DOR: 20.1001.1.2538161.1403.9.2.2.3.

DOI : 10.22034/csj.2024.203192

تابستان ۱۴۰۳، دوره ۹، شماره ۲، پیاپی ۳۳

دریافت مقاله: ۱۴۰۳/۰۱/۲۷

پذیرش مقاله: ۱۴۰۳/۰۲/۲۲

نوع مقاله: پژوهشی

یک معیار شباهت مبتنی بر محبوبیت برای بهبود کارایی خوشه‌بندی طیفی

حسن مطلبی*

استادیار، دانشکده مهندسی برق و کامپیوتر، دانشگاه تحصیلات تکمیلی صنعتی و فناوری پیشرفته، کرمان، ایران

پست الکترونیکی: h.motallebi.p@gmail.com

چکیده

روش‌های خوشه‌بندی طیفی، به دلیل قابلیتی که در تشخیص خوشه‌های با شکل‌های مختلف دارند، بسیار مورد توجه قرار گرفته‌اند. کارایی این روش‌ها وابستگی شدیدی به نحوه تعریف شباهت بین نمونه‌ها دارد. بنابراین، تلاش‌ها برای بهبود کارایی این الگوریتم‌ها بر روی ارایه معیار شباهت مناسب‌تر متمرکز بوده است. در این مقاله، به هر نمونه شاخصی تحت عنوان شاخص محبوبیت نسبت می‌دهیم که بیانگر میزان مرکزیت آن نمونه در مجموعه داده است. همچنین، معیار شباهت جدیدی مبتنی بر محبوبیت نمونه‌ها پیشنهاد و بر پایه آن الگوریتم خوشه‌بندی طیفی مبتنی بر محبوبیت را پیشنهاد می‌دهیم. از آنجا که شاخص محبوبیت پیشنهادی مستقل از چگالی محلی خوشه‌هاست، الگوریتم پیشنهادی می‌تواند در خوشه‌بندی داده‌های با چگالی‌های متفاوت موفق عمل کند. معیار پیشنهادی ویژگی سودمند دیگری نیز دارد؛ شباهت نمونه‌ها در معیار پیشنهادی با توجه به محبوبیت آنها و جایگاه یک نمونه در لیست همسایگان نمونه دیگر محاسبه می‌شود. این ویژگی به جداسازی خوشه‌های با همپوشانی بالا کمک بسیاری می‌کند. به دلیل سادگی تعریف پیشنهادی برای شاخص محبوبیت، الگوریتم محاسبه ماتریس شباهت پیشنهادی پیچیدگی محاسباتی بسیار پایینی دارد. برای مطالعه و مقایسه کارایی الگوریتم خوشه‌بندی پیشنهادی با همتاهای آن، از منظر معیار انطباق NMI، آزمایش‌هایی بر روی شش مجموعه داده مصنوعی و پانزده مجموعه داده واقعی انجام داده‌ایم. نتایج نشان می‌دهد که الگوریتم خوشه‌بندی پیشنهادی و معیار شباهت مطرح در آن کارایی بهتری نسبت به روش‌های همتای آن از جمله معیار شباهت مبتنی بر میانگین محلی، معیار خود- تنظیم و معیار شباهت محلی مبتنی بر همسایه‌های مشترک دارد و در اغلب موارد بهترین عملکرد را به‌دست می‌آورد.

واژه‌های کلیدی: خوشه‌بندی طیفی، شاخص محبوبیت، معیار شباهت مبتنی بر محبوبیت، گراف همسایگی، ماتریس شباهت.

۱. مقدمه

خوشه‌بندی عبارت است از دسته‌بندی نمونه‌ها در گروه‌های (خوشه‌های) جداگانه به نحوی که نمونه‌های درون هر خوشه به هم شبیه و متفاوت از نمونه‌ها در خوشه‌های دیگر باشند. از جمله اهداف خوشه‌بندی، به‌دست آوردن توصیفی از داده از طریق خلاصه‌سازی و همچنین شناسایی و قرار دادن نمونه‌های با ویژگی‌های مشترک و مشابه در گروه‌های جداگانه است. خوشه‌بندی در زمینه‌های مختلفی از جمله تحلیل اسناد[1] [۱]، شناسایی

1-Document analysis

* نویسندهٔ مسئول

الگو[۲] [۲]، بخش‌بندی تصاویر[۳] [۳] و غیره کاربرد دارد. یکی از چالش‌های مهم خوشه‌بندی در دنیای واقعی، تشخیص خوشه‌های با اندازه‌ها، چگالی‌ها و شکل‌های متفاوت در داده‌های شامل نوفه[۴] است.

تاکنون، الگوریتم‌های خوشه‌بندی بسیاری برای تشخیص خوشه‌های با شکل‌های مختلف، از جمله شکل‌های غیرکروی و نامنظم پیشنهاد شده است. به عنوان مثال، الگوریتم‌های خوشه‌بندی سلسله‌مراتبی[۵] [۴،۵] از جمله الگوریتم‌هایی هستند که می‌توانند خوشه‌های با شکل‌های دلخواه را تشخیص دهند. یک راه‌حل دیگر برای تشخیص خوشه‌ها با شکل‌های دلخواه، استفاده از الگوریتم‌های خوشه‌بندی طیفی [۶-۹] است. در این الگوریتم‌ها، با استفاده از ویژگی‌های طیفی[۶] ماتریس‌های خاصی، که از ماتریس (گراف) شباهت به دست می‌آیند، داده‌ها به فضایی با ابعاد بسیار کمتر نگاشت می‌شوند. مقادیر و بردارهای ویژه از جمله مهم‌ترین ویژگی‌های طیفی هستند. سپس، برای رسیدن به نتیجه نهایی، یک روش خوشه‌بندی سنتی مانند k-means بر روی داده‌ها در این فضای جدید اعمال می‌شود. ساخت ماتریس (گراف) شباهت[۷] مناسبی که شباهت بین نمونه‌ها را به خوبی منعکس کند برای رسیدن به کارایی مناسب در خوشه‌بندی طیفی بسیار کلیدی است. برای این کار، سه نوع گراف شباهت مرسوم متداولاً مورد استفاده قرار می‌گیرند [۱۰]

- گراف کاملاً متصل[۸]، که در آن همه گره‌ها به هم متصل هستند. این گراف تنها در صورتی سودمند است که وزن یال‌ها منعکس‌کننده شباهت بین نمونه‌ها باشد. معمولاً وزن یال بین x_i و x_j، که بیانگر شباهت این دو نمونه است، با استفاده از تابع شباهت گوسی[۹]، به صورت زیر محاسبه می‌شود:

$$exp\ (-d(x_i,x_j)^2/(2\sigma^2)) \quad (۱)$$

که $d(x_i,x_j)$ فاصله x_i و x_j و σ پارامتر هسته[۱۰] است که شتاب کاهش میزان شباهت با افزایش فاصله بین نقاط را کنترل می‌کند. از آنجا که انتخاب مقدار مناسب برای این پارامتر تاثیر قابل توجهی بر کارایی الگوریتم دارد، مطالعات بسیاری برای تعیین مقدار مناسب برای آن انجام شده است [۱۱].

- گراف همسایگی ε[۱۱]، که در آن یالی بین دو گره x_i و x_j کشیده می‌شود اگر فاصله آنها بیش از ε نباشد. پارامتر ε نقشی مانند پارامتر σ در گراف کاملاً متصل ایفا می‌کند.
- گراف-k نزدیک‌ترین همسایه[۱۲] (knn)، که در آن یالی از گره x_i به x_j کشیده می‌شود اگر و تنها اگر x_j یکی از k نزدیک‌ترین همسایه x_i باشد. از آنجا که رابطه همسایگی رابطه‌ای مقارن نیست گراف حاصل گرافی جهت‌دار خواهد بود. دو راه برای تبدیل این گراف به یک گراف بدون جهت وجود دارد. اولین روش این است که جهت یال‌ها را نادیده بگیریم، یعنی، اگر یکی از دو شرط $x_i \in knn(x_j)$ یا $x_j \in knn(x_i)$ برقرار باشد یک یال بدون جهت بین x_i و x_j می‌کشیم که $knn(\cdot)$ بیانگر مجموعه k نزدیک‌ترین همسایه یک نمونه است. روش دیگر این است که اگر هر دو شرط $x_i \in knn(x_j)$ و $x_j \in knn(x_i)$ برقرار بود یک یال بدون جهت بین x_i و x_j بکشیم. در این حالت، به گراف حاصل گراف k نزدیک‌ترین همسایه متقابل[۱۳] گفته می‌شود. در هر دو مورد، وزن یال‌های کشیده شده با توجه به شباهت گره‌های مربوطه مشخص می‌شود.

معمولاً، گراف‌های شباهت خلوت[۱۴]، از قبیل گراف همسایگی ε و گراف knn به دلیل ساختار ساده‌تر و واضح‌تر آنها ترجیح داده می‌شوند. بسیاری از کارهایی که با هدف بهبود کارایی خوشه‌بندی طیفی انجام شده‌اند

2-Pattern recognition
3-Image segmentation
4-Noise
5-Hierarchical clustering
6-Spectral properties
7-Affinity matrix
8-Fully connected graph
9-Gaussian similarity function
10-Kernel parameter
11-ε-neighborhood graph
12-k-nearest neighbor (kNN) graph
13-Mutual kNN graph
14-Sparse

این کار را با تولید گراف‌های شباهت مؤثرتر انجام می‌دهند [۱۰-۱۸]. بـه بیان دیگر، الگوریتم‌های خوشـه‌بندی طیفی متفاوت، برای رسـیدن به نتایج بهتر، معیارهای شـباهت متفاوتی را مورد استفاده قرار داده‌اند.

۱-۱. نوآوری تحقیق

در این مقاله، الگوریتم خوشه‌بندی طیفی جدیدی به نام «الگوریتم خوشه‌بندی طیفی مبتنی بر محبوبیت[15] (PBSC)» پیشـنهاد شده که در آن گراف همسایگی با استفاده از یک معیار شـباهت جدید مبتنی بر محبوبیت نمونه‌ها سـاخته می‌شـود. مهمتریـن نوآوری‌های این الگوریتـم در ادامه آورده شده است:

۱. مفهوم جدیـدی تحت عنوان «محبوبیت» معرفی شـده اسـت. در الگوریتم پیشـنهادی به هر نمونه یک میزان محبوبیت نسبت داده می‌شود. میزان محبوبیت یک نمونه بیانگر میزان مرکزیت آن نمونه در مجموعه داده است و با توجه به اعتباری که آن نمونه از طریق همسـایگی با دیگر نمونه‌ها دریافت می‌کند و همچنین میزان محبوبیت آن همسایه‌ها محاسبه می‌شود.

۲. بر مبنـای مفهوم محبوبیت، یک گراف شـباهت خلوت سـاخته می‌شـود که روابط همسـایگی نمونه‌ها را به خوبی منعکس می‌کند. شـباهت بیـن هر دو نمونه x_i و x_j در این گراف بر اساس محبوبیت آن‌ها و موقعیت x_i در لیست نزدیک‌ترین همسایگان x_j و بالعکس محاسبه می‌شود. معیار شـباهت پیشنهادی چند ویژگی سومند دارد: اولیـن ویژگی، که کمک به خوشـه‌بندی داده‌های چندمقیاسـه می‌کند، این اسـت که محبوبیـت نمونه‌ها مسـتقل از مقیاس و چگالی محلی آنها اسـت. ویژگی دیگر این است که در معیار شباهت پیشنهادی، شباهت بین نمونه‌ها در محدوده‌های خلوت کمتر از شباهت در محدوده‌های پرتراکم است. همچنین، این معیار شباهت به نقاط نزدیک به هم، در صورتی که هم‌خوشـه باشند میزان شـباهت بالا و اگر در خوشه‌های متفاوت باشند میان شـباهت پایینی نسـبت می‌دهد. این دو ویژگی به جداسازی خوشـه‌های دارای همپوشانی کمک می‌کند. نهایتاً، ویژگی دیگر معیار شـباهت پیشنهادی پیچیدگی محاسباتی بسیار پایین آن است.

برای مطالعه کارایی الگوریتم خوشـه‌بندی پیشنهادی آزمایش‌هایی بر روی شـش مجموعـه داده مصنوعی و پانزده مجموعه داده واقعی انجام داده‌ایم. کارایی الگوریتم PBSC پیشـنهادی را علاوه بر الگوریتم خوشه‌بندی طیفی سـنتی [۹] با الگوریتم‌های خوشـه‌بندی طیفـی‌ای که بر مبنای معیار شـباهت مبتنی بر میانگین محلی [۱۵]، معیار شـباهت خود-تنظیم [۱۲] و معیار شـباهت محلی مبتنی بر همسایه‌های مشترک اسـتوار هستند مقایسه کرده‌ایم. عـلاوه بر ایـن، الگوریتم پیشـنهادی را با پنـج الگوریتم خوشـه‌بندی دیگر، از گروه‌های مختلف، مقایسه کرده‌ایم. نتایج آزمایش‌های انجام شده و مقایسه اثربخشی و کارایی روش PBSC پیشـنهادی با همتاهای آن نشـان می‌دهد که این الگوریتم در مقایسـه با الگوریتم‌های پیشـین از منظر معیار انطبـاق NMI[16] موفق‌تر عمل می‌کند. به طور خاص، الگوریتم خوشـه‌بندی PBSC برای سیزده مجموعه داده از پانـزده مجموعه داده واقعی، بهتریـن عملکرد و برای دو مورد دیگر دومین بهترین عملکرد را دارد.

۱-۲. ساختار مقاله

در ادامـه ایـن مقاله، در قسـمت دوم به مـروری بر الگوریتم‌های خوشـه‌بندی طیفی موجـود می‌پردازیم. در قسمت سوم به شرح الگوریتم خوشه‌بندی طیفی مبتنی بر محبوبیت پیشنهادی و در قسمت چهارم، به مطالعه کارایی آن می‌پردازیم. نهایتاً، در قسمت پنجم به نتیجه‌گیری بحث خواهیم پرداخت.

۲. مرور ادبیات

بـرای رسـیدن بـه کارایـی مناسـب در روش‌های خوشه‌بندی طیفی، به‌دست آوردن گراف شباهت مناسبی

15-Popularity-Based Spectral Clustering (PBSC)

16-Normalized Mutual Information (NMI)

که شباهت و همسایگی نمونه‌ها را به خوبی منعکس کند بسیار کلیدی است. تابع شباهت گوسی یکی از متداول‌ترین توابع برای ساخت ماتریس شباهت است که در آن پارامتر σ عرض محدوده همسایگی را کنترل می‌کند. در [۱۰] مقدار مناسب برای این پارامتر برابر میانگین فاصله بین نقاط تا kامین نزدیک‌ترین همسایه آنها پیشنهاد شده و مقدار مناسب برای پارامتر k برابر $k \approx log(n) + 1$ در نظر گرفته شده است. مشکل این روش این است که برای بسیاری از داده‌ها در نظر گرفتن یک مقدار ثابت σ برای همه نقاط مناسب نیست.

در [۱۲] الگوریتم خوشه‌بندی طیفی خودتنظیم[17] معرفی شده که برای هر نقطه مثل x_i پارامتر هسته محلی جداگانه σ_i را در نظر می‌گیرد. شباهت بین x_i و x_j در الگوریتم مذکور با توجه به چگالی همسایگی x_i و x_j محاسبه می‌شود. در [۱۱] نویسندگان به جای تلاش برای یافتن مقدار مناسب σ ، یک تابع هسته گوسی توانی[18] پیشنهاد داده‌اند. در [۱۷] نویسندگان استفاده از یک مقدار سراسری σ، که با میانگین گرفتن از همه σ_iها به‌دست می‌آید را پیشنهاد کرده‌اند. در [۱۸] یک معیار فاصله مبتنی بر knn پیشنهاد شده است. در روش مذکور به جای پارامتر σ از پارامتر k استفاده شده که به دلیل صحیح بودن، انتخاب مقدار مناسب برای آن راحت‌تر است. در [۱۹] برای اجتناب از تأثیر نوفه‌ها ماتریس شباهتی مبتنی بر گراف k نزدیک‌ترین همسایه متقابل ساخته شده است.

در [۲۰] از یک گراف knn تغییریافته استفاده شده که در آن تعداد همسایه‌ها برای هر نقطه متناسب با شرایط آن نقطه مشخص می‌شود. همچنین، در [۲۱] نیز برای بهبود ساختار گراف، برای نقاط مختلف داده مقادیر مختلف k در نظر گرفته شده است. در [۱۴] نویسندگان روشی برای ساخت ماتریس شباهت بر اساس چگالی محلی نقاط پیشنهاد داده‌اند. همچنین، نویسندگان در [۱۵] یک معیار شباهت مبتنی بر میانگین محلی پیشنهاد داده‌اند که شباهت نمونه‌ها در آن بر اساس فاصله بین میانگین k نزدیک‌ترین همسایه آن‌ها محاسبه می‌شود. در [۱۶] نویسندگان یک معیار شباهت محلی با لحاظ کردن چگالی محلی و مبتنی بر وزن همسایگان مشترک در گراف جهت‌دار k نزدیک‌ترین همسایه پیشنهاد داده‌اند. همچنین، در [۲۲] با هدف ارایه یک معیار شباهت مناسب برای جداسازی خوشه‌های با تفکیک ضعیف[19] شباهت نقاط با توجه به تعداد و نزدیکی همسایگان مشترک آنها محاسبه شده است. در [۲۳] یک گراف شباهت بدون پارامتر ارائه شده که اطلاعات مربوط به فاصله، چگالی و اتصال نقاط را برای ساختن گراف همسایگی محلی ترکیب می‌کند. در [۲۴] گراف تطبیقی شباهت همسایگان[20] (ANSG) معرفی شده که به یال‌های بین یک نمونه و k نزدیک‌ترین همسایه‌اش احتمالاتی را نسبت می‌دهد. یال‌های با فواصل کوتاه در این گراف احتمال بالاتری به خود اختصاص می‌دهند که این منجر به اتصالات قوی بین گره‌های نزدیک به هم می‌شود. نویسندگان در [۲۵] گراف شباهت فازی[21] (FSG) را معرفی کرده‌اند که با تمام نقاط داده مانند مراکز برخورد می‌کند. گراف شباهت در تحقیق مذکور بر مبنای الگوریتم c-means و ماتریس عضویت فازی ساخته می‌شود. نهایتاً، برای مواجهه با نقاط پرت و نوفه در داده‌ها، در [۲۶] روشی مبتنی بر فاصله میانگین تعمیم یافته[22] پیشنهاد شده که حساسیت به مقدار پارامتر k یعنی تعداد همسایه‌ها را کاهش می‌دهد.

در جدول ۱ خلاصه‌ای از مشخصات، مزایا و معایب الگوریتم‌های خوشه‌بندی طیفی مذکور آورده شده است.

پرداختن همزمان به سه چالش ۱. تشخیص خوشه‌ها در داده‌های حاوی نوفه، ۲. خوشه‌بندی داده‌های با سطوح مختلف چگالی و۳. جداسازی خوشه‌های با تفکیک ضعیف مسئله‌ای است که در روش‌های ذکر شده فوق کمتر بر روی آن تمرکز شده است. در این مقاله به این ارایه روشی برای حل این چالش‌های سه‌گانه پرداخته شده است.

17-Self-turning spectral clustering
18-Powered Gaussian kernel function

19-Poor separation
20-Adaptive neighbors similarity graph
21-Fuzzy similarity graph
22-Generalized mean distance

جدول ۱- مشخصات، مزایا و معایب الگوریتم‌های خوشه‌بندی طیفی موجود

مرجع	محلی/سراسری	پارامتر(ها)	مزایا و معایب
[۱۰]	سراسری	عرض همسایگی، σ	عدم توانایی تشخیص خوشه‌های با سطوح مختلف چگالی
[۱۱]	سراسری	پارامتر توان، β	عدم توانایی تشخیص خوشه‌های با سطوح مختلف چگالی، ارایه روشی برای تخمین مقدار مناسب برای پارامتر β
[۱۲]	محلی	عرض همسایگی، σ	توانایی تشخیص خوشه‌های با سطوح مختلف چگالی
[۱۴]	محلی	عرض همسایگی، σ و شعاع همسایگی، ε	توانایی تفکیک خوشه‌های با تفکیک ضعیف، توانایی تطبیق با شرایط همسایگی محلی
[۱۵]	محلی	تعداد همسایه، k و عرض همسایگی، σ	توانایی تفکیک خوشه‌های با تفکیک ضعیف، عدم توانایی تشخیص خوشه‌های با سطوح مختلف چگالی
[۱۶]	محلی	تعداد همسایه، k	لحاظ کردن اطلاعات همسایگی محلی
[۱۷]	سراسری	عرض همسایگی، σ	عدم توانایی تشخیص خوشه‌های با سطوح مختلف چگالی
[۱۸]	سراسری	تعداد همسایه، k	مقاوم در برابر نوفه، توانایی کار با داده‌های چندمقیاسه
[۱۹]	سراسری	تعداد همسایه، k	مقاوم در برابر نوفه، عدم توانایی کار با داده‌های چندمقیاسه
[۲۰]	محلی	تعداد همسایه، k	مقاوم در برابر نوفه، توانایی یافتن خوشه های خلوت و کاهش زمان محاسبه
[۲۱]	محلی	حداکثر تعداد همسایه، k_{max}	بهبود ساختار گراف با استفاده از روش تطبیقی، تنظیم کردن تعداد همسایه‌های هر نمونه با توجه به موقعیت و شرایط آن
[۲۲]	محلی	تعداد همسایه، k	توانایی تفکیک خوشه‌های با تفکیک ضعیف
[۲۳]	محلی	بدون پارامتر	عدم نیاز به مقداردهی پارامترها، حساسیت به نوفه
[۲۴]	محلی	تعداد همسایه، k	توانایی کار با ابعاد بالای داده
[۲۵]	محلی	پارامتر توان، β، و پارامترهای فازی r و m	تشخیص هم‌زمان ماتریس فاصله و ساختار داده‌ها، عدم توانایی تشخیص خوشه‌های با سطوح مختلف چگالی
[۲۶]	سراسری	تعداد همسایه، k	عدم توانایی تشخیص خوشه‌های با سطوح مختلف چگالی، ارایه روش پیاده‌سازی موازی

۳. الگوریتم خوشه‌بندی طیفی مبتنی بر محبوبیت

در این قسمت، به توضیح الگوریتم خوشه‌بندی طیفی مبتنی بر محبوبیت پیشنهادی (PBSC) و مفاهیم مبنایی آن می‌پردازیم. نخستین مرحله از این الگوریتم محاسبه شاخص محبوبیت نمونه‌ها است. در ادامه، در ابتدا نحوه محاسبه محبوبیت نمونه‌ها شرح داده می‌شود. سپس به توضیح معیار شباهت مبتنی بر محبوبیت پیشنهادی و در نهایت به توضیح الگوریتم خوشه‌بندی پیشنهادی و پیچیدگی زمانی آن می‌پردازیم.

۳-۱. محاسبه محبوبیت نمونه‌ها

فرض کنید $X = \{x_1, \dots, x_n\}$ مجموعه داده‌ای شامل n نمونه و $d(x_i, x_j)$ بیانگر فاصله بین دو نمونه x_i و x_j در آن باشد. اگر rامین نزدیک‌ترین همسایه نمونه x_i را با $nn_r(x_i)$ نشان دهیم، مجموعه k نزدیک‌ترین همسایه x_i، که با $knn(x_i)$ نشان داده می‌شود، به صورت زیر تعریف می‌شود:

$$knn(x_i) = \{nn_0(x_i), nn_1(x_i), \dots, nn_k(x_i)\} \quad (۲)$$

که در آن صفرمین نزدیک‌ترین همسایه یک نمونه خود آن نمونه محسوب می‌شود.

تعریف ۱ (ماتریس توزیع اعتبار). ماتریس (توزیع) اعتبار $C = [c_{i,j}]_{n\times n}$، ماتریس مجموعه داده $X = \{x_1, \dots, x_n\}$ است که در آن $c_{i,j}$ مقدار اعتباری است که x_i به x_j می‌دهد. اگر rامین همسایه x_i نمونه x_j باشد خواهیم داشت:

$$c_{i,j} = w_r = \begin{cases} \alpha, & r = 0, \\ \alpha \cdot q^{k-r+1}, & r = 1, \dots, k, \\ 0, & r > k, \end{cases} \quad (۳)$$

که در آن $\alpha > 0$، $q > 1$ و w_r مقدار اعتباری است که هر نمونه به rامین نزدیک‌ترین همسایه خود می‌دهد $(r = 0, \dots, k)$. w_rها طوری مقداردهی می‌شوند که ۱) مجموع اعتباری که یک نمونه به همسایه‌های صفرم تا kام خود می‌دهد برابر با یک باشد، یعنی $\sum_{r=0}^{k} w_r = 1$

و اعتباری که یک نمونه به همسایه‌های نزدیک‌تر (غیر از خودش) می‌دهد بیشتر از همسایه‌های دورتر باشد، یعنی:

$$w_r \geq w_{r+1}, r = 1, \dots, k-1. \quad (٤)$$

توزیع اعتبار بین همسایه‌ها، که با دنباله $\langle w_1, w_2, \dots, w_{k-1}, w_k, w_0 \rangle$ نشان داده می‌شود، بر اساس یک توزیع هندسی بریده شده[۲۳] انجام می‌شود. داریم:

$$\langle w_1, \dots, w_{k-1}, w_k, w_0 \rangle = \langle q^k \alpha, \dots, q^2 \alpha, q\alpha, \alpha \rangle. \quad (٥)$$

با توجه به این که مقدار q بزرگتر از یک است، $w_0 = \alpha$ ، یعنی اعتباری که یک نمونه به خود می‌دهد کم‌ترین سهم را دارد و سهم آن کمتر از مقدار اعتباری است که یک نمونه به kامین نزدیک‌ترین همسایه خود می‌دهد. همچنین، اعتباری که یک نمونه به rامین نزدیکترین همسایه خود $(r = 1, \dots, k)$ می‌دهد برابر است با $w_r = \alpha q^{k-1+1}$. واضح است که اگر مقدار یکی از w_rها مشخص شود، می‌توان مقدار سایر w_rها را محاسبه کرد. در این مقاله، مقدار w_0 را به صورت زیر محاسبه می‌کنیم:

$$w_0 = \alpha = 10^{-(1+k/\sigma_{scale})} \quad (٦)$$

که در آن برای σ_{scale}، که پارامتر مقیاس نامیده می‌شود، مقدار $\sigma_{scale} = 20$ در نظر گرفته شده است. با استفاده از این پارامتر، در رابطه فوق می‌توان شتاب کاهش مقدار w_0 با افزایش مقدار k را کنترل کرد. نتایج آزمایش‌های انجام شده نشان می‌دهد که عملکرد الگوریتم پیشنهادی حساسیت چندانی به تغییرات نسبتاً اندک مقدار این پارامتر حول مقدار پیشنهادی ندارد و برای هر مقداری در بازه ۱۵ تا ۲۵ نتایج تقریباً یکسانی به‌دست می‌آید.

مفهوم محبوبیت در این مقاله را می‌توان همتایی برای مفهوم چگالی در نظر گرفت. میزان محبوبیت نمونه x_i با توجه به محبوبیت نمونه‌هایی که x_i را در لیست k نزدیک‌ترین همسایه خود دارند و همچنین جایگاه x_i در آن لیست، مطابق تعریف زیر محاسبه می‌شود.

تعریف ۲ (بردار محبوبیت). برای یک مجموعه داده مثل $X = \{x_1, \dots, x_n\}$، بردار محبوبیت، $\pi = (\pi_1, \dots, \pi_n)$ که در آن π_i شاخص محبوبیت نمونه x_i است، بردار ویژه متناظر با مقدار ویژه $\lambda = 1$ ماتریس توزیع اعتبار است.

از آنجا که ماتریس اعتبار، C، یک ماتریس سطری-تصادفی است، مسئلهٔ محاسبه بردار محبوبیت، π، همان مسئلهٔ یافتن بردار ویژه غالب[۲۴]، متناظر با مقدار ویژه $\lambda = 1$ این ماتریس است. تقریب t-ام از شاخص محبوبیت x_i، که با $\pi_i(t)$ نشان داده می‌شود، با استفاده از روش تکرار توان[۲۵]، به صورت زیر محاسبه می‌شود:

$$\pi_i(t) = \begin{cases} 1, & t = 0, \\ \sum_{j=1}^{n} \pi_j(t-1) \cdot c_{j,i}, & t = 1,2,\dots. \end{cases} \quad (٧)$$

طبق تعریف فوق، خواهیم داشت:

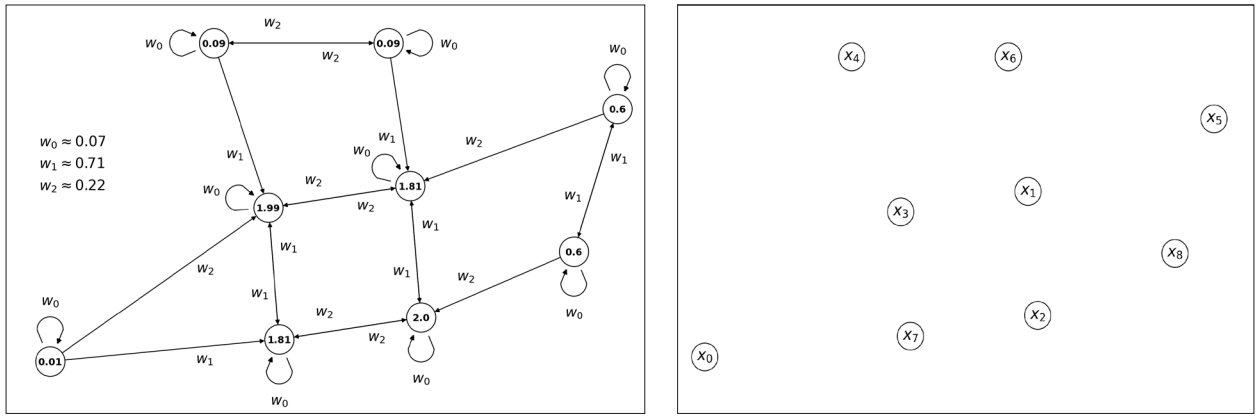
$$\pi(t) = \pi(t-1)C, t = 1,2,\dots \quad (٨)$$

که در آن $\pi(t)$ تقریب t-ام از بردار محبوبیت است. در واقع، بردار محبوبیت را می‌توان با شروع از تقریب اولیه $\pi(0) = (1, 1, \dots, 1)$ و به‌دست آوردن تقریب‌های متوالی با استفاده از رابطه ۸ تا رسیدن به همگرایی، محاسبه کرد.

در ادامه مقاله، در الگوریتم ۱ روال خوشه‌بندی طیفی پیشنهادی آورده شده است. در اولین گام از این الگوریتم، در سطرهای یکم تا دوازدهم، روند محاسبه ماتریس توزیع اعتبار و بردار محبوبیت آورده شده است. در راستای بررسی همگرایی الگوریتم، برای اندازه‌گیری اختلاف بین دو تقریب متوالی بردار محبوبیت، از معیار واگرایی کولبک- لیبلر[۲۶] استفاده می‌کنیم. معیار واگرایی کولبک-لیبلر [۲۷]، که با عنوان آنتروپی نسبی نیز شناخته می‌شود، یک مفهوم پایه‌ای در آمار و اطلاعات است که اختلاف بین دو توزیع احتمال را محاسبه می‌کند. بر اساس معیار واگرایی کولبک- لیبلر، اختلاف توزیع P نسبت به توزیع Q ، به شرط این که هر دو توزیع بر روی فضای نمونه X تعریف شده باشند، به صورت زیر محاسبه می‌شود:

23-Truncated geometric distribution

24-Dominant eigenvector
25-Power iteration method
26- Kullback–Leibler (KL) divergence



(ب) (الف)

شکل (۱): (الف) مثالی از مجموعه داده‌ای شامل نه نمونه و (ب) گراف متناظر با ماتریس توزیع اعتبار و محبوبیت نمونه‌ها در آن

$$D_{KL}(P \parallel Q) = \sum_{x \in X} P(x) \log\left(\frac{P(x)}{Q(x)}\right). \quad (۹)$$

لازم به ذکر است که در الگوریتم محاسبه بردار محبوبیت، برای تبدیل یک بردار محبوبیت به یک بردار توزیع جرم احتمال بردار محبوبیت نرمال‌سازی می‌شود تا مجموع عناصر آن یک شود. الگوریتم، تا وقتی که اختلاف بین دو تقریب متوالی مطابق معیار واگرایی کولبک- لیبلر بیشتر از $\varepsilon = 10^{-9}$ باشد به محاسبه تقریب‌های بعدی ادامه خواهد داد. همچنین، برای تضمین خاتمه این روند در صورت همگرایی کند، تعداد تکرارهای این روند را به $t_{max} = 100$ محدود کرده‌ایم. در این الگوریتم، اختلاف بین دو تقریب $\pi(t-1)$ و $\pi(t)$ از بردار محبوبیت به صورت زیر محاسبه می‌شود:

$$D_{KL}(\pi(t)/n \parallel \pi(t-1)/n). \quad (۱۰)$$

۳-۲. مثالی از محاسبه بردار محبوبیت

به عنوان یک مثال ساده از محاسبه محبوبیت نمونه‌ها، مجموعه داده کوچک نشان داده شده در شکل ۱- (الف) را در نظر بگیرید. در شکل ۱- (ب) ماتریس توزیع اعتبار و محبوبیت نمونه‌ها، برای $k = 2$ و $t = 2$، نشان داده شده است. طبق تعریف ۱ مقادیر اعتبار داده شده به صفرمین، اولین و دومین نزدیک‌ترین همسایه برای همه نمونه‌ها به ترتیب برابر ۰,۰۷، ۰,۷۱ و ۰,۲۲ است. در شکل ۱-(ب)، وزن یالی که از x_i به x_j کشیده شده برابر مقدار اعتباری است که x_i به x_j می‌دهد. توجه داشته باشید که جمع محبوبیت‌های تمام نمونه‌ها برابر تعداد آنها یعنی ۹ است. به عنوان یک مثال، نمونه x_0 را در نظر بگیرید. x_0 به اولین و دومین نزدیک‌ترین همسایه خود یعنی x_7 و x_3 اعتبار می‌دهد. از آنجا که نمونه‌های دیگر هیچ اعتباری به x_0 نمی‌دهند، این نمونه کمترین محبوبیت را در این مجموعه داده دارد، یعنی داریم $\pi_0 = 0.01$. همچنین، نمونه x_3، به دلیل این که در مرکز این مجموعه داده قرار دارد بیشترین محبوبیت را دارد که برابر است با $\pi_3 = 1.99$.

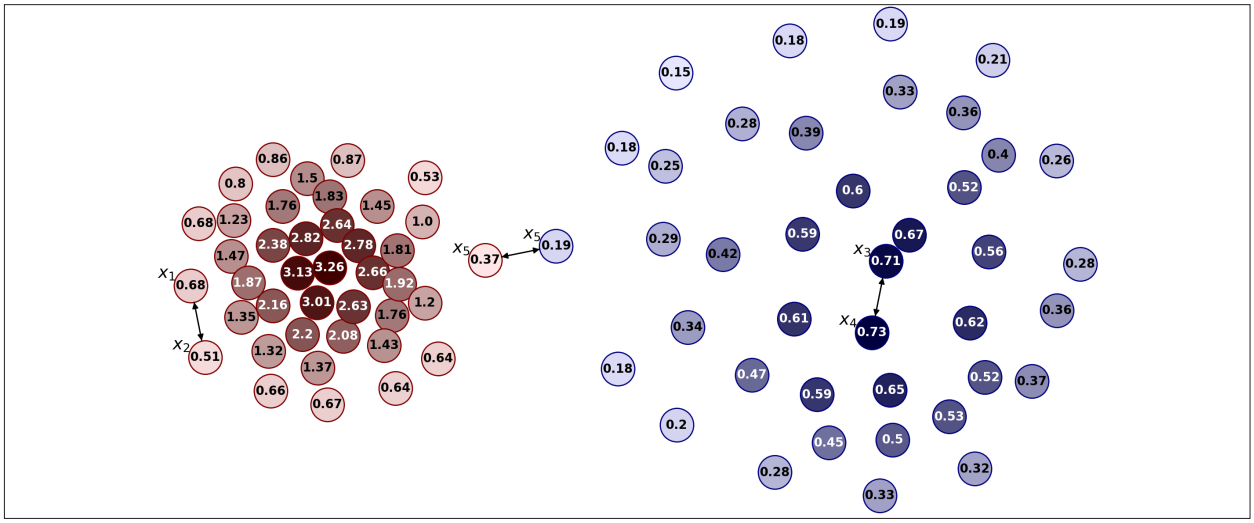
۳-۳. ساخت ماتریس شباهت مبتنی بر محبوبیت

تعریف معیار شباهت مبتنی بر محبوبیت پیشنهادی بر مبنای مفهوم «تمایل به ارتباط» استوار است:

تعریف ۳ (تمایل به ارتباط). فرض کنید x_i و x_j $(i \neq j)$، دو نمونه در X باشند، تمایل x_i برای ارتباط با x_j، که با $tendency(x_i, x_j)$ نشان داده می‌شود، به صورت زیر تعریف می‌شود:

$$tendency(x_i, x_j) = \pi_j \cdot c_{j,i} \quad (۱۱)$$

توجه کنید که تمایل x_i برای ارتباط با x_j، حاصل ضرب محبوبیت x_j در $c_{j,i}$، یعنی میزان اعتباری که x_j به x_i می‌دهد، است. رابطه ۱۱ ایده‌ای شبیه نحوه برقراری ارتباط افراد در



شکل (۲): مثالی از محاسبه شباهت بین سه جفت نقطه با فواصل یکسان، در محدوده‌های خلوت ($s(x_1,x_2)=0.051$)، در محدوده‌های پرتراکم ($s(x_3,x_4)=0.086$) و همچنین در خوشه‌های مختلف ($s(x_5,x_6)=0.025$)

یک جامعه دارد؛ به این معنی که در دنیای واقعی نیز، افراد معمولاً تمایل به ارتباط با افرادی دارند که هم محبوب باشند و هم آنها را به عنوان دوست نزدیک بپذیرند.

تعریف ٤ (شباهت مبتنی بر محبوبیت). فرض کنید که x_i و x_j، دو نمونه متفاوت از مجموعه داده X باشند، شباهت مبتنی بر محبوبیت x_i و x_j، که با $s(x_i,x_j)$ نشان داده می‌شود، به صورت زیر تعریف می‌شود:

$$s(x_i,x_j)=\min(tendency(x_i,x_j),\ tendency(x_j,x_i)). \quad (۱۲)$$

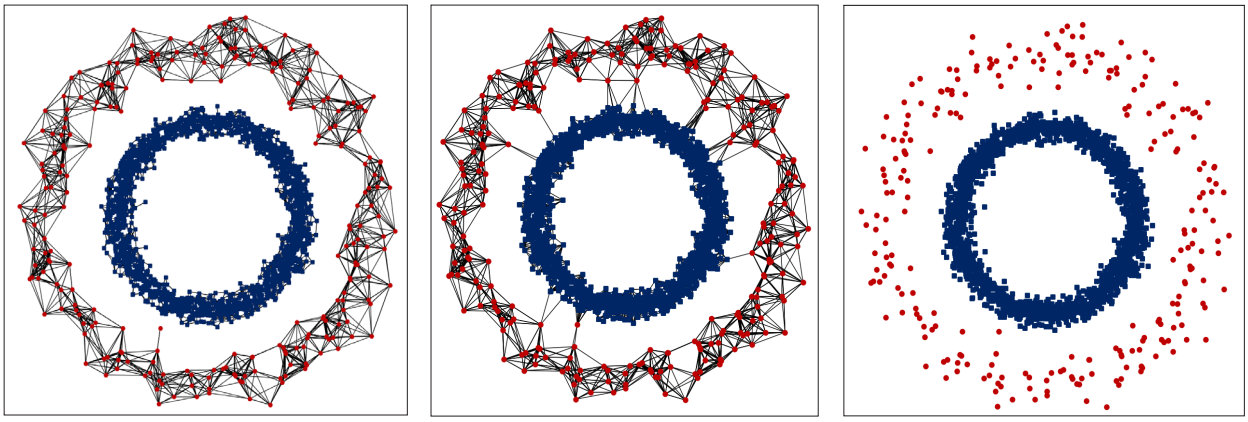
۴-۳- ویژگی‌های معیار شباهت مبتنی بر محبوبیت

معیار شباهت مبتنی بر محبوبیت پیشنهادی دارای چند ویژگی مفید است. اول این که، محبوبیت نمونه‌ها مستقل از مقیاس داده و چگالی محلی نمونه‌ها است. این ویژگی کمک بسیاری به تشخیص خوشه‌های با سطوح مختلف چگالی می‌کند. ویژگی دیگر این است که در معیار شباهت پیشنهادی، شباهت بین نمونه‌ها در محدوده‌های با چگالی و محبوبیت پایین (مثل حاشیه داده‌ها) کمتر از شباهت نمونه‌ها در محدوده‌های با چگالی و محبوبیت بالا (مثل مرکز داده‌ها) است. همچنین، معیار شباهت پیشنهادی به نقاط نزدیک به هم، در صورتی که هم‌خوشه باشند میزان شباهت بالا و در صورتی که در خوشه‌های متفاوت باشند میان شباهت پایین نسبت می‌دهد. این دو ویژگی به جداسازی خوشه‌های دارای هم‌پوشانی و با تفکیک ضعیف کمک بسیاری می‌کند. دیگر ویژگی معیار شباهت پیشنهادی پیچیدگی محاسباتی بسیار پایین آن است. به عنوان مثال، مجموعه داده شکل ۲ را در نظر بگیرید. این مجموعه داده شامل دو خوشه با چگالی متفاوت است. تیرگی رنگ هر نمونه در این شکل بیانگر میزان محبوبیت آن است. همچنین، میزان محبوبیت نمونه‌ها به صورت عددی نیز نشان داده شده است. اگر چگالی هر نقطه را برابر معکوس فاصله آن تا هفتمین نزدیک‌ترین همسایه‌اش در نظر بگیریم، چگالی نقاط در خوشه سمت چپ ۷٫۱ برابر چگالی نقاط در خوشه سمت راست است. با این وجود، میانگین محبوبیت نقاط دو خوشه (برای $k=3$ و $t=5$) تقریباً برابر است.

اکنون، سه جفت نقطه (x_1,x_2)، (x_3,x_4) و (x_5,x_6) را در این شکل در نظر بگیرید. فاصله بین هر سه جفت نقطه دقیقاً برابر است، یعنی:

$$d(x_1,x_2)=d(x_3,x_4)=d(x_5,x_6) \quad (۱۳)$$

جفت x_1 و x_2 را در محدوده کم چگالی لبه خوشه سمت چپ در نظر بگیرید. شباهت این دو نمونه با توجه به معیار شباهت مبتنی بر محبوبیت پیشنهادی برابر ۰٫۰۵۱ است. با این حال، شباهت بین جفت x_3 و x_4، با همان فاصله در



(الف) (ب) (ج)

شکل (۳): داده‌ای شامل دو خوشه دایره‌ای با تفکیک ضعیف و پراکندگی‌های بسیار متفاوت (قسمت الف)، گراف ده نزدیک‌ترین همسایه این داده بر اساس معیار شباهت سنتی (فاصله اقلیدسی) (قسمت ب)، و معیار شباهت مبتنی بر محبوبیت (قسمت ج)

محدوده‌ای پرتراکم (محبوبیت بالا)، به دلیل محبوبیت‌های بالای x_3 و x_4 بالاتر و برابر ۰,۰۸۶ است. اکنون، نمونه‌های x_5 و x_6 را با همان فاصله از دو خوشه مختلف در محدوده با محبوبیت پایین بین دو خوشه در نظر بگیرید. شباهت بین این دو نمونه بر اساس معیار شباهت مبتنی بر محبوبیت پیشنهادی برابر ۰,۰۲۵ است که به دلیل هم‌خوشه نبودن این دو نمونه، نسبت به شباهت دو نمونه هم‌خوشه x_1 و x_2 به طور قابل توجهی پایین‌تر است.

معیار شباهت مبتنی بر محبوبیت پیشنهادی ویژگی سودمند دیگری نیز دارد که به جدا کردن خوشه‌های با تفکیک ضعیف کمک می‌کند. به عنوان مثال، مجموعه داده شکل ۳- (الف) را، که شامل دو دایره هم مرکز با تفکیک ضعیف و پراکندگی‌های متفاوت است، در نظر بگیرید. گراف ده نزدیک‌ترین همسایه این داده، با توجه به معیار شباهت سنتی (فاصله اقلیدسی)، در شکل ۳-(ب) نشان داده شده است. دو خوشه در این شکل از هم تفکیک نشده‌اند. شکل ۳-(ج) گراف ده نزدیک‌ترین همسایه همان داده را بر اساس معیار شباهت مبتنی بر محبوبیت پیشنهادی نشان می‌دهد. در این شکل، شباهت نمونه‌ها بر اساس تعریف ٤ به ازای $k = 20$ محاسبه شده است. هر نمونه به ده نزدیک‌ترین همسایه‌اش متصل است مشروط به این که شباهت دو نمونه صفر نباشد. در این گراف دو خوشه از هم تفکیک شده‌اند. دلیل این موضوع این است که معیار پیشنهادی به نقاط نزدیک به هم، در صورتی که هم‌خوشه باشند میزان شباهت بالا و در صورتی که در خوشه‌های متفاوت باشند میان شباهت پایینی نسبت می‌دهد.

۵-۳. الگوریتم پیشنهادی

الگوریتم خوشه‌بندی طیفی مبتنی بر محبوبیت پیشنهادی، PBSC، شامل چند مرحله است. شبه کد این روش در الگوریتم ۱ آورده شده است. در اولین مرحله از این الگوریتم، در سطرهای اول تا دوازدهم، ماتریس توزیع اعتبار، C، و بردار محبوبیت نمونه‌ها، π، محاسبه می‌شود. سپس، در مرحله دوم، در سطرهای سیزده تا ۲۵، با توجه به ماتریس توزیع اعتبار و بردار محبوبیت، ماتریس شباهت، S، که نحوه محاسبه آن در تعریف ٤ آورده شد محاسبه می‌شود. در نهایت، در سومین مرحله، در سطرهای ۲٦ تا ۳۰، نتیجه خوشه‌بندی با استفاده از روند رایج در الگوریتم‌های خوشه‌بندی طیفی [۱۰] حاصل می‌شود. این روند به طور خلاصه در ادامه شرح داده شده است. در ابتدا L^{sym}، ماتریس لاپلاسین نرمال‌شده به صورت زیر محاسبه می‌شود:

$$L^{sym} = I - D^{-1/2} S D^{-1/2} \quad (۱٤)$$

که در آن، S ماتریس شباهت مبتنی بر محبوبیت و D یک ماتریس قطری، با درجه‌های d_1، d_2، ...، d_n بر روی قطر اصلی آن، است که ماتریس درجه نامیده می‌شود. در این

ماتریس d_i، درجه رأس x_i برابر است با $d_i = \sum_{j=1}^{n} s_{i,j}$. پس از محاسبه L^{sym}، نمونه‌ها بر اساس c بردار ویژه اول [۱۰] ماتریس L^{sym}، و با استفاده از الگوریتم کلاسیک k-$means$، به خوشه‌های C_1، C_2، ... C_c خوشه‌بندی می‌شوند که c بیانگر تعداد خوشه‌ها است.

۳-۶. انتخاب مقدار مناسب برای پارامتر k

پارامتر k، به عنوان تنها پارامتر الگوریتم خوشه‌بندی پیشنهادی، اندازه همسایگی نقاط را کنترل می‌کند. در ادامه، روشی برای تخمین مقدار مناسب، k^*، برای k پیشنهاد می‌دهیم. کار را با $k = 1$ شروع می‌کنیم و بردار محبوبیت را برای مقادیر متوالی k محاسبه می‌کنیم و تا جایی پیش می‌رویم که تفاوت بین بردارهای محبوبیت برای دو مقدار k متوالی، که با $\pi^{(k-1)}$ و $\pi^{(k)}$ نشان داده می‌شوند، به اندازه کافی کوچک شود. یعنی داشته باشیم:

$$D_{KL}\left(\pi'^{(k)} \parallel \pi'^{(k-1)}\right) \leq \varepsilon = 10^{-9}, \qquad (۱۵)$$

که در آن $\pi'^{(k-1)}$ و $\pi'^{(k)}$ به ترتیب بردار محبوبیت نرمال‌شده برای k و $k-1$ هستند که در آنها جمع مقادیر محبوبیت نرمال‌شده برابر با یک است.

۳-۷. پیچیدگی زمانی محاسبه ماتریس شباهت

با توجه به سادگی ساختار ماتریس توزیع اعتبار، پیچیدگی محاسباتی الگوریتم محاسبه ماتریس شباهت پیشنهادی برابر $O(n \cdot log\, n)$ و n بیانگر تعداد نمونه‌ها

Algorithm 1. Popularity Based Spectral Clustering (PBSC)
Input: Set $X = \{x_1, \dots, x_n\}$ of samples and neighborhood size, k
Output: The set of clusters

Step 1. Compute the credit matrix, $\mathrm{C} = [c_{i,j}]_{n\times n}$, and the popularity vector, $\pi = (\pi_1, \dots, \pi_n)$

1. for each pair of samples x_i and x_j $(i \leq j)$ do
2. compute $c_{i,j}$ and $c_{j,i}$ according to Definition 1
3. end for
4. $t = 0$
5. $\pi(0) = (1,1,\dots,1)$
6. *done* = false
7. while not *done* do
8. $t = t + 1$
9. $\pi(t) = \pi(t-1) \times C$
10. *done* = $(t = t_{max}) \vee D_{KL}(\pi(t)/n \parallel \pi(t-1)/n) \leq \varepsilon$
11. end while
12. $\pi = \pi(t)$

Step 2. Compute the popularity-based similarity matrix, S

13. for each pair of samples x_i and x_j $(i \leq j)$ do
14. if $d(x_i, x_j) = 0$ then
15. $s(x_i, x_j) = 1$
16. else
17. if $x_i \in kNN(x_j)$ and $x_j \in kNN(x_i)$ then
18. $tendency(x_i, x_j) = \pi_j \cdot c_{j,i}$
19. $tendency(x_j, x_i) = \pi_i \cdot c_{i,j}$
20. $s(x_i, x_j) = s(x_j, x_i) = \min\left(tendency\,(x_i, x_j), tendency(x_j, x_i)\right)$
21. else
22. $s(x_i, x_j) = s(x_j, x_i) = 0$
23. end if
24. end if
25. end for

Step 3. Compute the normalized Laplacian matrix L^{sym} and find the clusters

26. Compute the degree matrix D according to Section 3.5
27. $L^{sym} = I - D^{-1/2} S D^{-1/2}$
28. Compute the first c eigenvectors $v_1, \dots, v_c$ of L^{sym}
29. Cluster samples based on the computed eigenvectors
30. Return the set of clusters

است. این موضوع در ادامه به تفصیل شرح داده شده است.

پیچیدگی محاسباتی یافتن مجموعه $knn(\cdot)$ برای همه نقاط داده، در صورت استفاده از ساختار درخت k-d بسیار پایین و برابر $O(n \cdot log\ n)$ است [۲۸].

ماتریس توزیع اعتبار ساختار بسیار ساده‌ای دارد؛ هر سطر از این ماتریس تنها $k+1$ عنصر غیرصفر وجود دارد که مقادیر آن‌ها نیز برای همه سطرها یکسان است. بنابراین، این ماتریس $n \times n$ را می‌توان به صورت خلاصه و ضمنی در یک ماتریس $n \times k$ ذخیره کرد که عنصر i, r آن اندیس $nn_r(x_j)$ یعنی rامین نزدیک‌ترین همسایه نمونه x_i است. بنابراین، در محاسبه بردار محبوبیت با استفاده از روش تکرار توان، در هر تکرار، عملیات ضرب ماتریس اعتبار در بردار محبوبیت از مرتبه $O(n \cdot k)$ خواهد بود چرا که مقدار و موقعیت عناصر غیرصفر ماتریس اعتبار از پیش مشخص است. بنابراین، اگر m تکرار برای محاسبه بردار محبوبیت لازم باشد، محاسبه این بردار از مرتبه $O(n \cdot k \cdot m)$ خواهد بود و با توجه به این که $k \ll n$ و $m \ll n$، پیچیدگی زمانی این کار $O(n)$ خواهد بود. پس از محاسبه بردار محبوبیت، محاسبه ماتریس شباهت پیشنهادی با توجه به تعریف ٤ از مرتبه $O(n \cdot k)$ خواهد بود. در نتیجه، پیچیدگی محاسباتی الگوریتم محاسبه ماتریس شباهت مبتنی بر محبوبیت $O(n \cdot log\ n)$ خواهد بود.

برای مقایسه، پیچیدگی زمانی محاسبه معیارهای شباهت پیشنهاد شده در [۱۲، ۱۵ و ۱٦] را در ادامه شرح می‌دهیم. پیچیدگی زمانی محاسبه ماتریس شباهت پیشنهاد شده در [۱۲] برابر $O(n^2 \cdot k)$ است. در [۱۵] پیچیدگی محاسباتی یافتن مجموعه k نزدیک‌ترین همسایه برای تمام نمونه‌ها از مرتبه‌ای برابر $O(n \cdot \log n)$ و پیچیدگی زمانی محاسبه ماتریس شباهت برابر $O(n^2 \cdot k + n \cdot \log n)$ است. همچنین، در [۱٦] پیچیدگی محاسباتی یافتن همسایه‌های مشترک در گراف جهت‌دار k نزدیک‌ترین همسایه برابر $O(n^2 \cdot \log n + m \cdot k)$ است که در آن m بیانگر تعداد جفت نقاطی است که همسایه مشترک دارند و m به وضوح بسیار کوچکتر از n^2 است. بنابراین، پیچیدگی محاسباتی ساخت ماتریس وزن و ماتریس شباهت در [۱٦] برابر $O(n^2 + m)$ است. در نتیجه می‌توان گفت پیچیدگی محاسباتی روال محاسبه ماتریس شباهت مبتنی بر محبوبیت پیشنهادی، $O(n \cdot log\ n)$، در مقایسه با همتاهای آن ($O(n^2)$) کمتر است.

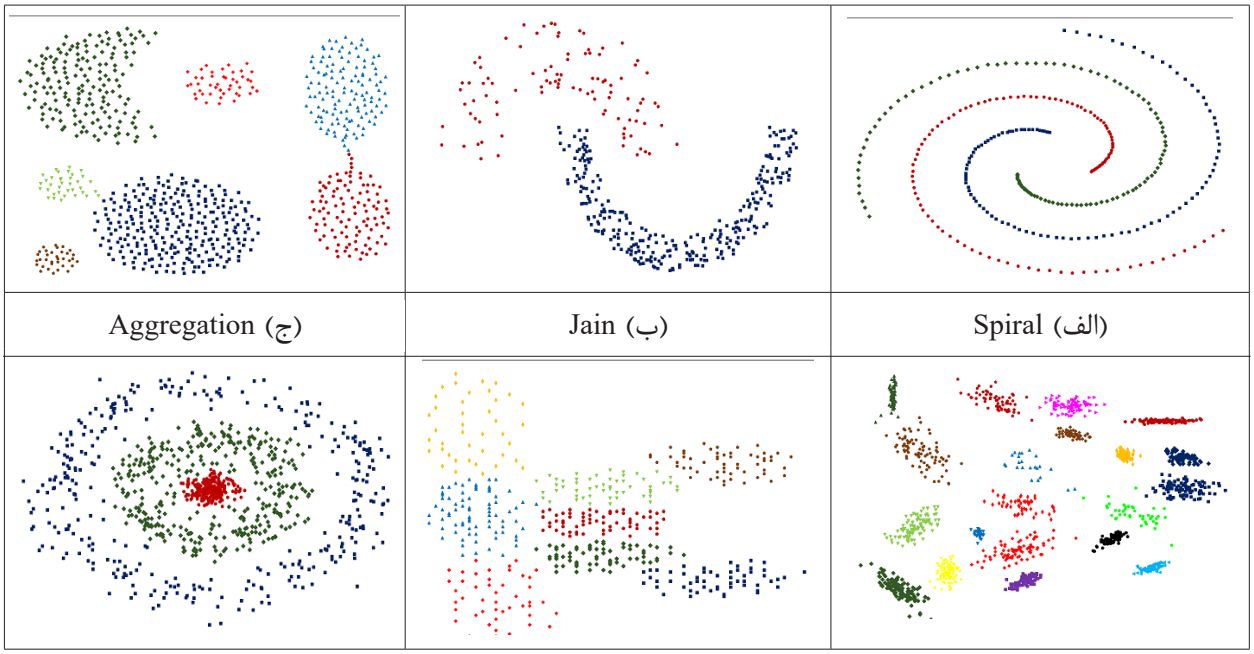
۴. ارزیابی و مقایسه

برای مطالعه کارایی الگوریتم خوشه‌بندی طیفی پیشنهادی، PBSC، ارزیابی‌هایی با استفاده از داده‌های مصنوعی و واقعی انجام می‌دهیم. برای توصیف میزان تطابق نتیجه الگوریتم‌های خوشه‌بندی با جواب صحیح از معیار انطباق NMI استفاده می‌کنیم. برای پیاده‌سازی الگوریتم خوشه‌بندی پیشنهادی از زبان برنامه‌نویسی پایتون و امکانات کتابخانه یادگیری ماشین Scikit-learn [۲۹] استفاده کرده‌ایم.

در ادامه، در ابتدا مطالعه کارایی الگوریتم پیشنهادی بر روی داده‌های مصنوعی و سپس مقایسه کارایی آن با برخی از مهم‌ترین الگوریتم‌های خوشه‌بندی موجود بر روی داده‌های واقعی انجام شده است.

۴-۱- داده‌های مصنوعی

ابتدا الگوریتم خوشه‌بندی طیفی مبتنی بر محبوبیت پیشنهادی، PBSC، را بر روی تعدادی داده مصنوعی شامل خوشه‌هایی با شکل‌ها، توزیع‌ها، چگالی‌ها و اندازه‌های گوناگون و دارای همپوشانی اعمال می‌کنیم. این داده‌ها عبارتند از Ds4c2sc8، 2d-20c-no0، Rings، Aggregation، Jain و Spiral [۳۰]. شکل ٤ خروجی الگوریتم پیشنهادی را برای این داده‌ها نشان می‌دهد. قسمت (الف) این شکل، خروجی الگوریتم خوشه‌بندی پیشنهادی را برای داده Spiral، که شامل سه خوشه مارپیچ است، نشان می‌دهد. این خروجی نشان می‌دهد که الگوریتم PBSC می‌تواند سه خوشه را به درستی تفکیک کند.



شکل ۴- خروجی الگوریتم خوشه‌بندی طیفی پیشنهادی PBSC بر روی داده‌های مصنوعی

برای ارزیابی الگوریتم خوشه‌بندی پیشنهادی در مواجهه با خوشه‌های با مقیاس‌های مختلف، در قسمت (ب) نتیجه الگوریتم PBSC را برای داده Jain نشان داده‌ایم. نتیجه نشان می‌دهد که الگوریتم پیشنهادی می‌تواند همه نقاط این داده را به درستی به خوشه‌ها منتسب کند. در قسمت (ج) نیز نتیجه الگوریتم PBSC برای داده Aggregation آورده شده است. الگوریتم خوشه‌بندی پیشنهادی بیش از ۹۹ درصد نمونه‌های این مجموعه داده را به درستی خوشه‌بندی می‌کند.

در قسمت‌های (د) و (ه) شکل ٤ خروجی الگوریتم خوشه‌بندی پیشنهادی به ترتیب برای داده‌های 2d-20c-no0 و Ds4c2sc8 آورده شده است. داده 2d-20c-no0 شامل خوشه‌هایی با پراکندگی‌ها و جهت‌ها مختلف است. الگوریتم پیشنهادی تقریباً همه نمونه‌های این داده را به درستی به خوشه‌ها منتسب می‌کند. همچنین، خروجی الگوریتم پیشنهادی برای مجموعه داده Ds4c2sc8، که شامل خوشه‌هایی با تفکیک بسیار ضعیف است، نشان می‌دهد که این الگوریتم بیش از ۹۷ درصد نمونه‌های این داده را به درستی به خوشه‌ها منتسب می‌کند. نهایتاً، خروجی الگوریتم پیشنهادی برای داده Rings، که شامل سه خوشه با همپوشانی بسیار بالا است، در قسمت (و) شکل ٤، نشان می‌دهد که این الگوریتم بیش از ۹٤ درصد نقاط را به درستی خوشه‌بندی می‌کند.

نتایج الگوریتم خوشه‌بندی طیفی پیشنهادی، PBSC، برای داده‌های مصنوعی نشان می‌دهد که این الگوریتم، در مواجهه با خوشه‌های با شکل‌ها، مقیاس‌ها، اندازه‌ها و توزیع‌های مختلف عملکرد بسیار موثری دارد.

۴-۲- داده‌های واقعی

برای بررسی عملکرد الگوریتم خوشه‌بندی PBSC، در محیط واقعی، الگوریتم پیشنهادی را بر روی پانزده مجموعه داده واقعی [۳۱] اعمال کرده‌ایم. مشخصات این داده‌ها، از جمله تعداد نمونه‌ها، ویژگی‌ها و خوشه‌ها، در جدول ۲ آورده شده است. ابعاد این داده‌ها از چهار تا ۳٤ و تعداد نمونه‌های آنها از ۱۵۰ تا ٦۳۲۱ متغیر است. برخی از این داده‌ها، شامل خوشه‌هایی با چگالی‌های

جدول ۲- مشخصات داده‌های واقعی مورد استفاده در ارزیابی

نام مجموعه داده	تعداد نمونه	تعداد ویژگی	تعداد خوشه
Dermatology	366	34	6
HCV	615	12	5
Iris	150	4	3
LiverDisorders	345	6	2
PLRX	182	12	2
Seeds	210	7	3
ShillBidding	6321	9	2
UserKnowledgeModeling	403	5	4
Vehicle	846	18	4
Vertebral2C	310	6	2
Vertebral3C	310	6	3
Vertebral3CW	310	6	2
WineQualityWhite	4898	11	7
WirelessIndoorLocalization	2000	7	4
Yeast	1484	8	10

بسیار متفاوت هستند. نسبت چگالی متراکم‌ترین خوشه به چگالی خلوت‌ترین خوشه در این داده‌ها از ۱,۱ برای داده LiverDisorders تا ۱۶,۲ برای داده WineQualityWhite متغیر است.

برای مقایسه کارایی، الگوریتم پیشنهادی PBSC را با نه الگوریتم خوشه‌بندی موجود، شامل مرتبط‌ترین الگوریتم‌ها از دسته‌های مختلف و بخصوص الگوریتم‌های جدید خوشه‌بندی طیفی، مقایسه کرده‌ایم. علاوه بر مقایسه با الگوریتم خوشه‌بندی مبتنی بر مرکز k-means [۳۲]، الگوریتم‌های مبتنی بر چگالی HDBSCAN [۳۳]، و RNN-DBSCAN [۳۴]، الگوریتم سلسله‌مراتبی Agglomerative [۴]، و الگوریتم بدون پارمتر FINCH [۳۵]، الگوریتم پیشنهادی را با الگوریتم‌های خوشه‌بندی طیفی زیر نیز مقایسه کرده‌ایم:

- الگوریتم خوشه‌بندی طیفی کلاسیک (SC-Classic) [۱۰]،
- الگوریتم خوشه‌بندی طیفی بر مبنای معیار شباهت مبتنی بر میانگین محلی (SC-LM) [۱۵]،
- الگوریتم خوشه‌بندی طیفی بر مبنای معیار شباهت خود-تنظیم (SC-ST) [۱۲] و نهایتاً
- الگوریتم خوشه‌بندی طیفی بر مبنای معیار شباهت محلی مبتنی بر همسایه‌های مشترک (SC-LSM) [۱۶].

در جدول ۳، مشخصات این الگوریتم‌ها از جمله نام پارامترهای آنها به همراه محدوده مقادیر هر پارامتر آورده شده است. برای هر کدام از داده‌های جدول ۲ هر کدام از الگوریتم‌های خوشه‌بندی را به ازای تمام مقادیر فضای پارامتری آنها اجرا کرده و بهترین نتیجه را از نظر شاخص NMI گزارش کرده‌ایم. همچنین، برای k، تنها پارامتر الگوریتم، PBSC، محدوده مقادیر زیر را در نظر گرفته‌ایم:

$$k \in \{1,3,5,\dots,15\} \cup \{20,25,30,\dots,50\} \quad (۱۶)$$

نتایج ارزیابی و مقایسه الگوریتم پیشنهادی، PBSC، با الگوریتم‌های خوشه‌بندی همتای آن، در قالب معیار NMI، در جدول ٤ آورده شده است. در مورد هر مجموعه داده، علاوه بر خروجی الگوریتم پیشنهادی اعداد بزرگتر یا مساوی آن نیز پررنگ شده‌اند. نتایج این ارزیابی و مقایسه نشان می‌دهد که برای هر کدام از این داده‌ها، الگوریتم پیشنهادی یا بهترین کارایی و یا دومین بهترین کارایی را دارد. در واقع برای سیزده مجموعه داده از پانزده مورد، الگوریتم پیشنهادی بهترین کارایی و برای دو مورد دیگر این الگوریتم دومین بهترین کارایی را دارد.

۵. نتیجه‌گیری

در این مقاله، الگوریتم خوشه‌بندی طیفی جدیدی بر مبنای یک معیار شباهت جدید مبتنی بر محبوبیت پیشنهاد شده است. در الگوریتم خوشه‌بندی پیشنهادی PBSC به هر نمونه یک شاخص محبوبیت نسبت داده می‌شود که مشخص می‌کند که آن نمونه تا چه حد در مرکز داده قرار دارد. سپس، بر مبنای محبوبیت نمونه‌ها، گراف

جدول ۳- مشخصات الگوریتم‌های خوشه‌بندی مورد استفاده در ارزیابی

نام الگوریتم	سال انتشار	پارامترها	مرجع
k-means	1967	type ∈ {regular, fcm, bisecting}, #clusters	[32]
HDBSCAN	2020	min_cluster_size = 5	[33]
RNN-DBSCAN	2018	k ∈ {3,6,9, ... ,60}	[34]
Agglomerative	2006	linkage ∈ {ward, complete, average, single}	[4]
FINCH	2019	req_clust = #clusters	[35]
SC-Classic	2001	assign_labels	[9]
SC-ST	2004	assign_labels	[12]
SC-LSM	2022	assign_labels	[16]
SC-LM	2022	assign_labels	[15]

جدول ۴- مقایسه الگوریتم پیشنهادی و الگوریتم‌های همتای آن بر روی مجموعه داده‌های جدول ۲ از منظر شاخص **NMI**

PBSC (Proposed)	SC-LM	SC-LSM	SC-ST	SC-Classic	FINCH	Agglomera-tive	RNN-DBSCAN	HDBSCAN	k-means	
92.	88.	91.	92.	86.	90.	86.	72.	45.	88.	Dermato
44.	26.	04.	02.	18.	04.	16.	21.	02.	06.	HCV
88.	85.	83.	81.	80.	72.	80.	70.	74.	76.	Iris
02.	01.	01.	01.	01.	00.	00.	02.	01.	00.	LiverDis
05.	00.	03.	00.	00.	00.	01.	04.	00.	00.	PLRX
78.	74.	78.	70.	69.	69.	71.	51.	38.	70.	Seeds
21.	09.	10.	00.	09.	00.	02.	19.	14.	00.	ShillBid
48.	38.	42.	34.	33.	19.	52.	20.	07.	44.	UserKno
24.	19.	21.	17.	13.	12.	15.	08.	02.	11.	Vehicle
18.	16.	15.	13.	15.	10.	16.	10.	08.	14.	Vert2C
29.	28.	30.	29.	27.	25.	27.	26.	16.	27.	Vert3C
18.	16.	15.	13.	15.	10.	16.	10.	08.	14.	Vert3CW
10.	10.	09.	08.	08.	08.	08.	07.	02.	09.	WineQW
92.	80.	92.	86.	72.	80.	88.	52.	05.	85.	WirelessIL
37.	36.	30.	31.	36.	31.	28.	07.	04.	31.	Yeast

شباهتی بر پایه یک معیار شباهت جدید ساخته می‌شود. معیار شباهت مبتنی بر محبوبیت پیشنهادی چند ویژگی سودمند دارد. اول این که، محبوبیت نمونه‌ها، مستقل از مقیاس داده و چگالی محلی نمونه‌ها است. این ویژگی سبب می‌شود که الگوریتم پیشنهادی در مواجهه با داده‌های با سطوح مختلف چگالی موفق عمل کند. دیگر ویژگی معیار شباهت پیشنهادی این است که در آن شباهت بین نمونه‌ها در محدوده‌های خلوت کمتر از شباهت در محدوده‌های پرتراکم است. همچنین، معیار شباهت پیشنهادی به نقاط نزدیک به هم که هم‌خوشه باشند میزان شباهت بالا و در صورتی که هم‌خوشه نباشند میزان شباهت پایینی نسبت می‌دهد. این دو ویژگی به جداسازی خوشه‌های دارای همپوشانی و با تفکیک ضعیف کمک بسیاری می‌کنند. دیگر ویژگی معیار شباهت پیشنهادی پیچیدگی محاسباتی بسیار پایین آن است.

نتایج ارزیابی‌های انجام شده بر روی داده‌های مصنوعی و واقعی نشان می‌دهد که الگوریتم پیشنهادی PBSC در مواجهه با خوشه‌های با شکل‌ها، مقیاس‌ها،

اندازه‌هـــا و توزیع‌های مختلف کارایی بهتری در مقایســـه با الگوریتم‌های خوشـــه‌بندی موجود دارد. در آینده، قصد داریم یـــک گونه سلســـله‌مراتبی از الگوریتم پیشـــنهادی ارایـــه دهیم که در آن فاصله خوشـــه‌ها با توجه به میزان چسبیدگی[۲۷] بین آنها مشخص می‌شود.

۶. مراجع

[1] Xudong Tang, Chao Dong, Wei Zhang: Contrastive author-aware text clustering. Pattern Recognit. 130: 108787 (2022)

[2] Junpeng Tan, Zhijing Yang, Yongqiang Cheng, Jielin Ye, Bing Wang, Qingyun Dai: SRAGL-AWCL: A two-step multi-view clustering via sparse representation and adaptive weighted cooperative learning. Pattern Recognit. 117: 107987 (2021)

[3] Hang Zhang, Haili Li, Ning Chen, Shengfeng Chen, Jian Liu: Novel fuzzy clustering algorithm with variable multi-pixel fitting spatial information for image segmentation. Pattern Recognit. 121: 108201 (2022)

[4] Guillaume Guenard, Pierre Legendre: Hierarchical Clustering with Contiguity Constraint in R. J. Stat. Softw. 103(7) (2022)

[5] Chunrong Wu, Qinglan Peng, Jia Lee, Kenji Leibnitz, Yunni Xia: Effective hierarchical clustering based on structural similarities in nearest neighbor graphs. Knowl. Based Syst. 228: 107295 (2021)

[6] Jianbo Shi, Jitendra Malik: Normalized Cuts and Image Segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 22(8): 888-905 (2000)

[7] Abhishek Kumar, Hal Daume: A Co-training Approach for Multi-view Spectral Clustering. ICML 2011: 393-400

[8] Jiexing Liu, Chenggui Zhao: Density Gain-Rate Peaks for Spectral Clustering. IEEE Access 9: 46000-46010 (2021)

[9] Ng A. Y, Jordan M. I, Weiss Y, On spectral clustering: Analysis and an algorithm, NIPS'01: Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic, 849-856 (2001).

[10] Von Luxburg U, A tutorial on spectral clustering, Statistics and Computing, 17(4), 395-416 (2007).

[11] Yessica Nataliani, Miin-Shen Yang: Powered Gaussian kernel spectral clustering. Neural Comput. Appl. 31(S-1): 557-572 (2019)

[12] Zelnik-Manor L, Perona P, Self-tuning spectral clustering, Neural Information Processing Systems (NIPS 2004) Vancouver, British Columbia, Canada, 1601-1608 (2004)

[13] Tong Liu, Jingting Zhu, Jukai Zhou, Yongxin Zhu, Xiaofeng Zhu: Initialization-similarity clustering algorithm. Multim. Tools Appl. 78(23): 33279-33296 (2019)

[14] Xianchao Zhang, Jingwei Li, Hong Yu: Local density adaptive similarity measurement for spectral clustering. Pattern Recognit. Lett. 32(2): 352-358 (2011)

[15] Hassan Motallebi, Rabeeh Nasihatkon, Mina Jamshidi: A Local Mean-based Distance Measure for Spectral Clustering. Pattern Analysis \& Applications 25(2), 351-359 (2022)

[16] Zongqi Cao, Hongjia Chen, Xiang Wang: Spectral clustering based on the local similarity measure of shared neighbors, ETRI Journal, 2022, 44(5): 769-779

[17] Paola Favati, Grazia Lotti, Ornella Menchi, Francesco Romani: Construction of the similarity matrix for the spectral clustering method: Numerical experiments. J. Comput. Appl. Math. 375: 112795 (2020)

[18] Malgorzata Lucinska, Slawomir T. Wierzchon: Spectral Clustering Based on k-Nearest Neighbor Graph. CISIM 2012: 254-265

[19] Tan M, Zhang S, Wu L, Mutual KNN based spectral clustering, Neural computing and applications, 32, 6435-6442 (2020)

[20] Mashaan A. Alshammari, John Stavrakakis, Masahiro Takatsuka: Refining a k-nearest neighbor graph for a computationally efficient spectral clustering. Pattern Recognit. 114: 107869 (2021)

[21] Yongda Cai, Joshua Zhexue Huang, Jianfei Yin: A new method to build the adaptive k-nearest neighbors similarity graph matrix for spectral clustering. Neurocomputing 493: 191-203 (2022)

[22] Xiucai Ye, Tetsuya Sakurai: Spectral clustering using robust similarity measure based on closeness of shared Nearest Neighbors. IJCNN 2015: 1-8

[23] Tulin Inkaya: A parameter-free similarity graph for spectral clustering. Expert Syst. Appl. 42(24): 9489-9498 (2015)

[24] F. Nie, X. Wang, H. Huang, Clustering and projected clustering with adaptive neighbors, in: Proceedings of the 20th ACM SIGKDD International Conference
on Knowledge Discovery and Data Mining, 2014, pp. 977–986.

[25] Z. Bian, H. Ishibuchi, S. Wang, Joint learning of spectral clustering structure and fuzzy similarity matrix of data, IEEE Trans. Fuzzy Syst. 27 (1) (2018) 31–44

[26] K.K. Sharma, A. Seal, Spectral embedded generalized mean based k-nearest neighbors clustering with s-distance, Expert Syst. Appl. 169 (2021) 114326.

[27] Ufuk Bahceci: New bounds for the empirical robust Kullback-Leibler divergence problem. Inf. Sci. 637: 118972 (2023)

[28] Jon Louis Bentley: Multidimensional Binary Search Trees Used for Associative Searching. Commun. ACM 18(9): 509-517 (1975)

[29] F. Pedregosa, et al. Scikit-learn: Machine learning in Python, J. Mach. Learn. Res., vol. 12, pp. 2825-2830, 2011.

[30] P. Fanti, and S. Sieranoja, "K-Means Properties on Six Clustering Benchmark Datasets," Appl. Intell., vol. 48, no. 12, pp. 4743-4759, 2018.

[31] D. Dua, and C. Graff, "UCI Machine Learning Repository," http://archive.ics.uci.edu/ml, 2019.

[32] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," Proc. 5th Berkeley Symp. Math. Statist. Prob., pp. 281–297, 1967.

[33] C. Malzer, and M. Baum, "A Hybrid Approach to Hierarchical Density-based Cluster Selection," Proc. IEEE Int. Conf. Multisens. Fusion Integr. Intell. Syst., pp. 223-228, 2020.

[34] A. Bryant, and K. J. Cios, "RNN-DBSCAN: A Density-Based Clustering Algorithm Using Reverse Nearest Neighbor Density Estimates," IEEE Trans. Knowl. Data Eng., vol. 30, no. 6, pp. 1109-1121, 2018.

[35] M. S. Sarfraz, V. Sharma, and R. Stiefelhagen, "Efficient Parameter-Free Clustering Using First Neighbor Relations," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 8934-8943, 2019.

27-Cohesion