DOR: https://csj.isi.org.ir/article_256402.html

DOI: https://doi.org/10.22034/csj.2026.585989.1157

پاییز ۱۴۰۵، دوره ۱۱، شماره ۳، پیاپی ۴۲

دریافت مقاله: ۱۴۰۵/۰۳/۲۵

پذیرش مقاله: ۱۴۰۵/۰۵/۱۰

نوع مقاله: پژوهشی

الگوریتم k–میانگین++ مبتنی بر فاصله شواهد و تحلیل پایداری در حضور نوفه و داده‌های پرت

وحیده احراری*

گروه علوم کامپیوتر، دانشکده علوم ریاضی، دانشگاه شهرکرد، شهرکرد، ایران

پست الکترونیکی: vahideh.ahrari@sku.ac.ir

زهرا بهدانی

گروه آمارو ریاضی، دانشکده انرژی و علوم داده، دانشگاه صنعتی خاتم الانبیاء بهبهان، بهبهان، ایران

پست الکترونیکی: behdani@bkatu.ac.ir

چکیده

در این مقاله، الگوریتم خوشه‌بندی شواهد–محور ED-K-means++ ارائه می‌شود که ترکیبی از مقداردهی اولیه K-means++ و فاصله شواهد ژوسلم در چارچوب نظریه دمپستر– شیفر است. هدف اصلی کاهش حساسیت به مراکز اولیه و بهبود پایداری در حضور نوفه و داده‌های پرت است. هر نمونه به تابع تخصیص احتمال پایه تبدیل شده و مراکز اولیه بر اساس مربع فاصله شواهد انتخاب می‌شوند. عملکرد روش پیشنهادی روی سه مجموعه‌داده استاندارد Wine، Iris و Breast Cancer در ۳۰ اجرای مستقل و در سه حالت بدون اغتشاش، نوفهٔ گاوسی و داده‌های پرت ارزیابی شد. نتایج نشان داد که اگرچه روش پیشنهادی لزوماً موجب افزایش شاخص بیرونی ARI در همه مجموعه‌داده‌ها نمی‌شود، اما نسبت به ED-K-means رفتار پایدارتر و ساختار درونی منسجم‌تری در فضای شواهد ایجاد می‌کند. در حالت بدون اغتشاش، ضریب سیلوئت برای Wine به ۰/۹۵۸ و برای Breast Cancer به ۰/۹۷۲ رسید، در حالی که در حالت داده‌های پرت (ρ=0.20) این شاخص برای Breast Cancer تا ۰/۹۸٤ نیز افزایش یافت. شاخص دیویس–بولدین نیز در حالت بدون اغتشاش برای Breast Cancer تا ۰/۳۲۹ کاهش یافت که بیانگر افزایش فشردگی درون‌خوشه‌ای و جدایی بهتر خوشه‌ها است. در تحلیل حساسیت پارامترهای ساخت BPA نیز مقدار مشابهی (۰/۳۹٤) برای این مجموعه‌داده به‌دست آمد. ارزیابی در سطوح مختلف نوفه و داده‌های پرت نشان داد، مقداردهی اولیه مبتنی بر فاصله شواهد از اجراهای ناپایدار جلوگیری می‌کند. در مجموع، نتایج نشان می‌دهد روش پیشنهادی با پایدارسازی مقداردهی اولیه در فضای شواهد، موجب بهبود انسجام درونی خوشه‌ها و کاهش حساسیت الگوریتم به شرایط نوفه‌ای، بدون افزایش پیچیدگی محاسباتی نسبت به روش پایه، می‌شود.

کلید واژه‌ها: خوشه‌بندی، K-means++، نظریه دمپستر– شیفر، فاصله ژوسلم، پایداری، داده‌های نوفه‌دار.

۱– مقدمه

خوشه‌بندی داده‌ها یکی از مسائل بنیادی در یادگیری ماشین است که هدف آن، تقسیم مجموعه‌داده به گروه‌هایی

* نویسندهٔ مسئول

با بیشترین شباهت درون‌خوشه‌ای و کمترین شباهت بین خوشه‌ها است. در میان روش‌های مختلف، الگوریتم K-means به دلیل سادگی، سرعت و سهولت پیاده‌سازی، از پرکاربردترین روش‌های خوشه‌بندی به شمار می‌رود [۳،٤]. با وجود این، دو محدودیت اصلی در این الگوریتم وجود دارد. نخست، استفاده از فاصله اقلیدسی که برای داده‌های دارای نوفه، ابهام یا عدم قطعیت همواره معیار مناسبی برای اندازه‌گیری شباهت نیست. دوم، وابستگی عملکرد الگوریتم به انتخاب مراکز اولیه خوشه، که می‌تواند به همگرایی به جواب‌های محلی و کاهش پایداری نتایج منجر شود [٤]. برای کاهش این مشکل، الگوریتم K-means++ با انتخاب هدایت‌شده مراکز اولیه، کیفیت مقداردهی اولیه و در نتیجه عملکرد خوشه‌بندی را بهبود می‌دهد [٥].

در بسیاری از کاربردهای واقعی، داده‌ها با نوفه، اطلاعات ناقص یا عدم قطعیت همراه هستند و نمایش آنها به‌صورت نقاط قطعی، بخشی از اطلاعات موجود را نادیده می‌گیرد. نظریه دمپستر- شیفر چارچوبی برای نمایش و ترکیب چنین اطلاعاتی فراهم می‌کند و امکان مدل‌سازی صریح عدم قطعیت را در تحلیل داده‌ها فراهم می‌سازد [٦،٧]. در این چارچوب، فاصله ژوسلم به‌عنوان معیاری برای اندازه‌گیری اختلاف میان بدنه‌های شواهد به‌کار می‌رود [٨]. با وجود پیشرفت‌های حاصل شده در خوشه‌بندی شواهدمحور، انتخاب مراکز اولیه همچنان یکی از عوامل مؤثر بر کیفیت و پایداری نتایج است. از این‌رو، طراحی یک راهکار مقداردهی اولیه مناسب برای این دسته از الگوریتم‌ها همچنان یک مسئله باز به شمار می‌رود. مرور پژوهش‌های پیشین و تحلیل این موضوع در بخش «کارهای مرتبط» ارائه شده است.

نوآوری این پژوهش، به‌کارگیری K-means++ برای مقداردهی اولیه در خوشه‌بندی شواهدمحور مبتنی بر نظریه دمپستر- شیفر است. در روش پیشنهادی، نمونه‌های اولیه بر اساس مربع فاصله ژوسلم انتخاب می‌شوند. انتظار می‌رود این شیوه با کاهش وابستگی الگوریتم به انتخاب تصادفی مراکز اولیه، پایداری نتایج را در اجراهای مختلف افزایش دهد و انسجام خوشه‌ها را، به‌ویژه در حضور نوفه و داده‌های پرت، بهبود بخشد.

اگرچه نظریه دمپستر- شیفر، فاصله ژوسلم و الگوریتم ED-K-means پیش‌تر معرفی شده‌اند، در روش ED-K-means مراکز اولیه خوشه همچنان به‌صورت تصادفی انتخاب می‌شوند. این موضوع می‌تواند باعث شود نتایج خوشه‌بندی در اجراهای مختلف نوسان داشته باشند. در این پژوهش، با انتقال ایده مقداردهی اولیه K-means++ به فضای شواهد و استفاده از فاصله ژوسلم در انتخاب مراکز اولیه، این محدودیت تا حد زیادی کاهش یافته است. بنابراین، نوآوری این مقاله در ارائه یک نظریه یا معیار فاصله جدید نیست، بلکه در به‌کارگیری و سازگار کردن یک روش موفق مقداردهی اولیه برای خوشه‌بندی شواهدمحور است. بر اساس بررسی منابع موجود، چنین رویکردی تاکنون برای الگوریتم‌های خوشه‌بندی مبتنی بر فاصله ژوسلم گزارش نشده است.

برای ارزیابی عملکرد روش پیشنهادی، آزمایش‌ها روی سه مجموعه‌داده استاندارد Wine، Iris و Breast Cancer و در سه حالت بدون اغتشاش، نوفهٔ گاوسی و آلودگی با داده‌های پرت انجام شد. علاوه بر مقایسه با الگوریتم‌های K-means++، K-means و ED-K-means، اثر پارامترهای ساخت تابع تخصیص احتمال پایه (BPA) نیز از طریق تحلیل حساسیت بررسی شد. نتایج نشان داد که اگرچه روش پیشنهادی لزوماً موجب بهبود شاخص ARI در همه مجموعه داده‌ها نمی‌شود، اما با پایدارسازی فرایند مقداردهی اولیه در فضای شواهد، خوشه‌هایی با انسجام درونی بیشتر و نوسان کمتر ایجاد می‌کند. این یافته نشان می‌دهد که مزیت اصلی روش پیشنهادی، نه در افزایش مطلق دقت خوشه‌بندی، بلکه در افزایش پایداری فرایند خوشه‌بندی و بهبود کیفیت ساختار درونی خوشه‌ها در فضای شواهد، به‌ویژه در حضور نوفه و داده‌های پرت، است.

ادامه مقاله به این صورت است که ابتدا مفاهیم و تعاریف موردنیاز مرور می‌شود، سپس الگوریتم پیشنهادی

و شبه‌کد آن ارائه می‌گردد. در بخش آزمایش‌ها، تنظیمات، سطوح مختلف نوفه و معیارهای ارزیابی معرفی شده و نتایج همراه با تحلیل ارائه می‌شود. در پایان نیز نتیجه‌گیری و پیشنهادهایی برای پژوهش‌های آتی مطرح خواهد شد.

۲- کارهای مرتبط

نظریه دمپستر- شیفر یکی از چارچوب‌های شناخته‌شده برای مدل‌سازی عدم قطعیت است و به همین دلیل، در سال‌های اخیر کاربرد گسترده‌ای در خوشه‌بندی داده‌های نوفه‌دار، ناقص و مبهم پیدا کرده است [۶،۷]. در این حوزه، فاصله ژوسلم به‌عنوان معیاری برای اندازه‌گیری اختلاف میان بدنه‌های شواهد معرفی شد و به‌تدریج به یکی از معیارهای پرکاربرد در روش‌های خوشه‌بندی شواهدمحور تبدیل شد [۸].

ژو و همکاران [۹] با جایگزین کردن فاصله اقلیدسی با فاصله ژوسلم، الگوریتم ED-K-means را پیشنهاد کردند. در این روش، استفاده از فاصله شواهد باعث شد عملکرد خوشه‌بندی برای داده‌های دارای عدم قطعیت نسبت به الگوریتم کلاسیک K-means بهبود یابد. با این حال، مراکز اولیه خوشه‌ها همچنان به‌صورت تصادفی انتخاب می‌شوند. بنابراین، حساسیت الگوریتم به مقداردهی اولیه و احتمال همگرایی به جواب‌های محلی همچنان در فضای شواهد وجود دارد

از سوی دیگر، آرتور و واسیلویتسکی [۵] با معرفی الگوریتم K-means++ نشان دادند که اگر مراکز اولیه به‌صورت هدفمند و بر اساس فاصله نمونه‌ها انتخاب شوند، کیفیت خوشه‌بندی، پایداری نتایج و سرعت همگرایی بهبود می‌یابد. با این حال، این ایده تاکنون برای الگوریتم‌های خوشه‌بندی مبتنی بر نظریه شواهد با استفاده از فاصله ژوسلم به‌کار گرفته نشده است.

پژوهش‌های جدیدتر در حوزه خوشه‌بندی شواهد محور، بیشتر بر افزایش تفسیرپذیری روش‌های خوشه‌بندی [۱۰،۱۱] و بهبود عملکرد آنها در داده‌های نوفه‌دار و محیط‌های نیمه‌نظارتی [۱۲] تمرکز داشته‌اند. با وجود پیشرفت‌های حاصل‌شده، مسئله انتخاب پایدار و هدایت‌شده مراکز اولیه خوشه در این روش‌ها همچنان بدون راهکار مشخص باقی مانده است.

بر اساس بررسی منابع موجود، تاکنون روشی برای به‌کارگیری ایده K-means++ در الگوریتم‌های خوشه‌بندی شواهدمحور مبتنی بر فاصله ژوسلم گزارش نشده است. به همین دلیل، در این پژوهش ایده K-means++ به فضای شواهد تعمیم داده شده و الگوریتم ED-K-means++ پیشنهاد می‌شود. در این روش، مراکز اولیه خوشه با استفاده از فاصله ژوسلم و تابع جرم شواهد انتخاب می‌شوند تا وابستگی الگوریتم به مقداردهی اولیه کاهش یابد و پایداری و کیفیت خوشه‌بندی بهبود پیدا کند.

۳- مبانی نظری و مباحث مقدماتی

در این بخش، ابتدا الگوریتم K-means و روش مقداردهی اولیه K-means++ معرفی می‌شوند. سپس مفاهیم بنیادین نظریه دمپستر- شیفر تشریح شده و در ادامه، فاصله شواهد به‌عنوان معیاری برای سنجش میزان شباهت یا عدم‌شباهت میان بدنه‌های شواهد توضیح داده می‌شود. در پایان، نحوه به‌کارگیری این فاصله در خوشه‌بندی شواهدمحور، به‌عنوان مبنای روش پیشنهادی این پژوهش، ارائه خواهد شد.

۳-۱- الگوریتم K-means

فرض کنید مجموعه‌داده‌ای شامل n مشاهده در فضای d-بعدی به‌صورت:

$$X = \{x_1, x_2, \dots, x_n\} \subset \mathbb{R}^d$$

در اختیار باشد. مسئله k-میانگین با هدف افراز این مجموعه به k خوشه مجزا $C_1, \dots, C_k$، به‌گونه‌ای تعریف می‌شود که پراکندگی درون‌خوشه‌ای (بر حسب فاصله اقلیدسی) با کمینه‌سازی مجموع مربعات فاصله نقاط از مراکز خوشه‌ها کاهش یابد [۳, ٤]:

$$J(U,C) = \sum_{j=1}^{k}\sum_{i=1}^{n} u_{ij} \|x_i - c_j\|_2^2, \qquad (1)$$

که در آن $c_j \in \mathbb{R}^d$ مرکز خوشه C_j و $U = [u_{ij}]$ ماتریس تخصیص با درایه‌های دودویی $u_{ij} \in \{0,1\}$ است، به‌طوری که برای هر i شرط $\sum_{j=1}^{k} u_{ij} = 1$ برقرار بوده و هر مشاهده دقیقاً به یک خوشه نسبت داده می‌شود.

الگوریتم k–میانگین را می‌توان یک فرایند کمینه‌سازی متناوب در نظر گرفت. با ثابت فرض کردن مراکز، تخصیص به‌صورت انتخاب نزدیک‌ترین مرکز بر حسب فاصله اقلیدسی به‌دست می‌آید:

$$u_{ij} = \begin{cases} 1, & j = \arg\min_{\ell\in\{1,\dots,k\}} \|x_i - c_\ell\|_2^2, \\ 0, & \text{در غیر این صورت.} \end{cases}$$

و با ثابت فرض کردن تخصیص‌ها، مقدار بهینه هر مرکز برابر میانگین نقاط منتسب به همان خوشه است:

$$c_j = \frac{\sum_{i=1}^{n} u_{ij}\, x_i}{\sum_{i=1}^{n} u_{ij}}, \qquad j = 1, \dots, k.$$

این دو گام به‌صورت تکراری اجرا می‌شوند تا یکی از معیارهای توقف، مانند کاهش ناچیز مقدار تابع هدف (۱) یا رسیدن به حداکثر تعداد تکرارهای مجاز، برقرار شود [۴]. از آنجا که در هر تکرار مقدار تابع هدف $J(U,C)$ کاهش می‌یابد یا بدون تغییر باقی می‌ماند و این تابع همواره نامنفی است، دنباله مقادیر آن غیرصعودی و از پایین کراندار خواهد بود؛ بنابراین، همگرایی الگوریتم تضمین می‌شود. با این حال، به دلیل غیرمحدب بودن مسئله خوشه‌بندی، الگوریتم لزوماً به کمینه سراسری همگرا نمی‌شود و معمولاً در یک کمینه محلی متوقف می‌گردد. در نتیجه، کیفیت جواب نهایی به انتخاب مراکز اولیه خوشه‌ها وابسته بوده و مقداردهی اولیه نقش مهمی در عملکرد الگوریتم ایفا می‌کند.

۳-۲- مقداردهی اولیه و الگوریتم K-means ++

برای کاهش وابستگی K-means به مقداردهی اولیه، آرتور و واسیلیویتسکی [۵] روش K-means++ را معرفی کردند که در آن مراکز اولیه به‌صورت تصادفی هدایت‌شده انتخاب می‌شوند؛ به‌گونه‌ای که نقاط دورتر از مراکز موجود، شانس بیشتری برای انتخاب دارند. این روش معمولاً کیفیت خوشه‌بندی را بهبود می‌بخشد.

برای پیاده‌سازی این ایده، ابتدا یک نقطه از مجموعه‌داده X به‌صورت یکنواخت انتخاب و به‌عنوان نخستین مرکز در نظر گرفته می‌شود. سپس برای هر نقطه x_i، کمترین فاصله آن تا مراکز انتخاب‌شده محاسبه می‌شود:

$$D(x_i) = \min_{1\le \ell \le t} d(x_i, c_\ell),$$

که در آن t تعداد مراکز انتخاب‌شده تا آن مرحله است. مرکز بعدی با احتمالی متناسب با مربع این فاصله انتخاب می‌شود:

$$\mathbb{P}\left(\text{به عنوان مرکز بعدی انتخاب شود } x_i\right) \propto D(x_i)^2,$$

در نتیجه، نقاطی که از مراکز فعلی دورتر هستند، شانس بیشتری برای انتخاب به‌عنوان مرکز جدید دارند. آرتور و واسیلیویتسکی [۵] نشان دادند که این راهبرد مقداردهی اولیه، مقدار تابع هدف نهایی را در امیدریاضی تا کران $O(\log k)$ نسبت به جواب بهینه سراسری تضمین می‌کند. از این‌رو، K-means++ معمولاً نسبت به نسخه کلاسیک، خوشه‌بندی‌های با کیفیت‌تر و پایدارتری تولید می‌کند.

در این مقاله، این ایده مقداردهی اولیه به فضای مبتنی بر نظریه دمپستر– شیفر تعمیم داده شده و معیار فاصله اقلیدسی $d(\cdot,\cdot)$ با فاصله شواهد جایگزین می‌شود. این جایگزینی امکان مقداردهی اولیه K-means++ را در قالب خوشه‌بندی شواهدمحور فراهم می‌کند.

۳-۳- نظریه دمپستر– شیفر و بدنه‌های شواهد

نظریه دمپستر– شیفر[1] (نظریه شواهد) چارچوبی

1-Dempster–Shafer Theory

ریاضی برای نمایش و ترکیب شواهد در حضور عدم‌قطعیت ارائه می‌کند [۶, ۷]. فرض کنید.

$$\Theta = \{\theta_1, \dots, \theta_m\}$$

مجموعه‌ای متناهی از فرضیه‌های ممکن باشد که چارچوب تمایز[۲] نامیده می‌شود. مجموعه توانی آن ($2^\Theta = \{A \mid A \subseteq \Theta\}$) تمام زیرمجموعه‌های ممکن از Θ را در بر می‌گیرد. برای نمایش میزان شواهد موجود درباره فرضیه‌های مختلف، در نظریه دمپستر- شیفر از تابع تخصیص احتمال پایه[۳] (BPA) استفاده می‌شود. این تابع به هر زیرمجموعه از چارچوب تمایز، مقداری بین صفر و یک اختصاص می‌دهد که بیانگر میزان حمایت شواهد از آن زیرمجموعه است. به عبارت دیگر، مقداری که به یک زیرمجموعه (A) اختصاص داده شود، نشان می‌دهد که شواهد موجود تا چه اندازه از این گزاره پشتیبانی می‌کنند که پاسخ واقعی در یکی از عناصر (A) قرار دارد.

تعریف ۱. تابع تخصیص احتمال پایه (BPA) تابعی به‌صورت

$$m: 2^\Theta \to [0,1]$$

است، به‌طوری که

$$m(\varnothing) = 0, \qquad \sum_{A \subseteq \Theta} m(A) = 1.$$

هر زیرمجموعه $A \subseteq \Theta$ که $m(A) > 0$ باشد، عضو کانونی[۴] نامیده می‌شود و مقدار $m(A)$ مقدار باور اولیه نسبت به این است که پاسخ واقعی در مجموعه A قرار دارد [۷].

در نظریه دمپستر- شیفر، برای نمایش کامل مجموعه شواهد از مفهوم بدنه شواهد[۵](BOE) استفاده می‌شود.

تعریف ۲. بدنه شواهد به‌صورت زوج مرتب (B, m) تعریف می‌شود که در آن $B \subseteq 2^\Theta$ مجموعه اعضای کانونی و m تابع تخصیص احتمال پایه متناظر است. این ساختار امکان می‌دهد تا عدم‌قطعیت و ابهام، از طریق تخصیص جرم به زیرمجموعه‌های مختلف (حتی شامل چندین فرضیه به‌طور هم‌زمان)، به‌صورت صریح مدل‌سازی شود. بر پایه تابع m، دو تابع استنتاجی رایج تعریف می‌شود. نخست، تابع باور[۶] که برای هر $A \subseteq \Theta$ به‌صورت:

$$\mathrm{Bel}(A) = \sum_{B \subseteq A} m(B), \qquad (۲)$$

است. این تابع میزان حداقل پشتیبانی شواهد از گزاره A را نشان می‌دهد. به بیان دیگر، تنها جرم‌هایی در آن لحاظ می‌شوند که به‌طور کامل درون A قرار دارند. دوم، تابع امکان‌پذیری[۷] که به‌صورت مکملی، میزان بیشینه پشتیبانی از A را بیان می‌کند و بر اساس جرم‌هایی تعریف می‌شود که با A ناسازگار نیستند. که مطابق رابطه

$$\mathrm{Pl}(A) = 1 - \mathrm{Bel}(\Theta \setminus A) = \sum_{B \cap A \neq \varnothing} m(B), \qquad (۳)$$

تعریف می‌شود. در نتیجه، برای هر $A \subseteq \Theta$ همواره نامساوی $\mathrm{Bel}(A) \le \mathrm{Pl}(A)$ برقرار است و بازه $[\mathrm{Bel}(A), \mathrm{Pl}(A)]$ به‌عنوان کران‌های پایین و بالای میزان پشتیبانی شواهد از A تفسیر می‌شود [۷]. در خوشه‌بندی معمولی مثل K-means، هر نمونه یک بردار عددی است:

$$\boldsymbol{x_i} = (\boldsymbol{x_{i1}}, \boldsymbol{x_{i2}}, \dots, \boldsymbol{x_{id}})$$

و فاصله بین نمونه‌ها با معیارهایی مانند فاصله اقلیدسی محاسبه می‌شود. اما در خوشه‌بندی شواهدمحور هر نمونه به‌جای یک نقطه قطعی، با یک بدنه شواهد (BOE) یا یک تابع جرم $\boldsymbol{m}$ نمایش داده می‌شود.

به عنوان یک مثال ساده فرض کنید سه خوشه داریم:

$$\boldsymbol{\Theta} = \{\boldsymbol{C_1}, \boldsymbol{C_2}, \boldsymbol{C_3}\}$$

برای یک نمونه خاص، اطلاعات موجود ممکن است چنین باشد:

$$\boldsymbol{m}(\{\boldsymbol{C_1}\}) = \boldsymbol{0.7}$$

2-Frame of Discernment
3-Basic Probability Assignment
4-Focal Element
5-Body of Evidence
6-Belief Function
7-Plausibility Function

یعنی شـــواهد نســـبتاً قوی از تعلق نمونه به خوشه C_1 وجود دارد. اما شاید برای نمونه‌ای دیگر نتوان با اطمینان تصمیم گرفت:

$$m(\{C_1, C_2\}) = 0.8$$

ایـــن مقدار می‌گویـــد می‌دانیم نمونه متعلـــق به یکی از خوشـــه‌های C_1 یا C_2 اســـت، اما نمی‌توانیم دقیقاً مشخص کنیم کدام یـــک. این نوع ابهـــام در K-means معمولی قابل نمایش نیست.

قاعده ترکیب دمپستر

در بسیاری از مسائل، اطلاعات مربوط به یک پدیده از چند منبع مختلف به دست می‌آید. از آنجا که هر منبع ممکن اســـت تنها بخشی از واقعیت را توصیف کند، لازم اســـت شـــواهد حاصل از منابع مختلف با یکدیگر ترکیب شوند تا ارزیابی جامع‌تری از وضعیت واقعی حاصل شـــود. نظریه دمپســـتر- شیفر برای این منظور قاعده ترکیب دمپســـتر را ارائـــه می‌کند. اگر دو منبع مســـتقل بـــا توابع تخصیص احتمـــال پایه m_1 و m_2 روی چارچـــوب تمایز Θ توصیف شده باشند. ایده اصلی قاعده دمپستر آن است که جرم‌های اختصاص‌یافته به زیرمجموعه‌هایی که با یکدیگر سازگار هســـتند، با هم ترکیب شـــوند و جرم‌های ناشی از شواهد متناقض به‌صورت جداگانه محاسبه و نرمال‌سازی گردند. تابع تخصیص ترکیبی m به‌صورت زیر تعریف می‌شـــود [۶]:

$$m(C) = \frac{1}{1-K} \sum_{\substack{A \cap B = C \\ A,B \subseteq \Theta}} m_1(A)\, m_2(B), \qquad C \neq \emptyset, \tag{۴}$$

که در آن

$$K = \sum_{\substack{A \cap B = \emptyset \\ A,B \subseteq \Theta}} m_1(A)\, m_2(B) \tag{۵}$$

ضریب ناســـازگاری بین دو منبع شواهد است. مقدار K نشـــان‌دهنده بخشی از جرم است که از تقاطع‌های تهی حاصل می‌شـــود و بنابراین بیانگر میزان تعارض میان دو منبع اســـت. هر چه K بزرگ‌تر باشـــد، ناسازگاری شواهد بیشـــتر خواهد بود. در صورتی که K<1 باشد، تابع m یک تابع تخصیص احتمال پایه معتبر خواهد بود.

۳-۴- فاصله شواهد بین بدنه‌های شواهد

در خوشـــه‌بندی شـــواهدمحور، نمونه‌هـــا به‌صورت بدنه‌های شـــواهد نمایـــش داده می‌شـــوند؛ بنابراین برای مقایســـه آنها باید معیاری برای سنجش میزان شباهت یا عدم‌شـــباهت میان دو بدنه شـــواهد در اختیار داشت. یکی از کاربردی‌ترین معیارها در این زمینه، فاصله پیشـــنهادی ژوسلم و همکاران اســـت [۸]. برخلاف فاصله‌های متداول که تنها اخـــتلاف مقادیر را در نظـــر می‌گیرند، این فاصله میزان همپوشـــانی میان زیرمجموعه‌های چارچوب تمایز را نیـــز لحـــاظ می‌کند. از ایـــن‌رو، دو بدنه شـــواهد که به زیرمجموعه‌های مشـــابه یا همپوشـــان جـــرم اختصاص داده‌اند، به یکدیگر نزدیک‌تر در نظر گرفته می‌شوند.

فرض کنیـــد $m_1, m_2 : 2^{\Theta} \to [0,1]$ دو تابـــع تخصیص احتمال پایه باشـــند. با در نظر گرفتن یک ترتیب ثابت برای تمام زیرمجموعه‌های $A \subseteq \Theta$، بردارهای ستونی $\mathbf{m}_1$ و $\mathbf{m}_2$ (با طول $2^{|\Theta|}$) از مقادیر $m_1(A)$ و $m_2(A)$ ساخته می‌شوند. در این صورت، فاصله ژوسلم میان m_1 و m_2 به‌صورت

$$d_J(m_1, m_2) = \sqrt{\frac{1}{2}(\mathbf{m}_1 - \mathbf{m}_2)^T \mathbf{D} (\mathbf{m}_1 - \mathbf{m}_2)}, \tag{۶}$$

تعریف می‌شـــود. در رابطه فـــوق، D ماتریسی با ابعاد $2^{|\Theta|} \times 2^{|\Theta|}$ است که شـــباهت میان زیرمجموعه‌های Θ را کدگـــذاری می‌کند. اگـــر زیرمجموعه‌های $A, B \subseteq \Theta$ مطابق همان ترتیب به‌کار رفته در ساخت بردارها شاخص‌گذاری شوند، درایه متناظر با (A,B) از $\mathbf{D}$ معمولاً بر مبنای ضریب ژاکارد[8] به‌شکل

$$D_{A,B} = \begin{cases} \dfrac{|A \cap B|}{|A \cup B|}, & A \cup B \neq \emptyset, \\ 0, & A \cup B = \emptyset, \end{cases} \tag{۷}$$

8- Jaccard Index

تعریف می‌گردد. بر این اساس، $d_J(\cdot,\cdot)$ یک فاصله معتبر (معیار) روی فضای توابع تخصیص احتمال پایه است و خواص نامنفی‌بودن، تقارن و نابرابری مثلثی را دارد [۸]. این فاصله معمولاً به‌عنوان معیار استاندارد عدم شباهت میان بدنه‌های شواهد به‌کار می‌رود و مبنایی برای تعریف فاصله شواهد میان نمونه‌ها در الگوریتم‌های خوشه‌بندی شواهدمحور محسوب می‌شود.

بنابراین، هر چه دو زیرمجموعه همپوشانی بیشتری داشته باشند، مقدار $D_{A,B}$ بزرگ‌تر خواهد بود و این شباهت در محاسبه فاصله لحاظ می‌شود. فاصله ژوسلم یک معیار معتبر روی فضای توابع تخصیص احتمال پایه است و خواص نامنفی بودن، تقارن و نابرابری مثلثی را دارد [۸]. به همین دلیل، این فاصله به‌طور گسترده به‌عنوان معیار استاندارد مقایسه بدنه‌های شواهد و مبنایی برای تعریف فاصله میان نمونه‌ها در الگوریتم‌های خوشه‌بندی شواهدمحور به‌کار می‌رود.

۴- الگوریتم پیشنهادی: ED-K-means ++

در خوشه‌بندی شواهدمحور مبتنی بر نظریه دمپستر-شیفر، هر نمونه داده به‌جای یک نقطه قطعی در $\mathbb{R}^d$ به‌صورت یک تابع تخصیص احتمال پایه نمایش داده می‌شود و عدم شباهت میان نمونه‌ها با فاصله میان بدنه‌های شواهد متناظر سنجیده می‌گردد. در روش [۹]، مقادیر ویژگی‌های هر نمونه به‌عنوان منابع شواهد در نظر گرفته می‌شوند و بر اساس یک قاعده ساخت از پیش تعیین‌شده، برای هر نمونه x_i یک BPA به‌صورت m_i ساخته می‌شود. سپس برای مقایسه دو نمونه، فاصله ژوسلم در رابطه (٦) به‌کار می‌رود.

به‌طور خلاصه، الگوریتم ED-K-means در [۹] چنین عمل می‌کند: ابتدا برای هر نمونه x_i یک بدنه شواهد تشکیل می‌شود. سپس k مرکز اولیه به‌صورت تصادفی انتخاب می‌گردد. در هر تکرار، فاصله شواهد بین هر نمونه و هر مرکز محاسبه شده و نمونه به خوشه‌ای اختصاص می‌یابد که این فاصله در آن کمینه است. پس از تخصیص، مراکز خوشه‌ها در فضای شواهد به‌روزرسانی می‌شوند، برای مثال می‌توان از میانگین‌گیری بردارهای $\mathbf{m}_i$ برای تعریف مرکز جدید استفاده کرد. این دو گام تا برقراری معیار توقف تکرار می‌شوند. نتایج گزارش‌شده در [۹] نشان می‌دهد که جایگزینی فاصله اقلیدسی با فاصله شواهد، در برخی داده‌های واقعی و مصنوعی می‌تواند به جداسازی بهتر خوشه‌ها و پایداری بیشتر منجر شود.

در این مقاله، الگوریتم ED-K-means [۹] با جایگزینی مقداردهی اولیه تصادفی توسط طرح K-means++ [۵] تعمیم داده می‌شود. الگوریتم حاصل که ED-K-means ++ نام‌گذاری می‌شود، مراحل اصلی ED-K-means را حفظ می‌کند، اما مراکز اولیه را در فضای شواهد و بر اساس فاصله شواهد انتخاب می‌نماید. این انتخاب هدفمند با هدف کاهش حساسیت به مقداردهی اولیه و افزایش پایداری در حضور نوفه انجام می‌شود، بدون آنکه مرتبه محاسباتی، نسبت به روش پایه تغییر کند. در ادامه مراحل انجام این الگوریتم را کامل توضیح می‌دهیم.

۴-۱- تعریف الگوریتم در فضای شواهد و ساخت تابع تخصیص احتمال پایه

فرض کنید مجموعه‌داده‌ای شامل n نمونه به‌صورت

$$X = \{x_1, \dots, x_n\} \subset \mathbb{R}^d$$

در اختیار باشد. در این روش، هر نمونه داده به‌جای آن که مستقیماً در فضای ویژگی‌ها مقایسه شود، ابتدا به یک BPA تبدیل می‌شود. هدف از این کار، نمایش میزان تعلق هر نمونه به خوشه‌های مختلف همراه با درجه‌ای از عدم‌قطعیت است. تا مقایسه نمونه‌ها به‌جای فضای اقلیدسی، در فضای شواهد انجام گیرد. فرض کنید.

$$\Theta = \{\theta_1, \theta_2, \dots, \theta_k\}$$

چارچوب تمایز باشد، که در آن k تعداد خوشه‌ها و هر θ_j نماینده یک خوشه است. برای سادگی، جرم‌ها تنها به فرضیه‌های تک‌عضوی و کل چارچوب تمایز اختصاص

داده می‌شوند؛ بنابراین فقط مقادیر

$$m_i(\{\theta_1\}), \dots, m_i(\{\theta_k\}), m_i(\Theta)$$

می‌توانند غیرصفر باشند و برای سایر زیرمجموعه‌های Θ، مقدار جرم صفر است. برای ساخت این جرم‌ها، ابتدا k نمونه پیش‌نمون[9]، $p_1, p_2, \dots, p_k \in \mathbb{R}^d$ از میان نمونه‌های استانداردشده انتخاب می‌شوند. این نمونه‌های پیش‌نمون با یک مقدار تصادفی ثابت انتخاب شده‌اند تا فرایند ساخت BPA در تمام اجراها بازتولیدپذیر باشد. سپس برای هر نمونه x_i، فاصله مربعی آن تا هر نمونه نماینده محاسبه می‌شود:

$$d_{ij}^2 = \| x_i - p_j \|_2^2, \qquad j = 1, \dots, k.$$

در ادامه، میزان نزدیکی نمونه x_i به هر نمونه پیش‌نمون با نگاشت نرم زیر محاسبه می‌شود:

$$q_{ij} = \frac{\exp(-\alpha d_{ij}^2)}{\sum_{\ell=1}^{k} \exp(-\alpha d_{i\ell}^2)}, \qquad j = 1, \dots, k,$$

که در آن $\alpha > 0$ پارامتر حساسیت نسبت به فاصله است. هرچه مقدار α بزرگ‌تر باشد، جرم بیشتری به نمونه پیش‌نمون نزدیک‌تر اختصاص می‌یابد. پس از محاسبه مقادیر q_{ij}، جرم هر نمونه بین خوشه‌های منفرد و یک مؤلفه عدم‌قطعیت توزیع می‌شود. به‌طور مشخص، سهم اختصاص‌یافته به هر خوشه متناسب با مقدار q_{ij} تعیین شده و بخش باقیمانده به کل چارچوب تمایز Θ اختصاص می‌یابد. براین اساس جرم‌های تک‌عضوی و جرم اختصاص‌یافته به کل چارچوب تمایز به‌صورت زیر تعریف می‌شوند:

$$m_i(\{\theta_j\}) = (1 - u) q_{ij}, \qquad j = 1, \dots, k,$$

و

$$m_i(\Theta) = u,$$

که در آن $u \in [0,1]$ میزان ناآگاهی کلی مدل را نشان می‌دهد. در این مقاله، مطابق تنظیمات آزمایش‌ها، مقادیر $\alpha = 1.5, u = 0.2$ ثابت در نظر گرفته شده‌اند با این تعریف، برای هر نمونه x_i داریم:

$$\sum_{j=1}^{k} m_i(\{\theta_j\}) + m_i(\Theta) = (1 - u) \sum_{j=1}^{k} q_{ij} + u = 1.$$

بنابراین m_i یک BPA معتبر است. پس از ساخت توابع جرم، فاصله میان دو نمونه x_i و x_ℓ با فاصله ژوسلم میان توابع جرم متناظر آن‌ها تعریف می‌شود:

$$d_E(m_i, m_\ell) = d_J(m_i, m_\ell).$$

در این نمایش، هر نمونه به‌صورت بردار زیر در نظر گرفته می‌شود:

$$\mathbf{m}_i = [m_i(\{\theta_1\}), \dots, m_i(\{\theta_k\}), m_i(\Theta)]^T$$

و فاصله ژوسلم بر همین بردارهای جرم محاسبه می‌گردد.

هدف الگوریتم، افراز نمونه‌ها به k خوشه در فضای شواهد است. مراکز خوشه نیز به‌صورت BPA نمایش داده می‌شوند:

$$m_{c_1}, \dots, m_{c_k}$$

اگر $z_i \in \{1, \dots, k\}$ برچسب خوشه نمونه x_i باشد، تابع هدف به‌صورت

$$\tilde{J}(Z, M_C) = \sum_{j=1}^{k} \sum_{i: z_i = j} d_E(m_i, m_{c_j})^2, \qquad (8)$$

تعریف می‌شود که در آن $Z = (z_1, \dots, z_n)$ و $M_C = (m_{c_1}, \dots, m_{c_k})$ مجموعه مراکز خوشه‌ها در فضای شواهد است. الگوریتم ED-K-means ++ با تکرار گام‌های تخصیص و به‌روزرسانی، مقدار تابع هدف (۸) را کاهش می‌دهد و تا رسیدن به معیار توقف ادامه می‌یابد.

در این مطالعه، مقدار u برای همه نمونه‌ها ثابت در نظر گرفته شده است. بنابراین هدف، یادگیری تطبیقی میزان ناآگاهی هر نمونه نیست، بلکه ایجاد یک بازنمایی شواهدی ساده و کنترل‌شده برای بررسی اثر مقداردهی اولیه در فضای شواهد است.

9-Prototype

۴-۲- مقداردهی اولیه و تکرارهای الگوریتم ED-K-means ++

الگوریتم ED-K-means ++ ساختار اصلی ED-K-means را حفظ می‌کند، با این تفاوت که مراکز اولیه به‌جای انتخاب تصادفی، با روش k–میانگین++ و بر مبنای فاصله شواهد d_E تعیین می‌شوند. ابتدا یک نمونه x_{i_1} به‌طور یکنواخت انتخاب و $m_{c_1} = m_{i_1}$ قرار داده می‌شود. سپس در مرحله tام، برای هر نمونه x_i مقدار

$$D_i = \min_{1 \le \ell \le t} d_E(m_i, m_{c_\ell})$$

محاسبه شده و سپس نمونه x_i با احتمال:

$$\mathbb{P}\left(x_i \text{ به عنوان مرکز جدید انتخاب شود}\right) = \frac{D_i^2}{\sum_{r=1}^{n} D_r^2} \qquad (۹)$$

به‌عنوان مرکز بعدی انتخاب می‌شود. پس از تعیین مراکز اولیه، الگوریتم به‌صورت تکراری شامل دو گام اجرا می‌شود:

گام تخصیص:

$$z_i = \arg\min_{1 \le j \le k} d_E\left(m_i, m_{c_j}\right), \qquad (۱۰)$$

و گام به‌روزرسانی، اگر

$$S_j = \{\, i \mid z_i = j \,\},$$

آنگاه مرکز جدید هر خوشه به‌صورت زیر محاسبه می‌شود:

$$(A) = \frac{1}{|S_j|} \sum_{i \in S_j} m_i\,(A), \qquad \forall A \subseteq \Theta, \qquad (۱۱)$$

که معادل برداری نیز به صورت میانگین BPAها است. از آنجا که میانگین‌گیری مؤلفه‌ای جرم‌ها نامنفی بوده و مجموع آنها برابر ۱ باقی می‌ماند، $m_{c_j}^{(new)}$ همچنان یک BPA معتبر خواهد بود.

تکرارها تا زمانی ادامه می‌یابد که:

$$\max_{1 \le j \le k} d_E\left(m_{c_j}^{(\text{new})}, m_{c_j}^{(\text{old})}\right) < \varepsilon$$

یا تعداد تکرارها به $T_{\max}$ برسد. در صورت جایگزینی d_E با فاصله اقلیدسی، مرحله مقداردهی اولیه به نسخه کلاسیک K-means ++ تقلیل می‌یابد [۵]. روند کلی الگوریتم در الگوریتم ۱ خلاصه شده است.

الگوریتم ۱: ED-K-means ++ مبتنی بر فاصله شواهد

ورودی: $X = \{x_1, \dots, x_n\}$، تعداد خوشه‌ها k، آستانه همگرایی ε، حداکثر تعداد تکرارها $T_{\max}$

خروجی: $Z = (z_1, \dots, z_n)$ و $M_C = (m_{c_1}, \dots, m_{c_k})$

۱. برای هر x_i، تابع جرم m_i را مطابق قاعده ساخت BPA ارائه‌شده در بخش ۳،۱ محاسبه کن.

۲. مراکز اولیه خوشه‌ها را با استفاده از روش $ED - K - means++$ و برپایه فاصله شواهد d_E انتخاب کن

برای $t = 1$ تا $T_{\max}$ مراحل زیر را تکرار کن:

۱،۳ برچسب هر نمونه را با استفاده از رابطه (۱۰) به‌روزرسانی کن.

۲،۳ مراکز فعلی خوشه‌ها را به عنوان مراکز قبلی ذخیره کن.

۳،۳ مراکز خوشه‌ها را با استفاده از رابطه (۱۱) به‌روزرسانی کن.

۴،۳ اگر

$$\max_j d_E\left(m_{c_j}^{(\text{new})}, m_{c_j}^{(\text{old})}\right) < \varepsilon$$

الگوریتم را خاتمه بده.

Z و M_C را بازگردان.

۵- نتایج عددی و تحلیل پایداری

در این بخش، عملکرد چهار روش ED-K-means، KME، KME++ و روش پیشنهادی ED-K-means ++ بر اساس $R = 30$ اجرای مستقل بررسی می‌شود. روش KME و KME ++ به‌ترتیب نسخه کلاسیک و نسخه دارای مقداردهی اولیه k–میانگین++ در فضای اقلیدسی هستند. در مقابل، ED-K-means و ED-K-means ++ در فضای شواهد و بر اساس فاصله ژوسلم عمل می‌کنند. هدف این مقایسه آن است که اثر مقداردهی اولیه از اثر تغییر فضای بازنمایی و معیار فاصله تفکیک شود.

برای ارزیابی بیرونی خوشه‌بندی، از شاخص ARI

اســتفاده شده اســت [۱۳]. همچنین برای بررسی ساختار درونی خوشــه‌ها، ضریب سیلوئت [۱] و شاخص دویس-بولدین [۲] گزارش می‌شــوند. از آن‌جا که شــاخص‌های درونی برای روش‌های اقلیدسی با فاصله اقلیدسی و برای روش‌های شــواهدی با فاصله شواهد محاســبه شده‌اند، مقایسه مستقیم مقدار آن‌ها میان دو خانواده روش‌ها باید با احتیاط انجام شــود. در نتیجــه، ARI معیار اصلی برای مقایسه بیرونی کیفیت خوشه‌بندی است و سیلوئت و DBI بیشتر برای تحلیل انسجام درونی در هندسه فاصله متناظر هر روش به‌کار می‌روند.

۵-۱- تحلیل نتایج

نتایج گزارش‌شده در جدول‌های ۱ تا ۹ نشان می‌دهد که ED-K-means ++ را نبایــد به‌عنوان روشی با برتری مطلق در همه داده‌ها و همه شــاخص‌ها تفسیــر کرد. در برخی مجموعه‌داده‌ها، به‌ویژه Wine و Breast Cancer، روش‌های اقلیــدسی KME و KME ++ از نظر ARI انطباق بیشــتری با برچســب‌های مرجــع دارند. بنابراین، ســهم اصلی روش پیشنهادی در افزایش عمومی دقت بیرونی نیست، بلکه در کاهش حساسیت نسخه شــواهدی الگوریتم به مقداردهی اولیه و کنترل نوسان نتایج است.

در حالــت بدون اغتشــاش، افزودن مقــداردهی اولیه k-میانگیــن++ بــه روش اقلیــدسی تغییــر چشــمگیری در عملکــرد ایجــاد نمی‌کنــد. در خانواده شــواهدی نیز ED-K-means ++ لزومــاً میانگیــن ARI را نســبت بــه ED-K-means افزایش نمی‌دهد، اما در برخی داده‌ها انحراف معیار کمتری نشــان می‌دهد. این الگو بیانگر آن اســت که مقداردهی اولیه پیشــنهادی بیش از آن که موجب افزایش مســتقیم دقت بیرونی شود، رفتار الگوریتم را در اجراهای مختلف یکنواخت‌تر می‌کند.

درحالت نوفهٔ گاوسی، میانگین ARI روش پیشــنهادی در بیشــتر موارد نزدیک به ED-K-means باقی می‌ماند. با این حال، کاهش نوسان برخی شاخص‌های درونی، به‌ویژه DBI، نشــان می‌دهد که ED-K-means ++ می‌تواند از برخی اجراهای ناپایــدار جلوگیری کند. این موضوع به‌خصوص در داده Iris و ســطوح بالاتر نوفهٔ روشن‌تر است؛ جایی که نسخه پایه شــواهدی دچار پراکندگی شدید در مقدار DBI می‌شود، اما روش پیشنهادی رفتار کنترل‌شده‌تری دارد.

در حالت آلودگی با داده پرت شبیه‌سازی‌شــده، نسبت نمونه‌های آلوده با ρ تغییر کرده و شــدت آلودگی ثابت و برابر با $\gamma = ۳$ در نظر گرفته شــده اســت. مطابق انتظار، افزایش ρ عموما باعث کاهش کیفیت خوشه‌بندی می‌شود. با این حــال، در برخی حالت‌ها، به‌ویــژه در مجموعه‌داده Wine , ED-K-means++ نسبت به مقداردهی اولیه تصادفی در فضای شواهد نوسان کمتری دارد. بنابراین، مقداردهی اولیه در فضای شواهد می‌تواند در برابر بخشی از بی‌ثباتی ناشی از نقاط آلوده نقش کنترل‌کننده داشته باشد، هر چند این اثر در همه داده‌ها و همه سطوح آلودگی یکسان نیست

شــاخص‌های درونی مانند ضریب سیلوئت و DBI در بسیاری از حالت‌ها ســاختار منظم‌تری را برای روش‌های شــواهدی، به‌ویــژه ED-K-means ++، نشــان می‌دهند. با وجــود این، چون این شــاخص‌ها در دو خانواده روش‌ها بــا فاصله‌های متفاوت محاســبه شــده‌اند، نبایــد از آنها برای ادعای برتری مطلق روش‌های شــواهدی نســبت به روش‌های اقلیدسی اســتفاده کرد. تفسیر دقیق‌تر آن است که روش پیشنهادی در هندسه فاصله شواهد، خوشه‌هایی با انسجام درونی بهتر یا نوسان کمتر ایجاد می‌کند.

۵-۲- تحلیل حساسیت نسبت به ساخت تابع تخصیص احتمال پایه

از آنجا که در روش پیشــنهادی، فاصله شواهد بر پایه تابع تخصیص احتمال پایه ســاخته می‌شود، کیفیت نهایی خوشه‌بندی می‌تواند به نحوه ســاخت BPA وابسته باشد. برای بررسی ایــن موضوع، حســاسیت ED-K-means ++ نسبت به دو عامل ارزیابی شد: نخست، مقدار اولیه نمونه‌های نماینده در ساخت BPA، و دوم، پارامترهای α و u.

نتایج جدول ۱۰ نشــان می‌دهد کــه انتخاب مقدار اولیه نمونه‌های نماینده یکی از عوامل اثرگذار در عملکرد روش

پیشنهادی است. تغییر محسوس مقدار ARI در برخی مجموعه‌داده‌ها نشان می‌دهد که کیفیت بازنمایی شواهدی، نقش مهمی در هم‌راستایی خوشه‌های تولید شده با برچسب‌های مرجع دارد. بنابراین، اگرچه مقداردهی اولیه k–میانگین++ در فضای شواهد می‌تواند نوسان ناشی از انتخاب مراکز اولیه را کاهش دهد، کیفیت نهایی خوشه‌بندی همچنان به نحوهٔ ساخت BPA وابسته است.

مطابق جدول ۱۱، اثر α در بازه بررسی شده محسوس‌تر از اثر u است. این نتیجه با ساختار BPA استفاده شده در این مقاله سازگار است، زیرا مقدار $m_i(\theta) = u$ برای همه نمونه‌ها ثابت در نظر گرفته شده و بنابراین نقش آن بیشتر در تنظیم کلی بازنمایی شواهدی ظاهر می‌شود تا در ایجاد تمایز مستقیم میان نمونه‌ها.

جدول ۱. میانگین و انحراف معیار ARI در حالت بدون اغتشاش

مجموعه‌داده	KME	KME++	ED-K-means	ED-K-means++
Iris	۰.۵۷۸ ± ۰.۰۷۸	۰.۵۹۱ ± ۰.۰۶۸	۰.۵۳۸ ± ۰.۰۱۵	۰.۵۲۸ ± ۰.۰۰۱
Wine	۰.۸۹۷ ± ۰.۰۰۰	۰.۸۹۷ ± ۰.۰۰۰	۰.۴۱۸ ± ۰.۰۰۴	۰.۴۱۸ ± ۰.۰۰۰
Breast Cancer	۰.۶۶۲ ± ۰.۰۰۹	۰.۶۶۴ ± ۰.۰۰۸	۰.۰۰۸ ± ۰.۰۰۰	۰.۰۰۸ ± ۰.۰۰۰

جدول ۲. میانگین و انحراف معیار ARI0 تحت نوفهٔ گاوسی با سطوح مختلف β

مجموعه‌داده	β	KME	KME++	ED-K-means	ED-K-means++
Iris	۰.۰۰	۰.۵۶۵ ± ۰.۰۸۵	۰.۵۹۶ ± ۰.۰۶۳	۰.۵۳۷ ± ۰.۰۱۵	۰.۵۲۸ ± ۰.۰۰۱
Iris	۰.۰۵	۰.۶۰۲ ± ۰.۰۵۸	۰.۶۱۴ ± ۰.۰۳۸	۰.۵۳۵ ± ۰.۰۱۲	۰.۵۳۰ ± ۰.۰۰۴
Iris	۰.۱۰	۰.۵۹۱ ± ۰.۰۷۲	۰.۶۱۰ ± ۰.۰۴۹	۰.۵۳۳ ± ۰.۰۱۵	۰.۵۳۰ ± ۰.۰۰۷
Iris	۰.۲۰	۰.۶۰۴ ± ۰.۰۶۲	۰.۶۰۹ ± ۰.۰۵۴	۰.۵۳۸ ± ۰.۰۲۷	۰.۵۳۴ ± ۰.۰۲۶
Wine	۰.۰۰	۰.۸۸۰ ± ۰.۰۹۴	۰.۸۹۷ ± ۰.۰۰۰	۰.۴۱۸ ± ۰.۰۰۲	۰.۴۱۸ ± ۰.۰۰۰
Wine	۰.۰۵	۰.۹۰۱ ± ۰.۰۰۹	۰.۸۹۹ ± ۰.۰۰۹	۰.۴۱۵ ± ۰.۰۰۵	۰.۴۱۵ ± ۰.۰۰۵
Wine	۰.۱۰	۰.۸۹۳ ± ۰.۰۱۸	۰.۸۷۷ ± ۰.۰۹۵	۰.۴۱۳ ± ۰.۰۱۰	۰.۴۱۳ ± ۰.۰۱۰
Wine	۰.۲۰	۰.۸۷۱ ± ۰.۰۲۲	۰.۸۵۳ ± ۰.۰۹۴	۰.۴۱۵ ± ۰.۰۱۳	۰.۴۱۵ ± ۰.۰۱۳
Breast Cancer	۰.۰۰	۰.۶۶۴ ± ۰.۰۰۹	۰.۶۶۶ ± ۰.۰۰۸	۰.۰۰۸ ± ۰.۰۰۰	۰.۰۰۸ ± ۰.۰۰۰
Breast Cancer	۰.۰۵	۰.۶۶۶ ± ۰.۰۱۰	۰.۶۶۸ ± ۰.۰۱۱	۰.۰۰۷ ± ۰.۰۰۵	۰.۰۰۷ ± ۰.۰۰۵
Breast Cancer	۰.۱۰	۰.۶۶۸ ± ۰.۰۱۳	۰.۶۶۹ ± ۰.۰۱۳	۰.۰۰۷ ± ۰.۰۱۱	۰.۰۰۷ ± ۰.۰۱۱
Breast Cancer	۰.۲۰	۰.۶۶۹ ± ۰.۰۱۷	۰.۶۷۱ ± ۰.۰۱۷	۰.۰۰۸ ± ۰.۰۲۱	۰.۰۰۸ ± ۰.۰۲۱

جدول ۳. میانگین و انحراف معیار ARI در حضور آلودگی با داده‌های پرت شبیه‌سازی شده با سطوح مختلف ρ و شدت ثابت $\gamma = 3$.

مجموعه‌داده	ρ	KME	KME++	ED-K-means	ED-K-means++
Iris	۰.۰۰	۰.۵۶۶ ± ۰.۰۸۴	۰.۵۷۷ ± ۰.۰۸۰	۰.۵۳۶ ± ۰.۰۱۵	۰.۵۲۸ ± ۰.۰۰۱
Iris	۰.۰۵	۰.۴۹۶ ± ۰.۱۷۰	۰.۵۴۲ ± ۰.۰۳۵	۰.۵۱۱ ± ۰.۰۱۹	۰.۵۰۴ ± ۰.۰۱۵
Iris	۰.۱۰	۰.۴۲۱ ± ۰.۱۹۳	۰.۴۶۱ ± ۰.۱۵۸	۰.۴۳۳ ± ۰.۱۴۰	۰.۴۴۷ ± ۰.۱۱۴
Iris	۰.۲۰	۰.۴۰۴ ± ۰.۰۸۶	۰.۳۷۶ ± ۰.۱۵۳	۰.۳۵۳ ± ۰.۱۶۲	۰.۳۴۸ ± ۰.۱۵۹
Wine	۰.۰۰	۰.۸۸۰ ± ۰.۰۹۴	۰.۸۹۷ ± ۰.۰۰۰	۰.۴۱۸ ± ۰.۰۰۰	۰.۴۱۸ ± ۰.۰۰۰
Wine	۰.۰۵	۰.۴۷۱ ± ۰.۳۱۷	۰.۳۲۴ ± ۰.۱۵۸	۰.۳۷۵ ± ۰.۱۰۳	۰.۳۹۸ ± ۰.۰۱۸
Wine	۰.۱۰	۰.۳۷۹ ± ۰.۲۳۹	۰.۲۴۵ ± ۰.۱۶۵	۰.۳۴۲ ± ۰.۰۸۳	۰.۳۴۲ ± ۰.۰۸۳
Wine	۰.۲۰	۰.۳۲۵ ± ۰.۲۰۴	۰.۱۷۶ ± ۰.۱۹۶	۰.۲۸۹ ± ۰.۰۹۲	۰.۲۹۸ ± ۰.۰۷۵
Breast Cancer	۰.۰۰	۰.۶۶۴ ± ۰.۰۰۹	۰.۶۶۷ ± ۰.۰۰۷	۰.۰۰۸ ± ۰.۰۰۰	۰.۰۰۸ ± ۰.۰۰۰
Breast Cancer	۰.۰۵	۰.۶۲۲ ± ۰.۱۲۰	۰.۶۰۱ ± ۰.۱۶۷	۰.۰۰۹ ± ۰.۰۰۴	۰.۰۰۹ ± ۰.۰۰۴
Breast Cancer	۰.۱۰	۰.۶۰۸ ± ۰.۱۱۷	۰.۵۸۹ ± ۰.۱۶۳	۰.۰۰۹ ± ۰.۰۰۹	۰.۰۰۹ ± ۰.۰۰۹
Breast Cancer	۰.۲۰	۰.۵۸۸ ± ۰.۰۲۶	۰.۵۸۸ ± ۰.۰۲۸	۰.۰۱۳ ± ۰.۰۲۴	۰.۰۱۳ ± ۰.۰۲۴

جدول ۴. میانگین و انحراف معیار شاخص دویس- بولدین در حالت بدون اغتشاش.

مجموعه‌داده	KME	KME++	ED-K-means	ED-K-means++
Iris	۲.۲۳۱ ± ۰.۶۱۸	۲.۳۱۳ ± ۰.۶۰۰	۱.۹۱۵ ± ۰.۴۸۰	۱.۶۶۱ ± ۰.۷۱۴
Wine	۲.۱۷۷ ± ۰.۴۶۷	۲.۲۷۳ ± ۰.۳۴۲	۱.۵۰۴ ± ۰.۶۴۹	۱.۳۹۴ ± ۰.۷۲۶
Breast Cancer	۱.۶۷۱ ± ۰.۵۰۹	۱.۷۷۶ ± ۰.۵۳۴	۰.۸۴۳ ± ۰.۹۶۱	۰.۳۲۹ ± ۰.۶۶۶

جدول ۵. میانگین و انحراف معیار شاخص دویس- بولدین تحت نوفهٔ گاوسی با سطوح مختلف β

مجموعه‌داده	β	KME	KME++	ED-K-means	ED-K-means++
Iris	۰.۰۰	۲.۱۸۷ ± ۰.۶۱۶	۱.۹۰۵ ± ۰.۷۹۳	۱.۶۴۸ ± ۰.۸۶۳	۱.۳۵۸ ± ۰.۸۶۹
Iris	۰.۰۵	۲.۱۶۹ ± ۰.۶۰۰	۲.۱۸۵ ± ۰.۶۳۹	۱.۹۱۰ ± ۰.۵۶۱	۱.۶۶۶ ± ۰.۷۱۳
Iris	۰.۱۰	۲.۰۳۳ ± ۰.۶۱۹	۲.۰۹۳ ± ۰.۶۸۳	۱.۶۴۱ ± ۰.۷۴۲	۱.۵۳۵ ± ۰.۷۷۹
Iris	۰.۲۰	۲.۰۱۴ ± ۰.۷۰۵	۱.۹۹۴ ± ۰.۶۴۷	۱.۴۸۱ ± ۰.۵۷۱	۱.۳۰۶ ± ۰.۸۸۸
Wine	۰.۰۰	۲.۲۸۰ ± ۰.۳۸۸	۲.۳۳۷ ± ۰.۲۳۹	۱.۷۱۴ ± ۱.۲۰۹	۱.۵۵۴ ± ۰.۶۷۰
Wine	۰.۰۵	۲.۱۴۶ ± ۰.۴۵۲	۲.۲۳۵ ± ۰.۴۱۲	۱.۴۱۴ ± ۰.۷۳۹	۱.۶۱۵ ± ۰.۶۱۹
Wine	۰.۱۰	۲.۱۷۳ ± ۰.۴۶۰	۲.۲۱۴ ± ۰.۴۴۸	۱.۵۸۷ ± ۰.۶۸۶	۱.۵۰۰ ± ۰.۶۴۳
Wine	۰.۲۰	۲.۲۰۳ ± ۰.۴۵۸	۲.۲۸۱ ± ۰.۳۷۶	۱.۴۹۰ ± ۰.۷۲۸	۱.۶۳۲ ± ۰.۵۳۷
Breast Cancer	۰.۰۰	۱.۹۱۸ ± ۰.۵۳۶	۲.۰۲۴ ± ۰.۵۰۸	۱.۱۶۵ ± ۰.۹۷۲	۰.۵۲۲ ± ۰.۸۲۹
Breast Cancer	۰.۰۵	۱.۹۵۵ ± ۰.۵۳۰	۲.۱۶۶ ± ۰.۴۳۳	۱.۲۲۵ ± ۰.۹۶۵	۰.۷۷۳ ± ۰.۹۴۹
Breast Cancer	۰.۱۰	۱.۹۶۰ ± ۰.۵۲۷	۲.۰۳۰ ± ۰.۵۰۸	۱.۲۲۴ ± ۰.۹۶۷	۰.۵۱۱ ± ۰.۸۳۵
Breast Cancer	۰.۲۰	۱.۸۰۸ ± ۰.۵۲۶	۱.۹۴۷ ± ۰.۵۲۵	۱.۲۲۳ ± ۰.۹۶۷	۰.۵۱۲ ± ۰.۸۳۵

جدول۶. میانگین و انحراف معیار شاخص دویس-بولدین در حضور آلودگی با داده‌های پرت شبیه‌سازی‌شده با سطوح مختلف ρ و شدت ثابت $\gamma = 3$.

مجموعه‌داده	ρ	KME	KME++	ED-K-means	ED-K-means++
Iris	۰.۰۰	۲.۰۳۹ ± ۰.۶۷۸	۲.۱۴۱ ± ۰.۶۸۰	۱.۷۲۳ ± ۰.۷۲۷	۱.۳۷۱ ± ۰.۸۹۲
Iris	۰.۰۵	۲.۱۱۳ ± ۰.۷۹۶	۲.۴۱۰ ± ۰.۴۵۴	۱.۷۸۰ ± ۰.۶۸۹	۱.۷۴۷ ± ۰.۶۲۸
Iris	۰.۱۰	۲.۳۲۴ ± ۰.۶۷۵	۲.۳۷۹ ± ۰.۴۴۵	۱.۷۹۴ ± ۰.۹۰۱	۱.۷۱۳ ± ۰.۵۹۰
Iris	۰.۲۰	۲.۲۲۲ ± ۰.۵۹۳	۲.۲۴۵ ± ۰.۶۴۱	۱.۹۸۹ ± ۱.۹۶۹	۱.۴۰۹ ± ۰.۷۹۹
Wine	۰.۰۰	۲.۳۰۶ ± ۰.۲۹۴	۲.۲۵۵ ± ۰.۳۷۸	۱.۶۳۸ ± ۰.۵۳۷	۱.۶۷۸ ± ۰.۵۵۳
Wine	۰.۰۵	۲.۶۹۵ ± ۰.۸۸۷	۳.۴۲۱ ± ۰.۴۰۶	۱.۶۰۴ ± ۰.۷۶۶	۱.۳۹۸ ± ۰.۷۵۰
Wine	۰.۱۰	۲.۹۴۶ ± ۰.۸۶۲	۳.۲۴۶ ± ۰.۵۹۷	۱.۴۷۸ ± ۰.۶۸۹	۱.۴۸۵ ± ۰.۶۵۸
Wine	۰.۲۰	۲.۸۷۱ ± ۰.۶۷۲	۳.۱۹۲ ± ۰.۶۳۶	۱.۶۱۴ ± ۰.۹۸۴	۱.۴۹۲ ± ۰.۷۱۳
Breast Cancer	۰.۰۰	۱.۸۸۲ ± ۰.۵۴۰	۱.۹۵۲ ± ۰.۵۲۹	۱.۴۲۲ ± ۰.۸۹۸	۰.۵۸۶ ± ۰.۸۶۷
Breast Cancer	۰.۰۵	۲.۰۳۰ ± ۰.۵۱۳	۱.۷۱۷ ± ۰.۴۶۶	۱.۳۵۷ ± ۰.۹۲۵	۰.۶۴۹ ± ۰.۸۹۹
Breast Cancer	۰.۱۰	۲.۳۱۸ ± ۰.۵۲۳	۲.۰۹۶ ± ۰.۵۵۹	۱.۴۹۹ ± ۰.۸۶۴	۰.۶۳۰ ± ۰.۹۰۶
Breast Cancer	۰.۲۰	۲.۴۳۰ ± ۰.۴۷۲	۲.۳۳۷ ± ۰.۴۵۹	۱.۴۷۶ ± ۰.۸۸۰	۰.۶۴۰ ± ۰.۹۲۵

به‌طورکلی، تحلیل حساسیت نشان می‌دهد که ED-K-means ++ یک اصلاح پایدارکننده برای مقداردهی اولیه در خوشه‌بندی شواهدمحور است، اما به‌تنهایی مسئله انتخاب بازنمایی شواهدی را حل نمی‌کند. طراحی روش‌های پایدارتر برای انتخاب نمونه‌های نماینده و تنظیم تطبیقی پارامترهای ساخت BPA می‌تواند مسیر مهمی برای پژوهش‌های آینده باشد.

جدول۷. میانگین و انحراف معیار ضریب سیلوئت در حالت بدون اغتشاش.

مجموعه‌داده	KME	KME++	ED-K-means	ED-K-means++
Iris	۰.۴۶۵ ± ۰.۰۰۹	۰.۴۶۴ ± ۰.۰۰۸	۰.۸۴۰ ± ۰.۰۲۳	۰.۸۵۷ ± ۰.۰۰۱
Wine	۰.۲۸۵ ± ۰.۰۰۰	۰.۲۸۵ ± ۰.۰۰۰	۰.۹۳۵ ± ۰.۰۷۶	۰.۹۵۸ ± ۰.۰۰۰
Breast Cancer	۰.۳۴۴ ± ۰.۰۰۱	۰.۳۴۴ ± ۰.۰۰۱	۰.۹۷۲ ± ۰.۰۰۰	۰.۹۷۲ ± ۰.۰۰۰

جدول۸: میانگین و انحراف معیار ضریب سیلوئت تحت نوفهٔ گاوسی با سطوح مختلف β .

مجموعه‌داده	β	KME	KME++	ED-K-means	ED-K-means++
Iris	۰.۰۰	۰.۴۶۵ ± ۰.۰۰۹	۰.۴۶۲ ± ۰.۰۰۷	۰.۸۴۲ ± ۰.۰۲۲	۰.۸۵۷ ± ۰.۰۰۱
Iris	۰.۰۵	۰.۴۶۰ ± ۰.۰۰۶	۰.۴۶۰ ± ۰.۰۰۴	۰.۸۴۷ ± ۰.۰۲۶	۰.۸۵۹ ± ۰.۰۱۷
Iris	۰.۱۰	۰.۴۵۵ ± ۰.۰۰۹	۰.۴۵۳ ± ۰.۰۰۷	۰.۸۵۱ ± ۰.۰۳۴	۰.۸۶۶ ± ۰.۰۱۸
Iris	۰.۲۰	۰.۴۳۴ ± ۰.۰۱۲	۰.۴۳۵ ± ۰.۰۱۱	۰.۸۳۳ ± ۰.۰۵۱	۰.۸۵۰ ± ۰.۰۴۵
Wine	۰.۰۰	۰.۲۸۴ ± ۰.۰۰۵	۰.۲۸۵ ± ۰.۰۰۰	۰.۹۵۲ ± ۰.۰۳۲	۰.۹۵۸ ± ۰.۰۰۰
Wine	۰.۰۵	۰.۲۸۴ ± ۰.۰۰۱	۰.۲۸۴ ± ۰.۰۰۱	۰.۹۵۷ ± ۰.۰۰۷	۰.۹۵۷ ± ۰.۰۰۷
Wine	۰.۱۰	۰.۲۸۱ ± ۰.۰۰۲	۰.۲۸۱ ± ۰.۰۰۵	۰.۹۵۷ ± ۰.۰۱۱	۰.۹۵۷ ± ۰.۰۱۱
Wine	۰.۲۰	۰.۲۷۰ ± ۰.۰۰۳	۰.۲۶۹ ± ۰.۰۰۷	۰.۹۶۱ ± ۰.۰۱۲	۰.۹۶۱ ± ۰.۰۱۲
Breast Cancer	۰.۰۰	۰.۳۴۴ ± ۰.۰۰۱	۰.۳۴۴ ± ۰.۰۰۱	۰.۹۷۲ ± ۰.۰۰۰	۰.۹۷۲ ± ۰.۰۰۰
Breast Cancer	۰.۰۵	۰.۳۴۳ ± ۰.۰۰۱	۰.۳۴۳ ± ۰.۰۰۱	۰.۹۷۱ ± ۰.۰۰۴	۰.۹۷۱ ± ۰.۰۰۴
Breast Cancer	۰.۱۰	۰.۳۴۰ ± ۰.۰۰۱	۰.۳۴۰ ± ۰.۰۰۱	۰.۹۷۳ ± ۰.۰۰۵	۰.۹۷۳ ± ۰.۰۰۵
Breast Cancer	۰.۲۰	۰.۳۳۲ ± ۰.۰۰۲	۰.۳۳۲ ± ۰.۰۰۲	۰.۹۷۲ ± ۰.۰۰۵	۰.۹۷۲ ± ۰.۰۰۵

جدول ۹:میانگین و انحراف معیار ضریب سیلوئت در حضور آلودگی با داده‌های پرت شبیه‌سازی‌شده با سطوح مختلف ρو شدت ثابت $\gamma = 3$.

مجموعه‌داده	ρ	KME	KME++	ED-K-means	ED-K-means++
Iris	۰.۰۰	۰.۴۶۶ ± ۰.۰۱۰	۰.۴۶۴ ± ۰.۰۰۸	۰.۸۴۲ ± ۰.۰۲۲	۰.۸۵۷ ± ۰.۰۰۱
Iris	۰.۰۵	۰.۵۲۸ ± ۰.۰۶۶	۰.۵۱۳ ± ۰.۰۵۲	۰.۸۵۴ ± ۰.۰۵۰	۰.۸۷۱ ± ۰.۰۳۲
Iris	۰.۱۰	۰.۴۸۵ ± ۰.۰۵۸	۰.۴۸۴ ± ۰.۰۶۹	۰.۸۴۱ ± ۰.۰۷۹	۰.۸۷۰ ± ۰.۰۴۸
Iris	۰.۲۰	۰.۴۱۵ ± ۰.۰۴۲	۰.۴۳۴ ± ۰.۰۲۵	۰.۸۶۴ ± ۰.۰۸۱	۰.۹۰۰ ± ۰.۰۵۵
Wine	۰.۰۰	۰.۲۸۴ ± ۰.۰۰۵	۰.۲۸۵ ± ۰.۰۰۰	۰.۹۵۸ ± ۰.۰۰۰	۰.۹۵۸ ± ۰.۰۰۰
Wine	۰.۰۵	۰.۳۰۱ ± ۰.۱۳۹	۰.۲۸۱ ± ۰.۱۰۰	۰.۹۲۲ ± ۰.۱۲۲	۰.۹۶۰ ± ۰.۰۰۶
Wine	۰.۱۰	۰.۲۶۳ ± ۰.۱۱۸	۰.۳۰۵ ± ۰.۱۳۹	۰.۹۵۹ ± ۰.۰۱۳	۰.۹۵۹ ± ۰.۰۱۳
Wine	۰.۲۰	۰.۲۲۰ ± ۰.۰۹۹	۰.۳۰۸ ± ۰.۱۳۱	۰.۹۳۸ ± ۰.۰۹۴	۰.۹۶۲ ± ۰.۰۱۳
Breast Cancer	۰.۰۰	۰.۳۴۴ ± ۰.۰۰۱	۰.۳۴۴ ± ۰.۰۰۱	۰.۹۷۲ ± ۰.۰۰۰	۰.۹۷۲ ± ۰.۰۰۰
Breast Cancer	۰.۰۵	۰.۳۱۹ ± ۰.۰۵۵	۰.۳۲۷ ± ۰.۰۶۷	۰.۹۷۲ ± ۰.۰۰۲	۰.۹۷۲ ± ۰.۰۰۲
Breast Cancer	۰.۱۰	۰.۲۸۵ ± ۰.۰۵۱	۰.۲۹۵ ± ۰.۰۷۲	۰.۹۸۰ ± ۰.۰۱۲	۰.۹۸۰ ± ۰.۰۱۲
Breast Cancer	۰.۲۰	۰.۲۲۲ ± ۰.۰۰۸	۰.۲۲۲ ± ۰.۰۰۸	۰.۹۸۴ ± ۰.۰۱۲	۰.۹۸۴ ± ۰.۰۱۲

جدول ۱۰. خلاصه تحلیل حساسیت ED-K-means ++ نسبت به مقدار اولیه انتخاب نمونه‌های نماینده در ساخت BPA با $(\alpha = 1.5)$ و $(u = 0.2)$. مقادیر ARI در ستون دوم بر اساس پنج مقدار اولیه (۱۱۱,۲۲۲,۳۳۳,۴۴۴,۵۵۵) خلاصه شده‌اند.

مجموعه‌داده	ARI روی مقدار اولیه‌ها	بهترین مقدار اولیه از نظر ARI	دامنه Silhouette	دامنه DBI
Iris	۰.۴۳۲ ± ۰.۰۸۸	(۱۱۱) (۰.۵۲۸)	۰.۶۴۳–۰.۸۸۰	۱.۳۹۴–۲.۱۲۸
Wine	۰.۴۱۵ ± ۰.۱۴۴	(۴۴۴) (۰.۶۰۸)	۰.۸۳۷–۰.۹۵۸	۱.۴۰۲–۱.۶۷۶
Cancer Breast	۰.۲۳۱ ± ۰.۱۹۵	(۴۴۴) (۰.۴۳۶)	۰.۹۷۰–۰.۹۹۲	۰.۷۱۵–۱.۸۰۲

جدول ۱۱. پیکربندی‌های منتخب در تحلیل حساسیت ED-K-means++ نسبت به پارامترهای ساخت BPA. مقادیر از میان $(\alpha \in \{0.50, 1.00, 1.50, 2.00\})$ و $(u \in \{0.10, 0.20, 0.30\})$ انتخاب شده‌اند.

مجموعه‌داده	(α)	(u)	ARI	Silhouette	DBI
Iris	(۲,۰۰)	(۰,۱۰)	۰.۵۳۵ ± ۰.۰۰۰	۰.۸۹۱ ± ۰.۰۰۰	۱.۳۵۱ ± ۰.۸۵۶
Wine	(۱,۵۰)	(۰,۲۰)	۰.۴۱۸ ± ۰.۰۰۰	۰.۹۵۸ ± ۰.۰۰۰	۱.۲۶۵ ± ۰.۷۷۰
Breast Cancer	(۱,۵۰)	(۰,۲۰)	۰.۰۰۸ ± ۰.۰۰۰	۰.۹۷۲ ± ۰.۰۰۰	۰.۳۹۴ ± ۰.۷۳۱

۵- نتیجه‌گیری

در این مقاله، نسخه‌ای از الگوریتم ED-K-means ارائه شد که در آن مقداردهی اولیه تصادفی با روش K-means ++ در فضای شواهد جایگزین می‌شود. نتایج نشان دادند که روش پیشنهادی لزوماً در همه مجموعه‌داده‌ها موجب افزایش ARI نمی‌شود، اما در چند حالت باعث کاهش نوسان و جلوگیری از اجراهای ناپایدار می‌گردد. بنابراین، مزیت اصلی ED-K-means ++ بیشتر در پایدارسازی مقداردهی اولیه در فضای شواهد است، نه در تضمین افزایش دقت بیرونی در همه داده‌ها. تحلیل حساسیت نیز نشان داد که کیفیت روش‌های شواهدی به نحوه ساخت BPA و انتخاب نمونه‌های نماینده وابسته است. از این‌رو، توسعه روش‌های پایدارتر برای انتخاب نمونه‌های نماینده، یادگیری تطبیقی پارامترهای α و u، و ارزیابی روش روی مجموعه‌داده‌های بزرگ‌تر و متنوع‌تر می‌تواند مسیر پژوهش‌های آینده باشد.

منابع

1. Rousseeuw, P. J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics, vol. 20, pp. 53–65.
2. Davies, D. L., & Bouldin, D. W. 1979. A cluster separation measure," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-1, no. 2, pp. 224–227.
3. MacQueen, J. B. 1967. Some methods for classification and analysis of multivariate observations, in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, Berkeley, CA, USA: University of California Press, 1967, pp. 281–297.
4. Lloyd, S. P. 1982. Least squares quantization in PCM, IEEE Transactions on Information Theory, vol. 28, no. 2, pp. 129–137.
5. Arthur, D., & Vassilvitskii, S. 2007. k-means++: The advantages of careful seeding, in Proceedings of the Eighteenth Annual ACM–SIAM Symposium on Discrete Algorithms (SODA), New Orleans, LA, USA, pp. 1027–1035.
6. Dempster, A. P. 1967. Upper and lower probabilities induced by a multivalued mapping, The Annals of Mathematical Statistics, vol. 38, no. 2, pp. 325–339.
7. Shafer, G. 1976. A Mathematical Theory of Evidence. Princeton, NJ, USA: Princeton University Press.
8. Jousselme, A.-L., Grenier, D., & Bossé, É. 2001. A new distance between two bodies of evidence, Information Fusion, vol. 2, no. 2, pp. 91–101.
9. Zhu, A., Hua, Z., Shi, Y., Tang, Y., & Miao, L. 2021. An improved k-means algorithm based on evidence distance," Entropy, vol. 23, no. 11, Art. no. 1550.
10. Lopes de Souza, V. F., Bakhti, K., Ramdani, S., Mottet, D. & Imoussaten, A. 2025. Explainable evidential clustering, arXiv preprint arXiv:2507.12192.
11. Adamyan, A., Hovanesyan, H., Radrigan, D., Baloian, N., & Harutyunyan, A.2025. Interpretable clustering using Dempster–Shafer theory, Journal of Universal Computer Science, vol. 31, no. 9, pp. 980–1003, 2025.
12. Gan, H., Yang, Z., & Zhou, R. 2023. Safe semi-supervised clustering based on Dempster–Shafer evidence theory, Engineering Applications of Artificial Intelligence, vol. 123, Art. no. 106334.
13. Hubert, L., & Arabie, P. 1985. Comparing partitions, Journal of Classification, vol. 2, no. 1, pp. 193–218.