پژوهش‌های نوین در تصمیم‌گیری
دوره ۲، شماره ۳، پاییز ۱۳۹۶

ارائه یک شاخص جدید اعتبار خوشه‌بندی بر مبنای کاردینالیته فازی

محمود دهقان نیری *

استادیار، گروه مدیریت صنعتی، دانشکده مدیریت و اقتصاد، دانشگاه تربیت مدرس، تهران، ایران

پذیرش: ۱۳۹۵/۱۰/۸ دریافت: ۱۳۹۶/۲/۲

چکیده

بسیاری از روش‌های خوشه‌بندی مستلزم تعیین تعداد خوشه‌های مورد جستجو می‌باشند. به مسئله تعیین تعداد خوشه‌های مناسب در خوشه‌بندی، مسئله اعتبار خوشه‌بندی می‌گویند. تخمین تعداد خوشه‌های بهینه از مهم‌ترین موضوعات مدنظر متخصصان خوشه‌بندی در سال‌های اخیر بوده و منجر به معرفی شاخص‌های اعتبار زیادی شده است. پیشرفته‌ترین این شاخص‌ها مبتنی بر تحلیل همزمان دو معیار میزان فشردگی(تراکم) درون خوشه‌ها و میزان جدایی خوشه‌ها از یکدیگر می‌باشد که عمدتاً درنتیجه عدم کارایی محاسباتی و پیچیدگی ریاضی ناکارآمد می‌شوند. به‌منظور رفع این کاستی، مقاله حاضر به پیشنهاد شاخص FCI که از مفهوم کاردینالیته در مجموعه‌های فازی بهره می‌برد، پرداخته است. این شاخص علاوه بر در نظر گرفتن همزمان دو معیار تراکم و جدایی، از کارایی محاسباتی بالایی برخوردار بوده و به‌دوراز تکلف ریاضی، با استفاده از کاردینالیته در خوشه‌بندی فازی به تعیین تعداد بهینه خوشه‌ها می‌پردازد. در این مقاله علاوه بر مرور شاخص‌های اعتبار خوشه‌بندی، به تشریح شاخص پیشنهادی FCI پرداخته شده و درنهایت، به‌منظور تبیین اثربخشی و کارایی شاخص، از مثال عددی استفاده شده است.

واژگان کلیدی: شاخص اعتبار خوشه‌بندی، خوشه‌بندی فازی، کاردینالیته مجموعه فازی، فشردگی و جدایی خوشه‌ها.

---

* نویسنده مسئول مقاله: E-mail: Mdnayeri@modares.ac.ir

۱- مقدمه

خوشه‌بندی یک روش یادگیری بدون نظارت[۱] است [۱] که به‌طور گسترده به‌عنوان یک روش شناسایی الگو[۲] استفاده می‌شود. بعد از معرفی نظریه مجموعه‌های فازی توسط پروفسور لطفی زاده [۲]، الگوریتم‌های خوشه‌بندی از این نظریه برای اختصاص دادن هر داده با درجاتی از عضویت به هر خوشه به‌جای اختصاص دادن آن فقط به یک خوشه استفاده کردند. دان[۳] روش خوشه‌بندی فازی C-میانگین(FCM)[۴] را برای خوشه‌بندی فازی داده‌ها ارائه [۳] و سپس بزدک[۵] آن را گسترش داد [۴، ۵]. روش خوشه‌بندی فازی C-میانگین شناخته‌شده‌ترین روش در خوشه‌بندی بوده و ازلحاظ محاسباتی قدرت و کارایی بالایی دارد [۶]. این الگوریتم نیازمند آن است که تعداد خوشه‌های (C) از قبل تعیین شود، اما در اغلب موارد تعداد خوشه‌ها قابل‌شناسایی نیستند [۷]؛ مسئله پیدا کردن تعداد بهینه خوشه‌ها اعتبار خوشه نامیده می‌شود[۴]. لذا از مهم‌ترین مسائل در خوشه‌بندی، انتخاب تعداد خوشه‌های مناسب است. در این خصوص تعداد خوشه‌ای مناسب است که اولاً نمونه‌های موجود در یک خوشه تا حد امکان شبیه یکدیگر بوده و ثانیاً نمونه‌های متعلق به خوشه‌های متفاوت تا حد امکان با یکدیگر نامشابه باشند. به‌بیان‌دیگر، خوشه‌ها باید بیشینه فشردگی را در درون خود داشته و تا حد امکان از یکدیگر جدا باشند [۸]. برای خوشه‌بندی مناسب هر دو معیار اهمیت دارند؛ چراکه اگر تنها معیار فشردگی مورداستفاده قرار گیرد در آن صورت هر داده می‌تواند به‌صورت یک خوشه در نظر گرفته شود و این بدان خاطر است که هیچ خوشه‌ای فشرده‌تر از خوشه با یک داده نمی‌باشد و اگر تنها معیار جدایی در نظر گرفته شود در آن صورت بهترین خوشه‌بندی، خوشه‌بندی است که کل داده‌ها در یک خوشه قرار بگیرند. با این توضیح که فاصله هر خوشه از خودش صفر است؛ بنابراین ضروری است ترکیب دو معیار فشردگی و جدایی به‌طور همزمان مدنظر قرار گیرد [۹]. جهت به دست آوردن مناسب‌ترین تعداد خوشه، بررسی‌های زیادی توسط محققین حوزه داده‌کاوی صورت گرفته است که در این میان می‌توان به پژوهش‌های انجام‌شده توسط بزدک [۴] و فوکویاما و سوگنو [۱۰]، زی و بنی [۱۱]، کیوون [۱۲]، وانگ و ژانگ [۱۳] و زالیک و زالیک [۱۴] اشاره کرد. در بسیاری از این مطالعات، شاخص‌های اعتبار خوشه‌بندی معرفی‌شده از تمامی اطلاعات موجود در خوشه استفاده نکرده و تنها معیار فشردگی و یا حداکثر جدایی

خوشه‌ها را در نظر می‌گیرند که ممکن است منجر به نتایج گمراه‌کننده شود. در مقاله حاضر از مفهوم کاردینالیته فازی در مجموعه‌های فازی، برای ارائه یک شاخص جدید (FCI) استفاده شده است. مفهوم کاردینالیته بیانگر میزان درجات عضویت عناصر متعلق به خوشه بوده و به‌نوعی می‌تواند انعکاس‌دهنده میزان تراکم خوشه (مجموعه فازی حاصل از خوشه‌بندی فازی) شود. در ادامه، ابتدا الگوریتم خوشه‌بندی فازی (FCM) و سپس عمده شاخص‌های اعتبار خوشه‌بندی موجود در ادبیات موضوع، تحلیل و درنهایت، شاخص پیشنهادی این پژوهش (FCI) تشریح شده و در قالب مثال عددی به کار گرفته می‌شود.

۲- ادبیات موضوع

۲-۱- الگوریتم‌های خوشه‌بندی فازی

روش‌های خوشه‌بندی را می‌توان با توجه به چگونگی تخصیص داده‌ها به خوشه‌ها، بررسی نمود. در تحلیل خوشه‌ای کلاسیک، هر واحد داده باید دقیقاً به یک خوشه تخصیص داده شود. این روش‌های کلاسیک، منجر به تعداد زیادی بخش‌بندی از مجموعه داده نمونه به زیرمجموعه‌های ناتُهی و دوبه‌دو مستقل می‌شود. چنین تخصیص مقیدی (سخت)[۶] از داده‌ها به خوشه‌ها، به علت امکان وجود داده‌هایی دارای فواصل یکسان از دو یا چند خوشه، می‌تواند نامناسب باشد. چنین داده‌های خاصی می‌توانند نشان دهنده موارد ترکیبی و یا هیبریدگونه‌ای باشند که به میزان یکسانی با دو یا چند گروه شباهت دارند. بخش‌بندی سخت[۷] (شدید و مقید) منجر به تخصیص کامل یک چنین داده‌هایی به یکی از خوشه‌ها می‌شود، درحالی‌که باید به‌طور مساوی به‌تمامی آن خوشه‌هایی که به آن‌ها شباهت دارند، تعلق داشته باشند. برای معرفی کامل‌تر و بهبود دادن جواب‌های خوشه‌ای، برخی از این روش‌های تحلیل خوشه قطعی[۸]، امکان همپوشانی خوشه‌ها را نیز فراهم می‌آورند. لذا هر واحد داده باید به حداقل یک خوشه تخصیص داده شود، اما امکان تخصیص همزمان به خوشه‌های بیشتر نیز فراهم است. اگرچه در این روش‌ها عضویت در چندین خوشه ممکن است، اما تخصیص سختِ داده‌ها در روش‌های خوشه‌بندی سنتی، قطعیت یا عدم قطعیت را در مورد تخصیص داده‌ها به کلاس‌های مختلف منعکس نمی‌کند [۱۵]. لذا تحلیل خوشه‌ای فازی[۹] امکان عضویت نسبی داده‌ها به خوشه‌ها را در قالب درجه

عضویت[۱۰] در بازه [0,1] فراهم می‌آورد. این امر، انعطاف‌پذیری لازم جهت تخصیص داده به بیش از یک خوشه به‌طور همزمان را ایجاد می‌کند. به‌علاوه، این درجات عضویت میزان دقیق‌تر و ریزتری از جزئیات مدل داده را ارائه می‌کنند. جدای از تخصیص یک واحد داده به چند خوشه با سهم‌های متفاوت، درجات عضویت می‌توانند چگونگی و میزان ابهام و قطعیت تعلق یک واحد داده به یک خوشه را نیز بیان نمایند. در خوشه‌بندی فازی، عضویت داده‌ها به خوشه‌ها فازی شده است تا امکان ایجاد فضاهای جواب با جزئیات بیشتر در قالب خوشه‌های فازی مجموعه‌ای از نمونه‌های مفروض $X = \{\overrightarrow{x_1}. \dots . \overrightarrow{x_n}\}$ فراهم آید. نظر به اینکه خوشه‌های C_i از داده‌ها، زیرمجموعه‌های کلاسیک در روش‌های سنتی بوده‌اند، اکنون به‌وسیله مجموعه‌های فازی $C^i_{(x)}$، خوشه‌های فازی از مجموعه داده X بیان شده‌اند. مطابق نظریه مجموعه فازی، u_{ij} بیانگر درجه عضویت یک واحد $\overrightarrow{x_j}$ به خوشه C^i می باشد $u_{ij} = C^i_{(\vec{x}_j)} \in [0.1]$. ازآنجایی‌که عضویت در خوشه‌ها فازی است یک عنوان (برچسب) منحصربه‌فرد که بیانگر شماره خوشه‌ای که داده به آن تعلق دارد، موجود نیست. در عوض، روش‌های خوشه‌بندی فازی، یک بردار برچسب فازی به هر داده $\overrightarrow{x_j}$ تخصیص می‌دهد که بیانگر عضویت آن در خوشه‌های c است.

$$\overrightarrow{u_{ij}} = (u_{1j}. \dots . u_{cj})^T$$

ماتریس U را، ماتریس بخش‌بندی فازی c×n می‌نامند که در آن c ارائه‌کننده تعداد خوشه‌ها(کلاسترها) و n تعداد داده‌ها را نشان می‌دهند. لذا به ازای هر داده به تعداد خوشه‌های تولیدشده و خاص هر خوشه درجه عضویتی بین ۰ تا ۱ برحسب میزان شباهت داده به آن خوشه توسعه تخصیص می‌یابد [۱۵].

طیف متنوعی از روش‌ها در مجلات و کتاب‌های متعدد با هدف یافتن خوشه‌های فازی موجود ممکن در یک مجموعه نمونه مفروض، پیشنهاد و ارائه شده است. برای نمایش یک چنین زمینه گسترده‌ای از روش‌ها، تمرکز روی ایده‌های بنیادین و اساسی آن‌ها مفید و مؤثر است؛ زیرا این روش‌ها می‌توانند با توجه به اصولی که بر مبنای آن بنا شده‌اند در کلاس‌های مختلف طبقه‌بندی شوند. یکی از این ایده‌ها، روش‌های پیدا کردن الگوی اولیه مناسب در خوشه‌بندی و افرازهای فازی (ارائه خوشه‌بندی‌ها)

با استفاده از ساختار جهانی بهینه‌سازی در قالب یک تابع هدف است. این عملیات خوشه‌بندی می‌تواند به‌عنوان یک مشکل بهینه‌سازی عملی فرموله شود. تابع هدف به هر دو عامل الگوهای اولیه خوشه‌بندی و عضویت داده‌ها در خوشه‌ها، بستگی دارد. این تابع هدف نمی‌تواند مستقیماً بهینه شود و درنتیجه معمولاً یک طرح AO برای بهینه‌سازی گروهی از متغیرها (به‌عنوان‌مثال درجه‌های عضویت) در سایر گروه‌ها (به‌عنوان مثال الگوی اولیه) که ثابت شده و بالعکس، مورداستفاده قرار می‌گیرد. این برنامه تکراری به‌روزرسانی برای دستیابی به بهینه‌سازی جهانی در تابع معیار، تکرار می‌شود. درنتیجه، یک بخش داده‌های فازی و توضیحی از خوشه‌ها که در ارتباط با تابع هدف انتخاب‌شده، بهینه در نظر گرفته‌شده‌اند، برای استفاده‌کننده حاصل می‌شود [۱۵]. در این میان، در ادامه به تشریح چند روش از این رویکرد خواهیم پرداخت که مشهورترین و پرکاربردترین آن‌ها الگوریتم C میانگین است.

۲-۱-۱- الگوریتم خوشه‌بندی c میانگین فازی (Fuzzy C-mean)

یکی از مهم‌ترین و پرکاربردترین الگوریتم‌های خوشه‌بندی، الگوریتم c میانگین است. این الگوریتم که با عنوان الگوریتم خوشه‌بندی K میانگین هم شناخته می‌شود، یک روش خوشه‌بندی بدون نظارت، جهت شناسایی گروه‌های داده با ویژگی‌های مشترک، در یک فضای چند شاخصه توسعه داده شده است [۱۶]. در این الگوریتم نمونه‌ها به c خوشه تقسیم می‌شوند و تعداد c از قبل مشخص شده است.

این الگوریتم در حقیقت توسعه‌یافته الگوریتم K-mean و یا ISODATA است [۱۷]. این الگوریتم به خوشه‌ها اجازه روی هم قرار گرفتن و تداخل را می‌دهد، لذا شرایط لازم برای خوشه‌بندی را ، زمانی که گروه‌ها با مرزهای قطعی از یکدیگر جدا نیستند، فراهم می‌آورد. همچنین خوشه‌های ایجادشده بهینه می‌باشند؛ زیرا که واریانس چندمتغیره درون خوشه‌های حاصل از الگوریتم، کمینه می‌باشد. واریانس کم، نشانه داشتن ویژگی‌های مشترک در دل خوشه‌ها است که به معنی تراکم بالا و کم بودن فاصله بین داده‌ها در فضای ویژگی‌هاست. یک الگوریتم خوشه‌بندی بهینه نقاط متراکم در داده‌ها را به‌عنوان مراکز خوشه و در محل ضعیف شدن این تراکم مرزهای خوشه را تعریف می‌نماید. هدف اصلی از الگوریتم خوشه‌بندی بخش نمودن یک مجموعه داده بزرگ (یک فضای چند شاخصه) به چند خوشه است. فاقد نظارت

بودن نیز نشان از آن داد که داده‌ها بدون داشتن هیچ‌گونه برچسب راهنما و یا نظارتی به درون خوشه‌ها تخصیص داده می‌شوند که برخلاف روش‌های طبقه‌بندی[11] است [۱۸]. الگوریتم FCM فرایندی تکراری را بکار می‌گیرد و از یک تخصیص تصادفی داده‌ها به خوشه‌ها شروع می‌کند و با محاسبه مراکز خوشه‌ها که در حقیقت میانگین موزون مقادیر شاخص‌های داده‌ها در خوشه است، الگوریتم به جریان می‌افتد. این فرایند تخصیص داده‌ها به خوشه‌ها آن‌قدر ادامه می‌یابد تا حداقل فاصله درون خوشه‌ها و حداکثر فاصله بین خوشه‌ها حاصل گردد که در این شرایط الگوریتم به شرایط پایداری رسیده و دیگر تغییری در خوشه‌ها حاصل نمی‌شود و همگرایی الگوریتم FCM نامیده می‌شود [۱۸]. تابع هدف الگوریتم خوشه‌بندی فازی c میانگین در رابطه ۱ ارائه شده است:

$$Min\, J = \sum_{i=1}^{c}\sum_{j=1}^{n}(c_{(x_j)}^{i})^m d_{ij}^2 = \sum_{i=1}^{c}\sum_{j=1}^{n}(u_{ij})^m \left\|x_j - v_i\right\|^2 \qquad (۱)$$

در این رابطه، m یک عدد حقیقی بزرگ‌تر از ۱ است که در اکثر موارد برای m عدد ۲ انتخاب می‌شود. اگر در فرمول فوق m را برابر ۱ قرار دهیم، تابع هدف خوشه‌بندی c میانگین (کلاسیک) غیر فازی به دست می‌آید [۱۹]. در فرمول فوق x_j نمونه (داده) j ام و v_i نماینده یا مرکز خوشه i ام و n تعداد نمونه‌هاست. u_{ij} میزان تعلق نمونه i ام در خوشه j ام را نشان می‌دهد. $\left\|x_j - v_i\right\|$ میزان تشابه (فاصله) نمونه j با (از) مرکز خوشه i ام است که می‌توان از هر تابعی که بیانگر تشابه نمونه و مرکز خوشه باشد استفاده کرد. از روی u_{ij} می‌توان یک ماتریس U را تعریف نمود که دارای c سطر و n ستون می‌باشد و مؤلفه‌های آن، هر مقداری بین ۰ تا ۱ را می‌توانند اختیار کنند. اگر تمامی مؤلفه‌های ماتریس U به‌صورت ۰ و یا ۱ باشند، الگوریتم مشابه c میانگین کلاسیک خواهد بود [۴]. بااینکه مؤلفه‌های ماتریس U می‌توانند هر مقداری بین ۰ تا ۱ را اختیار کنند اما مجموع مؤلفه‌های هریک از ستون‌ها باید برابر ۱ باشد (لذا چنانکه اشاره شد این الگوریتم از نوع احتمالی است) و داریم:

$$\sum_{i=1}^{c} u_{ij} = 1 \quad \forall j \in \{1. \ldots . n\} \qquad (۲)$$

معنای این شرط این است که مجموع تعلق هر نمونه به c خوشه باید برابر ۱ باشد. برای به دست آوردن فرمول‌های مربوط به u_{ij} و v_i، باید تابع هدف تعریف‌شده را کمینه کنیم. با استفاده از شرط فوق و برابر صفر قرار دادن مشتق تابع هدف خواهیم داشت:

$$v_i = \sum_{j=1}^{n} u_{ij}^m x_j \Big/ \sum_{j=1}^{n} u_{ij}^m \quad \text{(۴) رابطه} \quad u_{ij} = \frac{1}{\sum_{l=1}^{c} \left(\frac{d_{ij}}{d_{lj}}\right)^{2/(m-1)}} \qquad (۳)$$

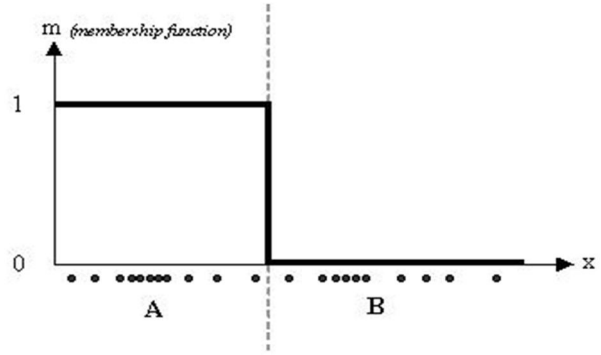
براساس توضیحات اشاره شده در فوق مراحل الگویتم خوشه‌بندی فازی c میانگین به‌صورت زیر خواهد بود[۲۰]:

۱. مقداردهی اولیه برای c، m و U^0. خوشه‌های اولیه حدس زده شوند؛
۲. مراکز خوشه‌ها محاسبه شوند (محاسبه v_iها).
۳. محاسبه ماتریس تعلق از روی خوشه‌های محاسبه‌شده در ۲؛
۴. اگر $\varepsilon \leq \|Ul+1-Ul\|$، الگوریتم خاتمه می‌یابد و در غیر این صورت بازگشت به مرحله ۲.

برای تبیین بهتر عملکرد خوشه‌بندی فازی توزیع یک‌بعدی از نمونه‌های ورودی، نمودار ۱ مفروض است.

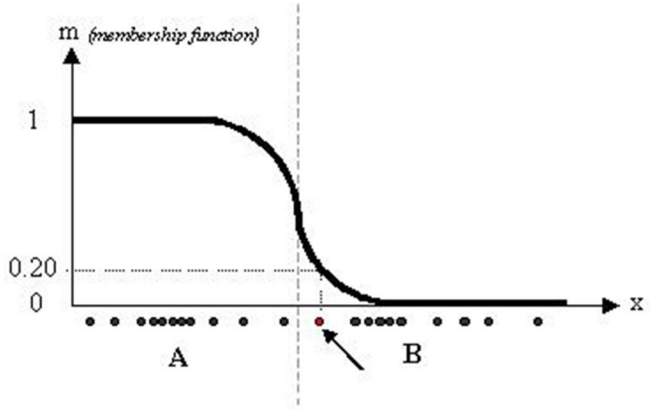
نمودار ۱ توزیع یک‌بعدی نمونه‌ها

اگر از الگوریتم c میانگین کلاسیک استفاده شود، داده‌های فوق به دو خوشه مجزا تقسیم خواهند شد و هر نمونه تنها متعلق به یکی از خوشه‌ها خواهد بود. به‌عبارت‌دیگر، تابع تعلق هر نمونه مقدار ۰ یا ۱ خواهد داشت. نتیجه خوشه‌بندی کلاسیک مطابق نمودار ۲ خواهد بود.



نمودار ۲ خوشه‌بندی کلاسیک نمونه‌های ورودی

نمودار ۲ تابع تعلق مربوط به خوشه A را نشان می‌دهد. تابع تعلق خوشه B متمم تابع تعلق A است. همان‌طور که مشاهده می‌کنید نمونه‌های ورودی تنها به یکی از خوشه‌ها تعلق دارند و به‌عبارت‌دیگر ماتریس U به‌صورت باینری است. حال اگر از خوشه‌بندی فازی استفاده کنیم خواهیم داشت:



نمودار ۳ خوشه‌بندی فازی نمونه‌ها

مشاهده می‌کنید که در این حالت منحنی تابع تعلق هموارتر است و مرز بین خوشه‌ها به‌طورقطع و یقین مشخص نشده است. به‌عنوان‌مثال، نمونه مشخص‌شده با درجه تعلق ۰/۲ به خوشه A و با درجه تعلق ۰/۸ به خوشه B نسبت داده شده است.

۲-۱-۲- الگوریتم خوشه‌بندی c میانگین مبتنی بر آنتروپی

برای الگوریتم خوشه‌بندی c میانگین فازی تابع هدف‌های متفاوتی تعریف شده است. یکی از این توابع با استفاده از مفهوم آنتروپی تعریف شده که در قالب رابطه ۵ ارائه شده است:

$$Min\, J = \sum_{i=1}^{c}\sum_{j=1}^{n}(c_{(x_j)}^{i})d_{ij}^{2} + n\sum_{i=1}^{c}\sum_{j=1}^{n}(c_{(x_j)}^{i})\, log(c_{(x_j)}^{i}) \qquad (۵)$$

در این رابطه n>0 است. جمله اول، تابع هدف الگوریتم c میانگین فازی با مقدار m=1 و جمله دوم رابطه n- برابر تابع آنتروپی فازی است که می‌تواند مقادیری بین ۰ تا 1/c را اختیار کند. اگر خوشه‌ها را مجموعه‌های فازی در نظر بگیریم، تابع آنتروپی فازی عدم قطعیت این که آیا x_j به یک خوشه خاص تعلق دارد یا خیر را می‌رساند [۲۱].

۲-۱-۳- الگوریتم خوشه‌بندی فازی حداکثر شباهت (FMLE)[۱۲]

مهم‌ترین ضعف الگوریتم FCM، با تمام گستردگی کاربرد، این است که مبتنی برفرض کروی[۱۳] بودن خوشه‌ها، آن‌هم با شعاع یکسان است و این در حالی است که بیشتر اوقات خوشه‌ها دارای شکل‌ها و ابعاد متفاوتی هستند. الگوریتم FMLE توسط گاتا و گوا در سال ۱۹۸۹ مطرح شد که بسیار منعطف‌تر بوده و اجازه تشکیل خوشه‌هایی بیضی[۱۴] با حدود متفاوت را می‌دهد. در رابطه زیر عضویت در خوشه با احتمال پیشین $P(c/g_i)$ جایگزین شده که منطبق بر نظریه بیز است. احتمال پیشین مبتنی بر احتمال شرطی $P(g_i/c)$ است و ارائه‌کننده در نظر گرفتن ارزش g در صورت تعلق داشتن به خوشه c می‌باشد. احتمال شرطی به‌عنوان یک تابع تراکم نرمال چندمتغیره[۱۵] در نظر گرفته می‌شود و فاصله اقلیدسی با فاصله ماهالونوبیس[۱۶] که یک فاصله جهت‌دار می‌باشد (رابطه ۶) جایگزین می‌شود.

$$d = \sqrt{(g_i - m_c)^T s_c^{-1}(g_i - m_c)} \qquad (۶)$$

که در آن s_c^{-1} ماتریس کوواریانس فازی است و از طریق رابطه ۷ محاسبه می‌شود.

$$s_c = \frac{\sum_{i=1}^{n} \mu_{ic}(g_i - m_c)(g_i - m_c)^T}{\sum_{i=1}^{n} \mu_{ic}} \quad (۷)$$

بنابراین میزان عضویت داده i به خوشه c به‌مانند رابطه ۸ محاسبه می‌شود.

$$\mu_{ic} = \frac{1}{\sqrt{|s_c|}} \exp\left[-\frac{1}{2}(g_i - m_c)^T s_c^{-1}(g_i - m_c)\right] \frac{n_c}{n} \quad (۸)$$

که در آن n_c، تعداد داده‌ها در خوشه c و همچنین n، تعداد کل داده‌هاست. این نسبت با استفاده از رابطه $\frac{1}{n}\sum_{i=1}^{n} \mu_{ic}$ محاسبه می‌شود. ماتریس کوواریانس و میانگین خوشه، تعیین کننده محل خوشه و حدود بیضی بودن آن، در فضای چند متغیره شاخص‌هاست.لازم است به یاد داشته باشیم الگوریتم FCM و FMLE، هردو ممکن است در مراحل اولیه به دلیل بهینه‌سازی محلی متوقف شوند [۲۳، ۲۲].

۲-۲- اعتبار خوشه‌بندی

به مسئله تعیین تعداد خوشه مناسب[۱۷] برای یک مجموعه داده مسئله اعتبار خوشه‌بندی[۱۸] می‌گویند [۲۴]. هدف، یافتن معیاری برای اعتبار خوشه‌بندی است که تعداد بهینه خوشه‌ها را ارائه نموده و درنهایت ساختار خوشه‌ای مناسب داده‌ها حاصل شود. اغلب معیارهای خوشه‌بندی، به بررسی میزان فشردگی[۱۹] درون خوشه و میزان جدایی[۲۰] بین خوشه‌ها می‌پردازند [۲۵، ۲۶]. طیف متنوع و گسترده‌ای از شاخص‌های اعتبار در ادبیات پیشنهاد شده‌اند [۲۷] که برای بحث مفصل و با جزئیات بیشتر، خواننده می‌تواند به منابع [۲۷، ۲۸] مراجعه نماید.

۲-۲-۱- معیارهای کارایی

اشاره شد که یکی از مهم‌ترین مسائل در خوشه‌بندی، انتخاب تعداد خوشه‌های مناسب است. تعداد خوشه‌ای مناسب است که نمونه‌های موجود در یک خوشه را تا حد امکان شبیه به یکدیگر و نمونه‌های متعلق به خوشه‌های متفاوت را تا حد امکان با یکدیگر نامشابه قرار دهد. عبارات فوق را بدین‌صورت نیز بیان می‌کنند که خوشه‌ها باید بیشینه فشردگی را درون خود داشته باشند و تا حد امکان از یکدیگر جدا باشند [۹]. اگرچه در روش c میانگین فازی و نسخه‌های مختلف آن، تعداد

خوشه‌ها از قبل مشخص شده است، ولی در ابتدای کار، تعداد خوشه‌ها برای طراح مشخص نیست و بیشتر با روش سعی و خطا تعداد مناسب خوشه‌ها تعیین می‌شود. تعیین میزان موفقیت و درستی خوشه‌های حاصل در طبقه‌بندی بدون نظارت (خوشه بندی) بسیار دشوار است [۲۳]. اغلب هیچ‌گونه اطلاعات قبلی و مرجعی برای ارزیابی میزان موفقیت وجود ندارد؛ لذا تنها می‌توان با استفاده از معیارهای آماری به بررسی و تعیین تعداد خوشه‌های مناسب بسنده نمود. برای مشخص کردن تعداد درست خوشه‌ها توابع ارزیابی مختلفی تعریف شده است که می‌توان با استفاده از آن‌ها تعداد خوشه‌ها را برای مسائل مختلف مشخص کرد. در خوشه‌بندی کلاسیک، معیارهای مورداستفاده عبارتند از: شاخص DB [۲۹]، شاخص دان [۱۶] و شاخص کالینسکی[۲۱] [۱۸]. در ادامه به شرحی مختصر از این موضوع و بررسی تعدادی از شاخص‌های معرفی‌شده در ادبیات موضوع می‌پردازیم و در بخش روش‌شناسی یک دیدگاه بسیار ساده، کارآمد و کارا را برای تعیین تعداد خوشه‌ها که نسبت به سایر شاخص‌های ارائه‌شده از سهولت استفاده بیشتری برخوردار است، پیشنهاد می‌شود.

تابع ضریب بخش‌بندی (PC)

یکی از ابتدایی‌ترین شاخص‌های تعیین اعتبار خوشه‌بندی فازی، شاخص PC می‌باشد که در این شاخص تنها به میزان فشردگی توجه شده است (رابطه ۹).

$$V_{PC}(U) = \frac{1}{n}(\sum_{i=1}^{c}\sum_{j=1}^{n}(c_{(x_j)}^{i})^2) = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}(u_{ij})^2 \qquad (۹)$$

انتخاب تعداد خوشه‌های مناسب با بیشینه کردن تابع فوق به دست می‌آید. یعنی برای تعداد خوشه‌های مختلف، خوشه‌بندی را اجرا می‌کنند و با استفاده از ماتریس تعلق به دست آمده، مقدار تابع فوق را محاسبه می‌کنند. تعداد خوشه‌هایی که این تابع به ازای آن بیشترین مقدار را داشته است، به‌عنوان تعداد خوشه‌های مناسب برای آن مسئله مورد استفاده قرار می‌گیرد. مقدار تابع فوق بین ۱/c و ۱ است که هرچه این مقدار به ۱ نزدیک‌تر باشد خوشه‌بندی ما بهتر است[۳۰].

تابع آنتروپی بخش‌بندی(PE)

در رابطه زیر، تابع آنتروپی بخش‌بندی ارائه شده است. در این تابع از ضرب درجات عضویت داده‌ها به خوشه‌ها در لگاریتمشان به‌جای توان دوم استفاده می‌شود.

$$V_{PE}(U) = -\frac{1}{n}\left(\sum_{i=1}^{c}\sum_{j=1}^{n}(c_{(x_j)}^{i})\log(c_{(x_j)}^{i})\right) = -\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}(u_{ij})log(u_{ij}) \qquad (۱۰)$$

انتخاب تعداد خوشه‌های مناسب با کمینه کردن تابع فوق به دست می‌آید. تعداد خوشه‌هایی که این تابع به ازای آن، کمترین مقدار را داشته است، به‌عنوان تعداد خوشه‌های مناسب برای مسئله مورد استفاده قرار می‌گیرد. مقدار این تابع بین ۰ تا $\log_2 c$ است. یک حالت دیگر نیز از این تابع تعریف شده است که به تابع بخش‌بندی آنتروپی نرمال‌شده معروف است. در این تابع، مقدار تابع ارزیابی فوق را بر لگاریتم تعداد خوشه‌ها (c) تقسیم می‌کنند.

نکته قابل‌توجه در مورد دو تابع معرفی‌شده در بالا این است که زمانی که PC برابر ۱ باشد PE برابر ۰ خواهد بود و در این حالت، خوشه‌بندی معادل خوشه‌بندی کلاسیک است. اگر PC برابر 1/c باشد، PE برابر log2c خواهد بود که در این حالت، خوشه‌بندی در فازی‌ترین حالت خود خواهد بود. از طرف دیگر، گفته شد که برای رسیدن به حالت خوشه‌بندی مطلوب، PC باید بیشینه و PE کمینه شود؛ بنابراین در خوشه‌بندی‌های فازی سعی می‌شود تا خوشه‌ها به خوشه‌های کلاسیک نزدیک‌تر باشند و نمونه‌ها با تعلق زیاد به خوشه‌ها نسبت داده شوند[۹].

هر دو معیار PC و PE با تعداد خوشه‌ها رابطه دارند و ازاین‌رو با افزایش تعداد خوشه‌ها، PC افزایش یافته و PE کاهش می‌یابد و این امر ممکن است محقق را به خطا دچار نماید لذا دیو[۲۲] پیشنهاد اصلاح شاخص PC را به شکل رابطه (۱۱) داد[۲۴]:

$$V_{\mathrm{MPC}} = 1 - \frac{\mathrm{C}}{\mathrm{C}-1}(1 - \mathrm{PC}) \qquad (۱۱)$$

که در آن MPC برای یافتن بهترین تعداد خوشه‌ها باید بیشینه شود.

نقاط ضعف موارد فوق این است که از خود داده‌ها به‌طور مستقیم برای ارزیابی خوشه‌بندی استفاده نشده است. در توابعی که در ادامه معرفی می‌شوند، خود نمونه‌ها نیز در تعریف تابع ارزیابی آمده‌اند.

تابع FS

یکی از دیگر شاخص‌های مورداستفاده در این حوزه، شاخص FS می‌باشد که توسط فوکایاما و سوگنو[۲۳] در سال ۱۹۸۹ ارائه شده است [۱۰]. نحوه محاسبه این شاخص به شرح رابطه ۱۲ است.

$$V_{FS} = (\sum_{i=1}^{c}\sum_{j=1}^{n}(c_{(x_j)}^{i})^m \left(\left\|x_j - v_i\right\|^2 - \|v_i - \bar{v}\|^2\right) \qquad (۱۲)$$

در تابع فوق، $\bar{v}$ میانگین کل نمونه‌ها است و انتخاب تعداد خوشه‌های مناسب با کمینه کردن تابع فوق به دست می‌آید. تعداد خوشه‌هایی که این تابع به ازای آن کمترین مقدار را داشته است، به‌عنوان تعداد خوشه‌های مناسب برای مسئله مورد استفاده قرار می‌گیرد. جمله اول در تابع فوق، معیاری برای فشردگی خوشه‌ها و جمله دوم، معیاری برای جدایی خوشه‌ها از هم است. هرچه خوشه‌ها فشرده‌تر باشند، جمله اول کوچک‌تر خواهد بود و هرچه جمله دوم بزرگ‌تر باشد، جدایی خوشه‌ها بیشتر است. بنابراین کمینه کردن تابع فوق می‌تواند معیار مناسبی برای ارزیابی خوشه‌بندی و تعداد خوشه‌ها باشد.

شاخص XB

از مزایای این معیار می‌توان به در نظر گرفتن همزمان هر دو معیار فشردگی و جدایی اشاره نمود. این شاخص توسط زی و بنی (۱۹۹۱)[۲۴] ارائه شد [۱۱]. شاخص XB از رابطه ۱۳ محاسبه می‌شود.

$$XB = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m}\left\|x_j - v_i\right\|^2}{n \times min_{i\neq j}\|x_i - x_j\|} \qquad (۱۳)$$

که در آن x_j بردار ویژگی‌های داده j ام و v_i مرکز خوشه i ام است. صورت کسر فوق نشانه میزان فشردگی بخش‌بندی فازی و مخرج کسر نشانه میزان جدایی بین

خوشه است. در اینجا نیز هدف کمینه نمودن شاخص XB است. لذا انتخاب تعداد خوشه‌های مناسب با کمینه کردن تابع فوق به دست می‌آید و تعداد خوشه‌هایی که این تابع به ازای آن کمترین مقدار را داشته است، به‌عنوان تعداد خوشه‌های مناسب برای مسئله مورد استفاده قرار می‌گیرد. در این رابطه هر چه خوشه‌ها فشرده‌تر باشند، صورت کسر کوچک‌تر خواهد بود و هرچه مخرج کسر بزرگ‌تر باشد، جدایی خوشه‌ها بیشتر است. بنابراین کمینه کردن تابع فوق می‌تواند معیار مناسبی برای ارزیابی خوشه‌بندی و تعداد خوشه‌ها باشد [۱۳].

شاخص PBMF

این شاخص بسیار به شاخص XB نزدیک است، لکن این شاخص نیز هر دو معیار فشردگی و جدایی را همزمان در نظر می‌گیرد. این معیار را از طریق رابطه ۱۴ محاسبه می‌نمایند.

$$\mathrm{PBMF} = \frac{1}{c} \times \frac{n \times \max_{ij} \|v_i - v_j\|}{\sum_{j=1}^{n} \sum_{i=1}^{c} u_{ij}^{m} \|x_j - v_i\|} \qquad (۱۴)$$

شاخص PBMF برای تعیین تعداد خوشه بهینه، کمینه می‌شود [۲۶].

شاخص PCAES

این شاخص توسط وو و یانگ معرفی شده است و از تابع نمایی به شرح رابطه ۱۵ استفاده می‌کند [۲۵].

$$V_{PCAES} = \sum_{i=1}^{C} \sum_{j=1}^{n} \frac{u_{ij}^2}{u_M} - \sum_{i=1}^{c} \exp\left(-min_{k \neq i} \left\{ \frac{\|v_i - v_k\|^2}{B_T} \right\}\right) \qquad (۱۵)$$

بخش اول این شاخص، فشردگی را با یک ضریب بخش‌بندی نرمال‌شده اندازه می‌گیرد و بخش دوم یک معیار جدایی به صورت نمایی است که مجموع حداقل فواصل بین مرکز خوشه‌ها را اندازه می‌گیرد.

۳- روش تحقیق

چنانکه اشاره شد در این تحقیق که از نوع تحقیقات توصیفی است، پس از مرور شاخص‌های مطرح در اعتبار خوشه‌بندی، به ارائه یک شاخص خوشه‌بندی نوین پرداخته شده است. این شاخص در خوشه‌بندی فازی کاربرد داشته و می‌تواند فرایند تعیین تعداد خوشه‌ها در این روش را مشابه سایر روش‌های مطرح، لکن با کارایی محاسباتی بیشتر، فراهم نماید. لذا پس از مرور ادبیات مرتبط با خوشه‌بندی فازی و شاخص‌های اعتبار آن در بخش ۲، در ادامه به معرفی شاخص پیشنهادی و به‌کارگیری آن در یک مثال عددی پرداخته شده و درنهایت، در بخش نتیجه‌گیری در خصوص نقاط قوت و ضعف روش پیشنهادی بحث خواهد شد. شاخص پیشنهادی در تمامی مجموعه‌های فازی که درواقع خروجی اصلی خوشه‌بندی فازی است، قابل‌محاسبه است؛ لذا ازنظر امکان‌پذیری، این شاخص به سهولت قابل‌اجراست و به‌منظور تحلیل روایی، شاخص پیشنهادی در قالب مثال با چند شاخص مطرح در ادبیات موضوع به‌طور همزمان تحلیل و مورد تأیید قرار گرفته است. می‌توان از مقایسه روش پیشنهادی با سایر روش‌های موجود با عنوان روایی همزمان این شاخص یاد نمود.

۳-۱- شاخص پیشنهادی FCI

چنانکه اشاره شد، دو معیار میزان تراکم درون هر خوشه[۲۵] و همچنین میزان پراکندگی خوشه‌ها از یکدیگر، دو معیار اصلی مورداستفاده در تمامی شاخص‌های اعتبار خوشه‌بندی برای یک مجموعه داده هستند. هدف از تعیین تعداد خوشه‌های مناسب در خوشه‌بندی فازی، بررسی خوشه‌هایی است که ماهیت مجموعه‌های فازی را دارند، به‌عبارت‌دیگر، هر خوشه یک مجموعه فازی است که در آن هریک از داده‌های مورد تحلیل، با درجه عضویت متفاوتی به آن خوشه تعلق دارند. بر این اساس است که می‌توان با استفاده از مفهوم کاردینالیته در مجموعه‌های فازی، میزان فشردگی و همچنین میزان پراکندگی خوشه‌های فازی را بررسی نمود. در تمامی روش‌های اعتبار خوشه‌بندی فازی، برای محاسبه هر دو معیار فشردگی و

جدایی یا حداکثر فاصله بین خوشه‌ها، ناگزیر از محاسبه فاصله هر داده از مرکز تمامی خوشه‌های حاصل می‌باشیم که این امر نیز مبتنی بر روش‌های متعددی برای محاسبه فاصله انجام‌پذیر است و درنهایت، با در نظر گرفتن درجه عضویت داده مذکور در خوشه مدنظر، معیار فشردگی درون خوشه و دوری از سایر خوشه‌ها محاسبه می‌شود. به‌طور خلاصه می‌توان اشاره نمود که محاسبه فواصل کلیه داده‌ها از مجموعه خوشه‌ها و... ازنظر محاسباتی بسیار وقت‌گیر بوده و با افزایش تعداد داده‌ها در مسائل دنیای واقعی منجر به کاهش کارایی شاخص‌های موجود می‌شود؛ لذا در مقاله با استفاده از مفهوم کاردینالیته به توسعه یک شاخص بسیار ساده برای این امر پرداخته شده است که در ادامه تشریح می‌شود.

در یک مجموعه غیر فازی، کاردینالیته A برابر است با مجموعه تعداد عناصر مجموعه A با فرض اینکه مجموعه جهانی U متناهی باشد. در یک مجموعه فازی، کاردینالیته عبارت است از مجموعه درجات عضویت اعضای مجموعه فازی که از طریق رابطه ۱۶ قابل‌محاسبه است.

$$|A| = \text{A مجموعه} \quad \text{تعداد اعضا} = \sum A(x).\, x \in \text{supp}(A) \qquad (۱۶)$$

کاردینالیته هر خوشه نشان دهنده میزان عضویت داده‌ها در آن خوشه است. لذا می‌توان نتیجه گرفت هرچه میزان کاردینالیته در یک مجموعه فازی (خوشه) بالاتر باشد، نشان دهنده میزان تراکم اعضای آن خوشه است. به‌عبارت‌دیگر، تلاش خوشه‌بندی فازی آن است که عناصر را در خوشه‌هایی قرار دهد که میزان عضویت و تراکم کاردینالیته فازی بیشتری در آن خوشه داشته باشند. از طرف دیگر هرچه میزان اشتراک کاردینالیته بیشتر باشد، نشان دهنده عدم جدایی بین خوشه‌های موردبررسی است. لذا بیشترین میزان کاردینالیته مشترک خوشه‌ها در مخرج کسر رابطه ۱۷ نشانگر عدم جدایی خوشه‌ها از یکدیگر است؛ به‌عبارت‌دیگر هرچه میزان کاردینالیته مشترک خوشه‌ها بیشتر باشد، می‌توان نتیجه گرفت که میزان جدایی در بین خوشه‌ها چندان مطلوب نیست. بر اساس این توضیحات شاخص FCI در قالب رابطه ۱۷ پیشنهاد می‌شود. هرچه میزان شاخص FCI

بزرگ‌تر باشد، نشان دهنده آن است که میزان تراکم درون خوشه‌ها بالاتر (صورت کسر) و میزان جدایی بین خوشه‌ها (اشتراک) در مخرج کسر بیشتر خواهد بود. لذا می‌توان اشاره نمود که در یک مسئله خوشه‌بندی، تعداد خوشه‌هایی که مقدار کسر رابطه ۱۷ را بیشینه نمایند، مناسب‌تر هستند؛ چراکه بنا بر توضیحات اشاره‌شده، مقدار FCI بیشتر، نشانگر تمایز بیشتری بین خوشه‌ها و همچنین فشردگی بیشتر درون خوشه‌ها است.

$$FCI = \frac{Mean\,/C_i/}{Max/C_i \cap C_j/} \quad i.j = 1.2.\ldots.C \quad i \neq j \qquad (۱۷)$$

در این رابطه $|C_i|$ اشاره به کاردینالیته خوشه i ام دارد و مخرج کسر نشان دهنده بزرگ‌ترین کاردینالیته اشتراک دوبه‌دوی خوشه‌ها است. در یک مسئله خوشه‌بندی هر تعداد خوشه که بتواند مقدار FCI بیشتری را حاصل نماید، تعداد خوشه بهینه خواهد بود. معیار FCI بین $+\infty$ در حالتی که خوشه‌های حاصل‌شده کاملاً از یکدیگر متمایز باشند (در این صورت حاصل صورت کسر برابر با ۱ و مخرج کسر برابر با ۰ خواهد بود) و مقدار صفر در حالتی که خوشه‌های حاصل‌شده کاملاً بر هم منطبق باشند (در این صورت متوسط کاردینالیته خوشه‌ها کوچک‌تر از مخرج کسر خواهد بود) تغییر می‌کند ($FCI \in (0.+\infty)$).

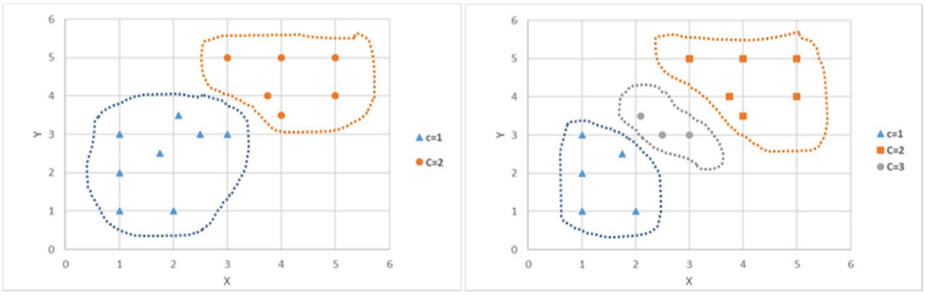
۴- مثال عددی

در ادامه، به‌منظور تشریح بیشتر شاخص FCI و کارایی عملیاتی آن، از یک مثال عددی استفاده شده است. مجموعه داده جدول ۱ را در نظر بگیرید که دارای دو مؤلفه X و Y است. خوشه‌بندی این مجموعه داده بر اساس دو مؤلفه X و Y با در نظر گرفتن دو (C=2) و سه خوشه (C=3) به شرح جدول ۱ خواهد بود. درصورتی‌که این مجموعه داده را در نمای دوبعدی ترسیم نماییم، با نمودار ۴ مواجه می‌شویم.

جدول ۱ مجموعه داده مثال و نتایج خوشه‌بندی آن

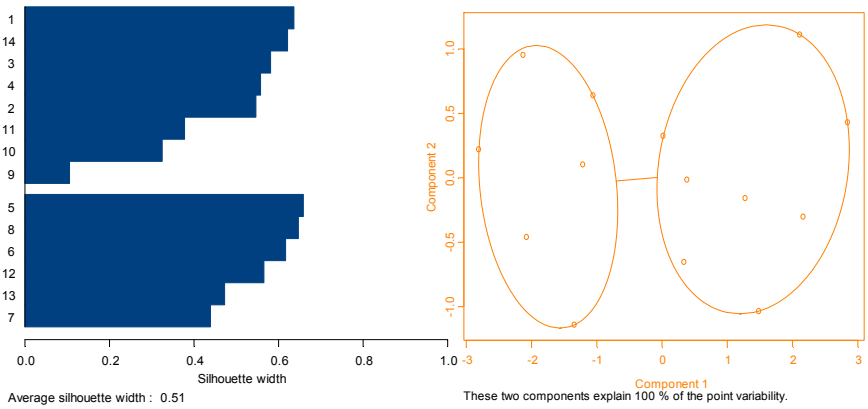
	C=3				C=2		داده‌ها	
خوشه ۳	خوشه ۲	خوشه ۱	خوشه‌بندی قطعی	خوشه ۲	خوشه ۱	خوشه‌بندی قطعی	y	x
۰/۱۶۰	۰/۰۶۸	۰/۷۷۱	۱	۰,۱۴	۰/۸۶	۱	۲	۱
۰/۲۳۰	۰/۱۱۷	۰/۶۵۲	۱	۰,۲۲۳	۰/۷۷۶	۱	۱	۲
۰/۱۶۵	۰/۰۸۷۱	۰/۷۴۷	۱	۰,۲۰۳	۰/۷۹۶	۱	۱	۱
۰/۳۷۶	۰/۱۴	۰/۴۸۳	۱	۰,۱۹۷	۰/۸۰۲	۱	۳	۱
۰/۱۶۱	۰/۷۵۳	۰/۰۸۴۶	۲	۰,۸۴۹	۰/۱۵	۲	۵	۴
۰/۱۸۷	۰/۷۰۴	۰/۱۰۷	۲	۰,۸۱۸	۰/۱۸۱	۲	۴	۵
۰/۳۲۴	۰/۵۱۸	۰/۱۵۶	۲	۰,۷۳۷	۰/۲۶۲	۲	۵	۳
۰/۱۶۸	۰/۷۲۹	۰/۱۰۱	۲	۰,۸۱۵	۰/۱۸۴	۲	۵	۵
۰/۶۹۳	۰/۱۶۲	۰/۱۴۳	۳	۰,۴۷۹	۰/۵۲	۱	۳	۳
۰/۶۹	۰/۱۳۵	۰/۱۷۳	۳	۰,۳۵۳	۰/۶۴۶	۱	۳,۵	۲/۱
۰/۷۹۳	۰/۰۸۶۸	۰/۱۱۹	۳	۰,۳۳۱	۰/۶۶۸	۱	۳	۲/۵
۰/۲۵۴	۰/۶۴۲	۰/۱۰۲	۲	۰,۸۵۶	۰/۱۴۳	۲	۴	۳/۷۵
۰/۳۱۱	۰/۵۵۵	۰/۱۳۲	۲	۰,۷۹۹	۰/۲۰	۲	۳,۵	۴
۰/۴۰۷	۰/۱۰۹	۰/۴۸۳	۱	۰,۱۳۳	۰/۸۶۶	۱	۲,۵	۱/۷۵

با اجرای خوشه‌بندی فازی به روش FCM، با در نظر گرفتن تعداد دو و سه خوشه چنانکه اشاره شد، نتایج جدول ۱ به همراه درجات عضویت هر داده به خوشه متناظر آن حاصل شده است. در این جدول علاوه بر ارائه نتایج خوشه‌بندی فازی، نتایج خوشه‌بندی قطعی متناظر آن نیز ارائه شده است.



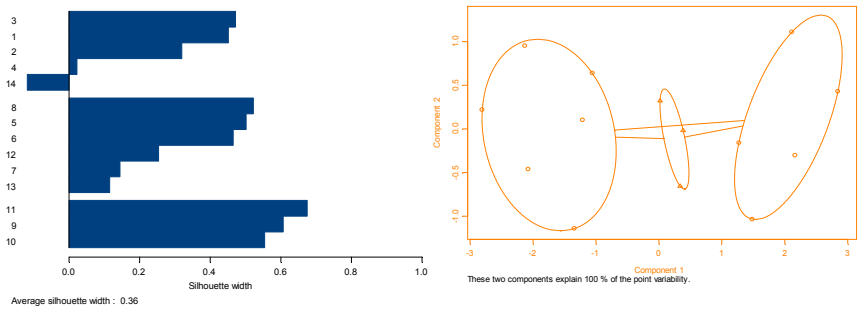
نمودار ۴ ترسیم دوبعدی مجموعه داده مثال

نمودار ۵ نشان دهنده خوشه‌بندی فازی حاصل روی‌داده‌های مثال عددی بوده و شاخص سیلهوات[۲۶] را نیز با در نظر گرفتن دو خوشه که برابر با ۰/۵۱ بوده است، ارائه می‌نماید. شاخص سیلهوات نیز ازجمله دیگر شاخص‌های کارایی خوشه‌بندی است.



نمودار۵ نتیجه خوشه‌بندی با تعداد ۲ خوشه C=2

نمودار ۶ نشان دهنده نتایج خوشه‌بندی فازی با C=3 به همراه شاخص سیلهوات برابر با ۰/۳۶ است.



نمودار ۶. نتیجه خوشه‌بندی با تعداد ۳ خوشه C=3

با توجه به ترسیم دوبعدی خوشه‌ها و همچنین در نظر گرفتن نمودار سیلهوات، می‌توان اشاره نمود که برای این مجموعه داده، در نظر گرفتن دو خوشه مناسب‌تر از سه خوشه است؛ چراکه مقدار متوسط سیلهوات در دو خوشه برابر با ۰/۵۱ بوده درحالی‌که برای ۳ خوشه برابر با ۰/۳۶ است و این امر نشان دهنده مناسب‌تر بودن ۲ خوشه نسبت به ۳ خوشه است. به‌عبارت‌دیگر، هرچه شاخص سیلهوات مقدار بیشتری را نشان دهد، بیانگر مناسب‌تر بودن تعداد خوشه مفروض در مسئله خوشه‌بندی است. به‌طورکلی، می‌توان اشاره نمود که سیلهوات بر اساس میزان نزدیکی و جانشینی خوشه‌ها برای هر داده محاسبه می‌شود و هرچه بیشتر باشد، یعنی تعداد خوشه انتخابی برای خوشه‌بندی داده‌ها مناسب‌تر است. در ادامه، علاوه بر موارد فوق‌الذکر، به بررسی چند معیار دیگر از معیارهای کارایی می‌پردازیم و نتایج را با شاخص FCI مقایسه می‌کنیم.

جدول ۲. مقایسه معیارهای کارایی

تعداد خوشه‌ها	PC	شاخص پیشنهادیFCI	PE	FS
جهت بهینه	Max	Max	Min	Min
C=۲	۰/۶۶	۲/۲	۰/۲۲	۷/۳۰
C=۳	۰/۵۱	۱/۸۹	۰/۳۶	۸/۴۶

چنانکه در جدول ۲ مشاهده می‌شود، معیار پیشنهادی به خوبیِ معیارهای PC و PE و FS، تعداد دو خوشه را انتخاب نموده است. البته ازنظر کارایی محاسباتی، این معیار به‌سادگی معیارهای تک‌بعدی یعنی PC و PE بوده و همچنین قابلیت معیار FS را از جهت در نظر گرفتن همزمان دو شاخص تراکم و جدایی خوشه‌ها از یکدیگر به‌طور همزمان دربر دارد.

۵- نتیجه‌گیری

در این مقاله شاخصی برای تعیین تعداد خوشه مناسب با استفاده از مفهوم کاردینالیته فازی پیشنهاد شده است. این شاخص که FCI نامیده شد، عدم کارایی محاسباتی شاخص‌های پیشین را برطرف نموده و به‌طور همزمان دو شاخص تراکم

و جدایی خوشه‌ها از یکدیگر را در قالب یک رابطه ساده کسری مبتنی بر کارینالیته مجموعه‌های فازی بررسی می‌کند. شاخص FCI، علاوه بر در نظر گرفتن همزمان بعد تراکم و جدایی، در تحلیل قدرت خوشه‌بندی مبتنی بر کاردینالیته فازی خوشه‌ها، بسیار ساده و سریع قابل‌محاسبه بوده و پیچیدگی و حجم محاسباتی شاخص‌های پیشین در محاسبه فواصل بین عناصر از مرکز هر خوشه را ندارد. تعداد محاسبات لازم برای شناسایی فاصله هر عضو خوشه از مرکز آن خوشه در حالت‌های مختلفی که اعضا به خوشه‌ها تخصیص داده می‌شوند، بسیار زیاد بوده و این مهم ضعف اصلی شاخص‌های پیشنهادی اعتبار خوشه‌بندی تاکنون بوده است. شاخص پیشنهادی این مقاله اگرچه این ضعف را برطرف می‌نماید، لیکن خود تنها مبتنی بر خوشه‌بندی فازی قابل‌اجرا بوده و برای مسائل خوشه‌بندی در شرایط قطعی مناسب نیست. لذا از نقاط ضعف شاخص پیشنهادی می‌توان به کاربردی نبودن آن در روش‌های خوشه‌بندی قطعی اشاره نمود. اگرچه گستردگی و قدرت الگوریتم‌های خوشه‌بندی فازی با توجه به تطابق بیشتری که با مفروضات دنیای واقعی دارند، مجالی برای تکنیک‌های قطعی باقی نمی‌گذارند.

لذا استفاده از این شاخص در تعیین تعداد خوشه‌ها در خوشه‌بندی فازی دارای اولویت بوده و می‌تواند به‌سرعت پاسخگوی محققان در تعیین تعداد خوشه و بهبود روند اجرای خوشه‌بندی فازی شود. درنهایت، پیشنهاد می‌شود از شاخص پیشنهادی در تعیین تعداد خوشه‌های مناسب در سایر روش‌های خوشه‌بندی فازی استفاده شده و نتایج با سایر شاخص‌های موجود مقایسه بیشتری شود تا بدین طریق بتوان محاسن شاخص پیشنهادی در اجرا را در کنار نقاط ضعف احتمالی آن شناسایی نمود.

۶- پی‌نوشت‌ها

1. Unsupervised learning
2. Pattern recognition
3. Dunn(1974)
4. Fuzzy c-mean clustering
5. Bezdek (1973)
6. Hard assignment
7. hard partition

8. Crisp
9. Fuzzy Cluster Analysis
10. Membership degree
11. Classification Method
12. Fuzzy Maximum Likelihood Estimation
13. Hyperspherical
14. Ellipsoidal
15. Multivariate Normal Density Function
16. Mahalanobis distance
17. Optimal c
18. Cluster validity
19. Compactness
20. Separation
21. Calinski and Harabasz (1974)
22. Dave (1996)
23. Fukuyama and Sugeno(1989)
24. Xie and Beni(1991)
25. Compactness
26. Silhouette

۷- منابع

[1] Boroufar, A., Rezaian, A., Shokohyar, S.(2017), Identifying the customer behavior model in life insurance Sector using data mining, Management Research in Iran, 20 (4), 65-94.

[2] Zadeh, L. A. (1965). Fuzzy sets, Information and control, 8(3), pp.338-353.

[3] Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions.*Journal of cybernetics*, *4*(1), 95-104.

[4] Bezdek JC. (1973), Fuzzy mathematics in pattern classification, PhD dissertation, Cornell University, Ithaca, NY.

[5] Bezdek, J. C., Coray, C., Gunderson, R., & Watson, J. (1981). Detection and characterization of cluster substructure i. linear structure: Fuzzy c-lines. *SIAM Journal on Applied Mathematics*, *40*(2), 339-357.

[6] De Oliveira, J. V., & Pedrycz, W. (Eds.). (2007). *Advances in fuzzy clustering and its applications*. New York: Wiley.

[7] Zhang, Y., Wang, W., Zhang, X. (2008). A cluster validity index for fuzzy clustering, Information Sciences, 178(4), 1205-1218.

[8] Sohrabi, B., Raeesi, V. I., Zare, M. F. (2016). Designing a Recommender System for Optimizing and Managing Bank Facilities through the Utilization of Clustering and Classification Algorithms, Modern Researches in Decision Making, 1(2), 53-76.

[9] Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of intelligent information systems*, *17*(2-3), 107-145.

[10] Fukuyama, Y., & Sugeno, M. (1989, June). A new method of choosing the number of clusters for the fuzzy c-means method. In *Proc. 5th Fuzzy Syst. Symp* (Vol. 247, pp. 247-250).

[11] Xie, X. L., & Beni, G. (1991). A validity measure for fuzzy clustering. *IEEE Transactions on pattern analysis and machine intelligence*, *13*(8), 841-847.

[12] Kwon, S. H. (2004). Threshold selection based on cluster analysis. *Pattern Recognition Letters*, *25*(9), 1045-1050.

[13] Wang, W., & Zhang, Y. (2007). On fuzzy cluster validity indices. *Fuzzy sets and systems*, *158*(19), 2095-2117.

[14] Žalik, K. R. and Žalik, B.(2010), Validity index for clusters of different sizes and densities, Pattern Recognition Letters, 43(10), 3374 -3390.

[15] Döring, C., Lesot, M. J., & Kruse, R. (2006). Data analysis with fuzzy clustering methods. *Computational Statistics & Data Analysis*, *51*(1), 192-214.

[16] Dunn, J.C., (1973), A fuzzy relative of the isodata process and its use in detecting compact well separated clusters ,J. Cybern, No.28, pp.32–57.

[17] Duda, R. O., Hart, P. E., & Stork, D. G. (2001). Unsupervised learning and clustering. *Pattern classification*, 519-598.

[18] Lucieer, V., & Lucieer, A. (2009). Fuzzy clustering for seafloor classification.*Marine Geology*, *264*(3), 230-241.

[19] Fisher, P., Wood, J.,(1998), "What is a Mountain ? Or the Englishman who went up a Boolean geographical concept but realised it was fuzzy", Geography, No.83, pp.247–256.

[20] Chiu, S.L. (1994), Fuzzy model identification based on cluster estimation, J. Intell. Fuzzy Systems, No. 2,pp.267- 278

[21] Yao, J., Dash, M., Tan, S. T., & Liu, H. (2000). Entropy-based fuzzy clustering and fuzzy modeling. *Fuzzy Sets and Systems*, *113*(3), 381-388.

[22] Gath, I., & Geva, A. B. (1989). Unsupervised optimal fuzzy clustering. *IEEE Transactions on pattern analysis and machine intelligence*, *11*(7), 773-780.

[23] Duda, T., & Canty, M. (2002). Unsupervised classification of satellite imagery: choosing a good algorithm. *International Journal of Remote Sensing*, *23*(11), 2193-2212.

[24] Dave, R.N.(1996), "Validating fuzzy partition obtained through c-shells clustering", Pattern Recognition, No.17, pp.613–623.

[25] Wu, K. L., Yang, M. S. (2005). A cluster validity index for fuzzy clustering.*Pattern Recognition Letters*, *26*(9), 1275-1291.

[26] Pakhira, M. K., Bandyopadhyay, S., & Maulik, U. (2005). A study of some fuzzy cluster validity indices, genetic clustering and application to pixel classification. *Fuzzy Sets and Systems*, *155*(2), 191-214.

[27] Hoppner, F., Klawonn, F., Kruse, R., Runkler, T., 1999. Fuzzy Cluster Analysis.Wiley, Chichester, UK.

[28] Bezdek, J.C., Keller, J.M., Krishnapuram, R., Kuncheva, L.I., Pal, N.R.(1999), Will the Real Iris data please stand up? IEEE Trans. Fuzzy Systems 7, pp.368-369.

[29] Davies, David L.; Bouldin, Donald W. (1979). "A Cluster Separation Measure". IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI-1 (2): 224–227. *doi:10.1109/TPAMI.1979.4766909*

[30] Tsekouras, G. and Haralambos, S.(2004),A new approach for measuring the validity of the fuzzy c-means algorithm, Advances in Engineering Software, No.35,pp.567–575.