دقت ۰/۹۰، نسبت به الگوریتم پیشنهادی با NMI=۰/٤٦۰ و دقت ۰/٦۸۷، نتایجی بهتر ارائه داده است. این به دلیل سادگی ساختار مجموعه داده‌ها است که الگوریتم‌های مبتنی بر طیف قادر به شناسایی خوشه‌ها با دقت بیشتری باشند.

ارزیابی عملکرد الگوریتم پیشنهادی و سایر الگوریتم‌ها با مجموعه‌داده‌های مصنوعی

دیتاست Blobs مجموعه‌ای از داده‌هاست که به صورت خوشه‌های گوسی در فضای ویژگی‌ها توزیع شده‌اند. ابعاد قابل تنظیم به طور پیش‌فرض ۲بُعدی هستند. تعداد نمونه‌ها به طور پیش‌فرض ۱۰۰ نمونه است. تعداد خوشه‌ها به طور

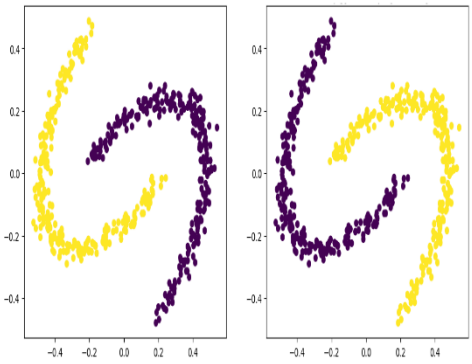
پیش‌فرض ۳ خوشه است.



شکل (۲): دادهٔ مصنوعی Blobs

دیتاست Moons شامل دو نیم‌دایرهٔ درهم‌تنیده است که به شکل دو هلال ماه در فضای دوبُعدی قرار دارند. این ساختار غیرخطی برای آزمایش الگوریتم‌هایی مناسب است که توانایی شناسایی مرزهای پیچیده را دارند. این داده دوبُعدی است. تعداد نمونه‌ها به طور پیش‌فرض ۱۰۰ نمونه است. تعداد خوشه‌ها ۲ خوشه است. نقاط داده در دو نیم‌دایرهٔ درهم‌تنیده با امکان افزودن نویز گوسی توزیع می‌شوند.

شکل‌های (۲) و (۳) اثربخشی الگوریتم خوشه‌بندی پیشنهادی را بیشتر نشان می‌دهند و عملکرد خوشه‌بندی بی‌نقصی را در دو مجموعه‌دادهٔ مصنوعی متنوع Blobs و Moons نشان می‌دهند. همهٔ معیارهای ارزیابی به حداکثر مقادیر خود رسیده‌اند که این امر نشان می‌دهد الگوریتم با موفقیت ساختارهای خوشه‌بندی اساسی را حتی در توزیع‌های دادهٔ پیچیده و غیرمحدب بدون خطا شناسایی می‌کند. این امر سازگاری الگوریتم پشنهادی را در مدیریت انواع اشکال داده برجسته می‌کند.



شکل (۳): دادهٔ مصنوعی Moons

جدول (۵): ARI در مجموعه‌داده‌های مصنوعی

مجموعه‌داده‌ها	k میانگین استاندارد	الگوریتم خوشه‌بندی طیفی	روش پیشنهادی
R15	۰/۵۷	۰/٦٤	۰/۸۸
Blobs	۱	۱	۰/٦۷٦
Moons	۱	۱	۱
± SD میانگین	۰/۲۰۳±۰/۸۵۷	۰/۸۸۰ ۰/۱٦٦±	۰/۱٦۷±۰/۸۵٤

جدول (٦): NMI در مجموعه‌داده‌های مصنوعی

مجموعه‌داده‌ها	k میانگین استاندارد	الگوریتم خوشه‌بندی طیفی	روش پیشنهادی
R15	۰/۵۷	۰/۷۱	۰/۹٤
Blobs	۱	۱	۰/۷
Moons	۰/۳۸	۰/۹٦	۱
± SD میانگین	۰/۳۱۸±۰/٦۵۰	۰/۱۵۵±۰/۸۹۰	۰/۱۲۲±۰/۹۰

جدول (۷): دقت در مجموعه‌داده‌های مصنوعی

مجموعه‌داده‌ها	k میانگین استاندارد	الگوریتم خوشه‌بندی طیفی	روش پیشنهادی
R15	۰/۸۷	۰/۹۱	۰/۹۲۸
Blobs	۱	۱	۰/۸٦۰
Moons	۰/۳۸	۰/۹۷	۱
± SD میانگین	± ۰/۷۵۰ ۰/۲۹٦	۰/۰٤۵±۰/۹٦۰	۰/۰۷۱±۰/۹۲۹

برای مقایسه عملکرد الگوریتم‌های خوشه‌بندی بر روی مجموعه‌داده‌های مصنوعی R15، Blobs وMoons در جدول‌های (۵)، (٦) و (۷)، سه معیار ارزیابی NMI،ARI و دقت بررسی شده‌اند. نتایج نشان می‌دهد الگوریتم پیشنهادی، الگوریتم خوشه‌بندی طیفی و الگوریتم k میانگین استاندارد در هر یک از این مجموعه‌داده‌ها عملکردی متفاوت دارند.

در مجموعه‌دادهٔ R15، الگوریتم پیشنهادی باARI برابر ۰/۸۸، NMI برابر ۰/۹٤ و دقت برابر ۰/۹۲، عملکردی بهتر نسبت به الگوریتم‌های دیگر داشته است. الگوریتم خوشه‌بندی طیفی با ARI برابر ۰/٦٤، NMI برابر ۰/۷۱ و دقت برابر ۰/۹۱ در رتبهٔ دوم قرار دارد، در حالی که الگوریتم k میانگین استاندارد با ARI برابر ۰/۵۷، NMI برابر ۰/۵۷ و دقت برابر ۰/۸۷ کمترین عملکرد را نشان داده است. این نتایج نشان می‌دهد الگوریتم پیشنهادی در شناسایی خوشه‌های پیچیده‌تر موفق‌تر عمل کرده است. مجموعه‌دادهٔ Blobs در این مجموعه‌داده، الگوریتم خوشه‌بندی طیفی و k میانگین استاندارد هر دو با ARI وNMI برابر ۱ و دقت برابر ۱، عملکردی بسیار خوب داشته‌اند. الگوریتم پیشنهادی با ARI برابر ۰/٦۷٦،NMI برابر ۰/۷۵۹ و دقت برابر ۰/۸٦۰، عملکردی کمتر نسبت به دو الگوریتم دیگر داشته است. این نتایج نشان می‌دهد در داده‌هایی با ساختار ساده‌تر، الگوریتم‌های سنتی‌تر مانند k میانگین می‌توانند عملکردی مناسب داشته باشند. مجموعه‌دادهٔ Moons در این مجموعه‌داده، الگوریتم پیشنهادی و الگوریتم خوشه‌بندی طیفی هر دو با ARI وNMI برابر ۱ و دقت برابر ۱، عملکردی بسیار خوب داشته‌اند. الگوریتم kمیانگین استاندارد با ARI برابر ۰/۳۸،NMI برابر ۰/۷۸ و دقت برابر ۰/۷۸، عملکردی با دقت کمتر نسبت به دو الگوریتم دیگر داشته است. این نتایج نشان می‌دهد در داده‌هایی با ساختار غیرخطی و پیچیده‌تر، الگوریتم‌های پیشرفته‌تر مانند خوشه‌بندی طیفی می‌توانند عملکردی بهتر داشته باشند.

در مجموع، این مقایسه‌ها نشان می‌دهند انتخاب الگوریتم خوشه‌بندی مناسب به ساختار داده‌ها بستگی دارد. در داده‌هایی با ساختار ساده‌تر، الگوریتم‌های سنتی‌تر مانند k میانگین می‌توانند عملکردی مناسب داشته باشند، در حالی که در داده‌هایی با ساختار پیچیده‌تر، الگوریتم‌های پیشرفته‌تر مانند خوشه‌بندی طیفی و الگوریتم پیشنهادی می‌توانند عملکردی بهتر ارائه دهند.

تحلیل آماری کیفی (Wilcoxon & Friedman)

نتایج در جدول (۸) نشان می‌دهد الگوریتم پیشنهادی در مقایسه با دو روش دیگر، یعنی خوشه‌بندی طیفی و K میانگین استاندارد، عملکردی بسیار بهتر از خود نشان داده است. در بخش میانگین NMI، مقدار به‌دست‌آمده برای روش پیشنهادی بیشتر از دو روش دیگر است که نشان می‌دهد الگوریتم پیشنهادی توانسته است ساختارهای درونی داده‌ها را بهتر شناسایی کند و خوشه‌هایی با شباهت درونی بیشتر ایجاد کند. همچنین، مقدار انحراف معیار کمتر بیانگر پایداری بهتر این روش در میان تکرارهای مختلف و بر روی مجموعه‌داده‌های گوناگون است. این نتایج نشان می‌دهد روش پیشنهادی علاوه بر دقت بیشتر، از نظر پایداری و سازگاری نیز برتر است.

در ادامه، آزمون Signed-Rank Wilcoxon برای مقایسهٔ زوجی الگوریتم‌ها انجام شده است. بر اساس مقادیر p، اختلاف میان روش پیشنهادی با هر دو الگوریتم دیگر در سطح اطمینان ۹۵ درصد معنادار است؛ به طوری که مقدار p بین روش پیشنهادی و خوشه‌بندی طیفی برابر ۰/۰۳۱ و بین روش پیشنهادی و K میانگین استاندارد برابر ۰/۰۱٦ بوده است. این بدان معناست که روش پیشنهادی به طور آماری در شاخص NMIعملکردی بهتر دارد. تنها تفاوت بین خوشه‌بندی طیفی وK *میانگین* از نظر آماری معنادار نیست که نشان می‌دهد عملکرد این دو روش در سطحی مشابه قرار دارد.

نتایج آزمون Friedman نیز این یافته‌ها را تأیید می‌کند. مقدار آمارهٔ χ^2 برابر ۱۰/۳۳ است و همچنین p برابر ۰/۰۰۵۷است که نشان‌دهندهٔ وجود تفاوت معنادار در عملکرد سه الگوریتم است. میانگین رتبه‌های فریدمن نشان می‌دهد الگوریتم پیشنهادی با رتبهٔ ۱/۱۷ بهترین عملکرد را دارد، در

حالی که خوشه‌بندی طیفی و K میانگین استاندارد به ترتیب با رتبه‌های ۲/۰۰ و ۲/۸۳ در جایگاه دوم و سوم قرار گرفته‌اند. در آزمون Nemenyi Post-hoc نیز تفاوت بین روش پیشنهادی و K میانگین استاندارد از نظر آماری معنادار است ، اما اختلاف بین روش پیشنهادی و خوشه‌بندی طیفی معنادار نیست. بنابراین، در سطح اطمینان ۹۵ درصد، روش پیشنهادی نسبت به K میانگین استاندارد برتری آشکار دارد و در مقایسه با خوشه‌بندی طیفی نیز عملکردی بهتر، ولی نه به صورت معنادار از نظر آماری نشان می‌دهد. در مجموع، می‌توان نتیجه گرفت الگوریتم پیشنهادی پایدارترین و دقیق‌ترین روش خوشه‌بندی در این آزمایش‌ها بوده است.

جدول (۸): تحلیل آماری الگوریتم‌ها بر اساس شاخص NMI

نتیجهٔ آماری/تفسیر	K میانگین استاندارد	خوشه‌بندی طیفی	الگوریتم پیشنهادی	معیار	شاخص/آزمون
روش پیشنهادی بیشترین مقدار میانگین را دارد	۰/۲٦±۰/٤۷	۰/۲۷±۰/۵۱	۰/۲۹±۰/٦۲	میانگین ± انحراف معیار	میانگین NMI واقعی
p برابر ۰/۰۳۱→ اختلاف معنادار به نفع الگوریتم پیشنهادی	—	—	—	p(Proposed vs Spectral)	Wilcoxon Signed-Rank Test
p برابر ۰/۰۱٦→ اختلاف معنادار به نفع الگوریتم پیشنهادی	—	—	—	p(Proposed vs K-Means)	
p برابر ۰/۳۱۳→ اختلاف معنادار نیست	—	—	—	p(Spectral vs K-Means)	
تفاوت معنادار بین الگوریتم‌ها→ p کمتر از ۰/۰۱	—	—	—	χ^2برابر ۱۰/۳۳و p برابر ۰/۰۰۵۷	Friedman Test
الگوریتم پیشنهادی بهترین میانگین رتبه را دارد	۲/۸۳	۲/۰۰	۱/۱۷	رتبه (۱=بهتر)	**میانگین رتبهٔ فریدمن**
CD کمتر از ۰/۸۵ → تفاوت معنادار نیست	—	—	۰/۸۳	Rank (Proposed–Spectral)	Nemenyi Post-hoc
CD بیشتر از ۰/۸۵ → تفاوت معنادار	—	—	۱/٦٦	Rank (Proposed–K-Means)	
CD برابر ۰/۸۵ در سطح اطمینان ۹۵ درصد	—	—	—	مقدار	Critical Difference (CD)
الگوریتم پیشنهادی از K- Meansبه طور معنادار بهتر و از Spectralکمی بهتر است	ضعیف‌تر در تمام داده‌ها	مشابه الگوریتم پیشنهادی در برخی از داده‌ها	الگوریتم پیشنهادی بهترین عملکرد کلی	تفسیر کلی	**نتیجهٔ نهایی**

جدول (۹): مطالعهٔ ابلیشن الگوریتم پیشنهادی در داده‌های واقعی

تحلیل کیفی	تغییر عملکرد نسبت به مدل اصلی	دقت ± SD میانگین	NMI ± SD میانگین	توضیح مؤلفهٔ جایگزین	تغییر اعمال‌شده
بهترین عملکرد، پایداری زیاد	—	۰/۲۱٤±۰/۷۳۳	۰/۲۲±۰/۷۳۹	پیکربندی پایهٔ الگوریتم پیشنهادی	مدل اصلی آپولونیوس + L + kNN متقابل + تعبیه طیفی
افت جزئی در داده‌های پیچیده، کرنل آپولونیوس به فواصل بلند حساس‌تر است	%-٦/٥	۰/۲۱±۰/٦٨٤	۰/۱۹±۰/٦۹۱	$S_{ij} = \exp-\left(\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$	جایگزینی کرنل آپولونیوس با کرنل گوسی
پایدارتر در داده‌های نامتوازن، اما بدون بهبود معنادار	%-۲/۱	۰/۲۰±۰/۷۲٤	۰/۲۰±۰/۷۲٦	نرمال‌سازی متقارن	تغییر لاپلاسیَن از L به L_sym
کاهش پایداری به ویژه در داده‌های نویزی	%-۷/٤	۰/۱۹±۰/٦۹۳	۰/۱۸±۰/۷۰٥	حذف شرط تقارن در همسایگی	تغییر گراف kNN متقابل ↔ ساده
افت شدید عملکرد، تأیید اهمیت تعبیهٔ طیفی	%-۱/۲٤	۰/۲۱±۰/٥٤۱	۰/۲۳±۰/٥٥۲	حذف گام طیف نگاری	اجرای K میانگین استاندارد در فضای اصلی (بدون تعبیهٔ طیفی)
بهینه‌ترین ترکیب از دید NMI و ACC	—	۰/۲۱٤±۰/۷۳۳	۰/۲۲±۰/۷۳۹	پیکربندی نهایی الگوریتم	اجرای K میانگین استاندارد پس از تعبیهٔ طیفی (مدل کامل)

نتایج جدول (۹) نشان می‌دهد هستهٔ آپولونیوس نقشی کلیدی در عملکرد مدل دارد. جایگزینی آن با هستهٔ گوسی باعث کاهش میانگین NMI از ۰/۷۳۹ به ۰/٦۹۱ شده است. این افت به ویژه در داده‌هایی با ساختار غیرخطی (مانند *Digits* و *Olivetti Faces*) بیشتر بوده است، زیرا کرنل آپولونیوس حساسیت کمتری به تغییرات محلی دارد و روابط فاصله‌ای را در بازه‌های طولانی‌تر به صورت دقیق‌تر مدل می‌کند.

تغییر نوع لاپلاسیَن از غیرنرمال‌شده به نرمال‌شدهٔ متقارن تأثیر کمی بر میانگین عملکرد گذاشت (حدود ۲ درصد کاهش)، اما در داده‌های نامتوازن مانند *Diabetes*، پایداری بیشتری در همگرایی خوشه‌ها ایجاد کرد. در مقابل، تغییر ساختار گراف از kNN متقابل به kNN ساده باعث افت محسوس عملکرد شد. گراف متقابل باعث حذف یال‌های غیرمعنادار می‌شود و در نتیجه، ساختار خوشه‌ای تمیزتری به دست می‌دهد.

در نهایت، مهم‌ترین بخش مطالعهٔ ابلیشن مربوط به گام تعبیهٔ طیفی است. اجرای مستقیم K میانگین استاندارد در فضای اصلی منجر به افت شدید (در NMI از ۰/۷۳۹ به ۰/٥٥۲) شد. این امر نشان می‌دهد فضای تعبیهٔ طیفی با کاهش نویز و آشکارسازی ساختارهای غیرخطی، برای خوشه‌بندی ضروری است.

در مجموع، آزمایش‌ها نشان می‌دهند هر چهار مؤلفهٔ کرنل آپولونیوس، لاپلاسیَن مناسب، گراف متقابل و تعبیهٔ طیفی، در

کنار هم، موجب افزایش چشـــمگیر دقت و پایداری الگوریتم پیشنهادی شده‌اند.

تحلیل پیچیدگی زمانی الگوریتم پیشنهادی

در الگوریتم خوشـــه‌بندی طیفی بدون اســـتفاده از PCA، ابتدا داده‌ها نرمال‌ســـازی می‌شـــوند که هزینه‌ای در حد $O(n*d)$ دارد. سـپس، فاصـلهٔ زوجی بین تمام نمونه‌ها در بُعد d محاسبه می‌شود که یکی از پرهزینه‌ترین مراحل است و زمان اجرای آن $O(n^2*d)$ اســـت. ســـاخت گراف-k نزدیک‌ترین همســـایه‌ها نیز معمولاً همان مرتبهٔ پیچیدگی را دارد، مگر از روش‌های سریع‌تر استفاده شود. تشکیل ماتریس شباهت و محاسبهٔ لاپلاسیان گراف نیز به ترتیب نیازمند زمان و حافظه‌ای در حدود $O(n*t*k*r)$ اســت. مرحلهٔ بعدی تجزیهٔ ویژهٔ لاپلاسیان است که اگر روی ماتریس چگال انجام شـــود، پیچیدگی زمانی $O(n^3)$ و حافظه‌ای $O(n^2)$ دارد. در پایان، الگوریتم K میانگین اســـتلندارد روی embedding حاصـــل اجرا می‌شـــود که در هر تکرار هزینه‌ای در حد $O(t*n*k*r)$ (که در t تکرار $r=k-1$) دارد. در مجموع، پیـــچـــیـــدگـــی کـــلـــی الـــگـــوریـــتـــم بـــرابـــر $O(n*d)+O(n^2*d)+O(n^3)+O(n*t*k*r)$ است.

۵- نتیجه‌گیری

الگوریتم پیشـــنهادی با اســـتفاده از هســـتهٔ آپولونیوس و ســـاختار گراف، در برخی از مجموعه‌داده‌ها عملکردی بهتر نســـبت به خوشـــه‌بندی طیفی دارد، اما در برخی دیگر، خوشـــه‌بندی طیفی نتایجی بهتر ارائه می‌دهد. بنابراین، انتخاب الگوریتم مناسب به ویژگی‌های داده‌ها و ساختار آنها بستگی دارد. بررســـی عملکرد الگوریتم پیشـــنهادی مبتنی بر هســـتهٔ آپولونیوس در مقایسـه با خوشـه‌بندی طیفی نشـان می‌دهد این الگوریتم در برخی از مجموعه‌داده‌ها مانندWine و Digits عملکردی بهتر دارد. برای مثـال، در مجموعـه‌دادهٔ Wine، الگوریتم پیشـــنهادی باNMI برابر ۰/۸۵۳ و دقت برابر ۰/۹۵۵، نسبت به خوشه‌بندی طیفی با NMI برابر ۰/٤۵ و دقت ۰/۷۳، نتایجی بهتر ارائه داده اسـت. این بهبود ممکن اسـت ناشـی از استفاده از هستهٔ آپولونیوس در ساخت گراف شباهت باشد که ســـاختارهای محلی و غیرخطی داده‌ها را بهتر مدل‌ســـازی می‌کند. با این حال، در مجموعه‌داده‌هایی مانند Iris، Cancer و Breast، خوشـــه‌بندی طیفی عملکردی بهتر نســـبت به الگوریتم پیشـــنهادی دارد. برای مثال، در مجموعـه‌دادهٔ Iris، خوشـــه‌بندی طیفی با NMI (۰/۷٦) و دقت ۰/۹۰، نسـبت به الگوریتم پیشـنهادی با NMI (۰/٤٦۰) و دقت ۰/٦۸۷، نتایجی بهتر ارائه داده اســـت. این امر ممکن اســـت به دلیل ســـادگی ساختار داده‌های این مجموعه‌ها باشـد که در آن، الگوریتم‌های مبتنی بر طیف قادر به شـناسـایی خوشـه‌ها با دقت بیشـتری هستند.

مراجع

[1] A. Y. Ng, M. I. Jordan, Y. Weiss, “On spectral clustering: Analysis and an algorithm”, in T. G. Dietterich, S. Becker, Z. Ghahramani (Eds.), Advances in Neural Information Processing Systems 14 (NIPS 2001), Cambridge, MA: MIT Press, 2002. [Online]. Available: https://papers.nips.cc/paper_files/paper/2001/hash/801272ee79cfde7fa5960571fee36b9b-Abstract.html

[2] J. MacQueen, “Some methods for classification and analysis of multivariate observations”, in Proc. 5th Berkeley Symp. Math. Statist. Probab., Vol. 1, pp. 281–297, Berkeley, CA: University of California Press, 1967. https://www.semanticscholar.org/paper/ac8ab51a86f1a9ae74dd0e4576d1a019f5e654ed

[3] B. Park, C. Park, S. Hong, H. Choi, “Sparse kernel k-means clustering”, J. Appl. Stat., Vol. 52, No. 1, pp. 158–182, 2025. https://doi.org/10.1080/02664763.2024.2362266

[4] H. He, “Clustering algorithm based on Hopkins statistics and K-means”, Multimedia Tools Appl., pp. 1–26, Feb. 2025. https://doi.org/10.1007/s11042-025-20693-6

[5] X. Liu, “SimpleMKKM: Simple multiple kernel K-means”, IEEE Trans. Pattern Anal. Mach. Intell., Vol. 45, No. 4, pp. 5174–5186, Apr. 2023. https://doi.org/10.1109/TPAMI.2022.3198638

[6] Z. Li, C. Tang, X. Zheng, Z. Wan, K. Sun, W. Zhang, X. Zhu, “Mutual structure learning for multiple kernel clustering”, Inf. Sci., Vol. 647, Art. No. 119445, Nov. 2023. https://doi.org/10.1016/j.ins.2023.119445

[7] S. Pourbahrami, M. A. Balafar, L. M. Khanli, “ASVMK: A novel SVMs Kernel based on Apollonius function and density peak clustering”, Engineering Applications of

[11] S. Pourbahrami, M. A. Balafar, L. M. Khanli, Z. A. Kakarash, “A survey of neighborhood construction algorithms for clustering and classifying data points”, Computer Science Review, Vol. 38, p. 100315, 2020.
https://doi.org/10.1016/j.cosrev.2020.100315
[12] S. Pourbahrami, L. M. Khanli, S. Azimpour, “A novel and efficient data point neighborhood construction algorithm based on Apollonius circle”, Expert Systems with Applications, Vol. 115, pp. 57-67, 2019.
https://doi.org/10.1016/j.eswa.2018.07.066
[13] N. Abdolmaleki, L. M. Khanli, M. Hashemzadeh, S. Pourbahrami, “ACQC: Apollonius Circle-based Quantum Clustering”, Journal of Computational Science, Vol. 64, p. 101877, 2022.
https://doi.org/10.1016/j.jocs.2022.101877

Artificial Intelligence, Vol. 126, p. 106704, 2023.
https://doi.org/10.1016/j.engappai.2023.106704
[8] Z. Zhang, X. Liu, L. Wang, “Spectral clustering algorithm based on improved Gaussian kernel function and beetle antennae search with damping factor”, Comput. Intell. Neurosci., Vol. 2020, Art. No. 1648573, 2020. https://doi.org/10.1155/2020/1648573
[9] H. Zhou, Z. Wang, H. Chen, X. Wang, “A novel spectral clustering algorithm based on neighbor relation and Gaussian kernel function with only one parameter”, Soft Comput., Vol. 28, No. 2, pp. 981–989, Jan. 2024. https://doi.org/10.1007/s00500-023-09309-z
[10] S. Pourbahrami, M. Hashemzadeh, “A geometric-based clustering method using natural neighbors”, Information Sciences, Vol. 610, pp. 694-706, 2022. https://doi.org/10.1016/j.ins.2022.08.047

[1] Spectral Clustering
[2] k-Means
[3] Hopkins Statistic
[4] MSL-MKC
[5] Laplacian regularization
[6] Support Vector Machines (SVM)
[7] Spectral Clustering with Neighbor Relations
[8] Sparse Kernel *k*-Means for Feature Selection in Nonlinear Clustering (SKKM)
[9] smoothing factor
[10] Accuracy
[11] Indicator Function
[12] https://archive.ics.uci.edu
https://www.kaggle.com