فصلنامه علمی ـ پژوهشی مطالعات مدیریت صنعتی سال دهم، شماره ۲۷، زمستان ۹۱
صفحات ۱۷ تا ۳۷

ارائه یک شاخص اعتبار خوشه‌بندی جدید با استفاده از معیار فاصله جاکارد

محمدحسین فاضل زرندی *
سولماز غضنفر اهری**
نادر غفاری‌نسب***

چکیده

تخمین تعداد بهینه خوشه‌ها در دسته‌بندی بدون نظارت داده‌ها، از زمینه‌های چالش برانگیز برای محققان در سالهای اخیر بوده که منجر به ارائه شاخص‌های اعتبار خوشه‌بندی متعدد شده است. این شاخص‌ها اغلب از دو معیار فشردگی و جدایش برای ارزیابی خوشه‌بندی انجام شده استفاده می‌کنند. در این مقاله، یک معیار جدایش جدید برای شاخص اعتبار خوشه‌بندی ECAS که توسط فاضل و همکاران [۱] ارائه شده است، معرفی می‌گردد، که در آن از معیار فاصله جاکارد استفاده شده است. فاصله جاکارد از اندازه اشتراک و اجتماع دو مجموعه فازی استفاده می‌کند. بنابراین اطلاعات بیشتری در مورد هم پوشانی و جدایش خوشه‌ها در اختیار شاخص اعتبار خوشه‌بندی قرار می‌دهد. این قابلیت باعث می‌شود که شاخص جدید در مقابل تغییرات درجه فازی بودن خوشه‌بندی، پایداری بیشتری نسبت به شاخص ECAS داشته باشد. برای مقایسه عملکرد شاخص جدید با ۹ شاخص موجود در ادبیات، از ۱۵ مجموعه داده (۳ مجموعه داده مرسوم و ۱۲ مجموعه داده مصنوعی) به عنوان داده‌های آزمون استفاده شده است. نتایج به دست آمده نشان‌دهنده پایداری و قابلیت بالای شاخص ارائه شده در مقایسه با سایر شاخص‌ها است.

کلید واژگان: شاخص اعتبار خوشه‌بندی، فاصله جاکارد، الگوریتم خوشه‌بندی فازی c- میانگین (FCM)، فشردگی و جدایش نمایی.

---

* استاد دانشکده مهندسی صنایع و سیستم‌های مدیریت، دانشگاه صنعتی امیرکبیر، تهران، ایران. zarandi@aut.ac.ir
** کارشناس ارشد مهندسی مالی دانشگاه صنعتی امیرکبیر، تهران، ایران. solmaz.ahari@gmail.com
*** دانشجوی دکتری مهندسی صنایع دانشگاه علم و صنعت ایران، تهران، ایران. (نویسنده مسئول) ngnasab@iust.ac.ir
تاریخ دریافت: ۹۱/۷/۳ تاریخ پذیرش: ۹۱/۱۰/۱۰

مقدمه

خوشه‌بندی یک روش یادگیری بدون نظارت[1] است که به طور گسترده به عنوان یک تکنیک شناخت الگو[2] استفاده شده است. بعد از معرفی تئوری مجموعه‌های فازی توسط لطفی‌زاده [۲]، الگوریتم‌های خوشه‌بندی از این تئوری برای اختصاص دادن هر داده با درجه‌ای از عضویت به هر خوشه، به جای اختصاص دادن آن فقط به یک خوشه استفاده کردند. دان، روش خوشه‌بندی فازی $c-$ میانگین[3] (FCM) را برای خوشه‌بندی فازی داده‌ها ارائه کرد [۳] و بزدک آن را گسترش داد [۴]. روش خوشه بندی فازی $c-$ میانگین اولین مدلی به شمار می‌رود که از لحاظ محاسباتی کارا و قدرتمند است و به همین علت شناخته‌شده‌ترین رویکرد خوشه‌بندی است که به طور گسترده مورد استفاده قرار گرفته است [۵]. این الگوریتم نیازمند آن است که تعداد خوشه‌ها (c) از قبل تعیین شود، ولی در اغلب موارد این تعداد به راحتی قابل شناسایی نیست [۶]. مساله پیدا کردن تعداد بهینه خوشه‌ها، معمولا مساله اعتبار خوشه نامیده می شود [۷]. به همین منظور مطالعات زیادی توسط محققین مختلف صورت گرفته است که در این میان می‌توان به تحقیقات انجام شده توسط بزدک [۷] [۸]، فوکویاما و سوگنو [۹]، زی و بنی [۱۰]، کیوون [۱۱]، وانگ و زانگ [۱۲] و زالیک [۱۳] اشاره کرد. شاخص‌های اعتبار خوشه‌بندی اغلب از دو معیار فشردگی و جدایش برای ارزیابی خوشه‌بندی انجام شده استفاده می‌کنند [۱۲]. این شاخص‌ها سعی در محاسبه فشردگی و جدایش بین خوشه‌ها و در برخی موارد هم‌پوشانی آن‌ها و ساخت ترکیب مناسبی از آن‌ها برای پیدا کردن مناسب‌ترین خوشه‌بندی دارند. بسیاری از شاخص‌های اعتبار خوشه‌بندی از تمام اطلاعات موجود در مورد شکل خوشه استفاده نمی‌کنند. همین موضوع باعث می‌شود که آن‌ها در برخی موارد، نتیجه درست را به دست ندهند. همچنین برخی از آن‌ها نمی‌توانند در مجموعه داده‌های دارای اغتشاش[4]، تعداد مناسب خوشه‌ها را تشخیص دهند [۱۲] [۶]. برخی از آن‌ها نیز تمام خصوصیات موجود در خوشه‌ها را در نظر نمی‌گیرند، به عنوان مثال شاخص ارائه شده توسط کیم، فشردگی

1- Unsupervised learning
2- Pattern recognition
3- Fuzzy c-mean clustering
4- Noise

خوشه‌ها را در نظر نمی‌گیرد.

در این مطالعه، از فاصله جاکارد که از اندازه اجتماع و اشتراک دو مجموعه فازی برای اندازه‌گیری فاصله آن‌ها استفاده می‌کند، برای معرفی یک معیار فاصله نمایی جدید استفاده می‌شود. این معیار فاصله سپس با یک معیار فشردگی که قبلا توسط فاضل و همکاران ارائه شده است [۱]، ترکیب می‌شود تا یک شاخص جدید برای تعیین تعداد بهینه خوشه‌ها بدست آید. خوشه‌بندی بهینه زمانی به دست می‌آید که معیار فشردگی تا حد امکان بالا و همچنین معیار جدایش تا جایی که ممکن است پایین باشد. شاخص معرفی شده، ECASJ، زمانی بیشترین مقدار خود را دارد که خوشه‌بندی ایجاد شده دارای بیشترین فشردگی و بیشترین جدایش باشد.

در بخش بعد، ابتدا الگوریتم FCM شرح داده می‌شود. سپس تعدادی از شاخص-های اعتبار خوشه‌بندی که در گذشته معرفی شده‌اند مرور و انگیزه استفاده از فاصله جاکارد در معیار جدایش بیان می‌شود. بعد از آن معیار جدایش جدید معرفی و خصوصیات آن توضیح داده می‌شود. در ادامه شاخص جدید بر روی چندین مجموعه داده مرسوم و چندین مجموعه داده مصنوعی آزمایش می‌شود و نتایج آن‌ها با چند شاخصی که قبلا معرفی شده‌اند مقایسه می‌شود. در نهایت مزیت شاخص جدید نسبت به شاخص‌های قبلی مورد بررسی قرار می‌گیرد. در بخش انتهایی مقاله نیز نتایج و پیشنهادهایی برای مطالعات آینده آورده شده‌اند.

پیشینه تحقیق

الگوریتم خوشه بندی فازی c- میانگین

روش خوشه‌بندی FCM سعی دارد مجموعه داده بدون برچسب $X=\{x_1,x_2,\ldots,x_n\}\subseteq R^P$ را (که در آن n و p به ترتیب تعداد داده‌ها و ابعاد داده هستند) به c خوشه تقسیم‌بندی کند، به طوری که تابع هدف $J_m(U,V)=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}\,\|x_j-v_i\|^2$ حداقل شود، که در آن u_{ij} درجه عضویت نقطه x_j در iامین خوشه است. v_i مرکز خوشه iام و $\|x_j-v_i\|^2$ فاصله اقلیدسی بین x_i و v_i است [۱]. الگوریتم خوشه‌بندی FCM که توسط بزدک ارائه شده است به صورت زیر است [۷].

گام ۱. ماتریس پایه $U_{t-1}=[u_{ij}^{(t-1)}]$ (که در ابتدا، $t\leftarrow 1$) را بسازید. ماتریس U نشان

می‌دهد که هر داده x_j با چه درجه‌ای به خوشه c_i تعلق دارد (برای $1\le i\le k$ به طوریکه $\sum_{i=1}^{k} u_{ij} = 1$).

گام ۲. مراکز هسته را با رابطه زیر به روز کنید ($v_t=[v_1^{(t)},\ldots,v_k^{(t)}]$ برای $1\le i\le k$):

(رابطه ۱)

$$v_i^{(t)} = \frac{\sum_{j=1}^{n} (u_{ij}^{(t-1)})^m x_j}{\sum_{j=1}^{n} (u_{ij}^{(t-1)})^m}$$

گام ۳. فاصله بین x_j و $v_i^{(t)}$ را برای $1\le j\le n$, $1\le i\le k$ با استفاده از رابطه زیر محاسبه کنید:

(رابطه ۲) $\|x_j-v_i\|^2=(x_j-v_i)^T(x_j-v_i)$

گام ۴. $U_t=[u_{ij}^{(t)}]$ را با روش زیر به روز کنید (برای هر x_j که $1\le j\le n$) :

(الف) اگر $\|x_j-v_i\|^2>0$، $1\le i\le k$، آنگاه عضویت x_j، در t به صورت زیر محاسبه می شود:

(رابطه ۳)

$$u_{ij}^{(t)} = \left[\sum_{l=1}^{k} \left(\frac{\| x_j - v_i \|^2}{\| x_j - v_l \|^2} \right)^{2/(m-1)} \right]^{-1},$$

گام ۵. اگر $\|U_t-U_{t-1}\|\le\varepsilon$، توقف کنید؛ در غیر این صورت، $t\to t+1$ و به مرحله ۲ بروید.

مروری بر شاخص‌های اعتبار خوشه‌بندی

الگوریتم FCM تعداد خوشه‌ها را به عنوان ورودی دریافت می‌کند. بنابراین قبل از شروع الگوریتم لازم است که تعداد بهینه خوشه‌ها را در یک مجموعه داده بدانیم. به همین منظور شاخص‌های اعتبار خوشه‌بندی زیادی ساخته شده‌اند. در این بخش ما تعدادی از شناخته‌شده‌ترین آنها را مرور می‌کنیم.

شاخص ضریب تقسیم‌بندی[1] (PC) که یک شاخص ساده است به صورت زیر تعریف می‌شود:

(رابطه ۴)

$$V_{PC} = \frac{1}{n} \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^2$$

1- Partition Coefficient

ماکزیمم $V_{PC}(U,c_i,m)$ تعداد مناسب خوشه‌ها را به دست می‌دهد [۴] [۱۴]. اشکال این شاخص این است که نسبت به درجه فازی بودن وابستگی یکنواخت دارد.

دیو[1]، این شاخص را تصحیح کرد و نام آن را ضریب تقسیم بندی تصحیح شده[2] (MPC) گذاشت. این شاخص نسبت به درجه فازی بودن، وابستگی یکنواخت ندارد و به صورت زیر تعریف می شود [۱۵]:

$$V_{MPC} = 1 - \frac{c}{c-1}(1 - V_{PC})$$

همچنین بزدک انتروپی تقسیم‌بندی[3] را معرفی کرد. این شاخص با رابطه زیر تعریف شده است [۷] [۸]:

$$V_{PE} = -\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij} \log u_{ij} \qquad \text{(رابطه ۵)}$$

که در آن، مینیمم $V_{PE}(U,c_i,m)$ تعداد بهینه خوشه‌ها را نشان می‌دهد.

شاخص‌های بسیار دیگری توسط محققان در دو دهه گذشته معرفی شده‌اند، مانند، V_{FS}، V_{XB}، V_{MPC}، V_K، V_{PBMF}، V_W، V_{DB}، V_{PCAES} و V_{ECAS}. بسیاری از آنها درجات عضویت و ساختار هندسی مجموعه داده‌ها را بررسی می‌کنند. مینیمم شش شاخص اول و ماکزیمم دو شاخص آخر، مناسب‌ترین خوشه‌بندی را نشان می‌دهند. برخی از آن‌ها در زیر توضیح داده شده‌اند [۹] [۱۰] [۱۵] [۱۱] [۶] [۱۶] [۲۳] [۱].

V_{FS} در بخش اول خود فشردگی و در بخش دوم جدایش را اندازه می‌گیرد [۹]:

$$V_{FS} = \sum_{i=1}^{c}\sum_{j}^{n} u_{ij}^{m} \| x_j - v_i \|^2 - \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m} \| v_i - \bar{v} \|^2 \qquad \text{(رابطه ۶)}$$

V_{XB}، فشردگی و جدایش کل را اندازه‌گیری می‌کند که به صورت رابطه زیر تعریف شده است [۱۰].

$$V_{XB} = \frac{J_m(U,V)/n}{Sep(V)} = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m} \| x_j - v_i \|^2}{n \min_{i \neq j} \| v_i - v_j \|^2} \qquad \text{(رابطه ۷)}$$

V_{PCAES} توسط وو و یانگ معرفی شده است و از تابع نمایی استفاده می‌کند [۱۶]:

$$V_{PCAES} = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{2} / u_M - \sum_{i=1}^{c} \exp\left(-\min_{k \neq i}\left\{ \| v_i - v_k \|^2 / \beta_T \right\}\right) \qquad \text{(رابطه ۸)}$$

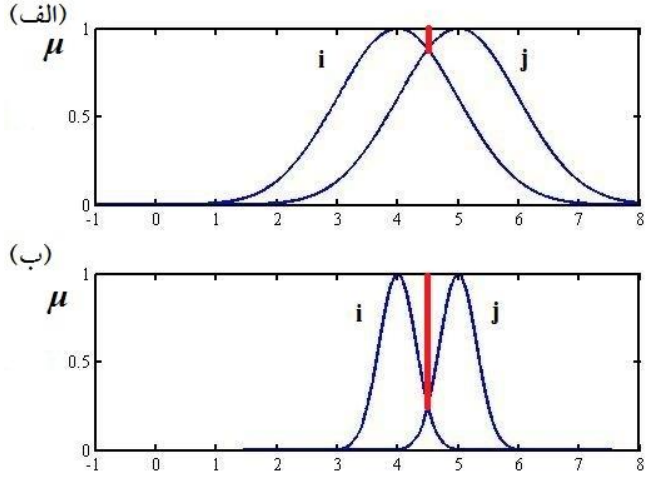
1- Dave
2- Modified Partition Coefficient
3- Partition Entropy

بخش اول V_{PCAES} فشردگی را با یک ضریب تقسیم بندی نرمال شده اندازه می‌گیرد و بخش دوم یک معیار جدایش نمایی است که مجموع حداقل فواصل بین مراکز خوشه‌ها را اندازه می‌گیرد.

V_W که توسط زانگ و همکاران ارائه شده است، شاخص دیگری است که به صورت زیر تعریف شده است [۶]:

$$V_W(U,V)=\frac{Var^N(U,V)}{Sep^N(c,V)}=\left(\frac{Var(U,V)}{Var_{\max}}\right)\Big/\left(\frac{Sep(c,U)}{Sep_{\max}}\right) \qquad \text{(رابطه ۹)}$$

V_W دارای دو بخش است که فشردگی و جدایش خوشه‌ها را بررسی می‌کند. در قسمت جدایش، این شاخص از رابطه $1-\max_{i\neq j}\left[\max_{x_k}\min(u_{ik},u_{jk})\right]$ استفاده می کند که در واقع اندازه خط قرمز رنگ در شکل (۱) است. اگر مقدار $\mathrm{Sep}(c,U)$ بزرگ باشد، تقسیم‌بندی فازی خوبی صورت گرفته است. یعنی خوشه‌ها به خوبی جدا از هم هستند. اما همان‌طور که مشاهده می‌شود، این معیار جدایش، شکل کلی خوشه‌ها را در نظر نمی‌گیرد. در واقع انواع مختلفی از خوشه‌ها می‌تواند وجود داشته باشد که دارای همان مقدار از معیار جدایش باشند.



شکل ۱. در این شکل، درجه عدم مشابهت بین خوشه‌ها با رنگ قرمز نشان داده شده است. (الف) دو خوشه که هم پوشانی آنها زیاد است. (ب) دو خوشه فازی که هم پوشانی آنها کم است.

شاخص دیویس ـ ولدین[1] (V_{DB}) که کاربرد زیادی هم دارد از معیار شباهت بین دو خوشه (R_{ij}) استفاده می کند که بر اساس پراکندگی یک خوشه (s_i) و عدم شباهت بین دو خوشه (d_{ij}) تعریف می شود. شباهت بین دو خوشه به صورت

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \qquad \text{(رابطه ۱۰)}$$

که در آن s_i و d_{ij} به صورت زیر تعریف می شوند:

$$d_{ij} = d(v_i, v_j)$$

$$s_i = \frac{1}{\|c_i\|} \sum_{x \in c_i} d(x, v_i)$$

با توجه به مطالب بیان شده و تعریف شباهت بین دو خوشه شاخص دیویس ـ بولدین به صورت زیر تعریف می‌شود.

$$V_{DB} = \frac{1}{n_c} \sum_{i=1}^{n_c} R_i \qquad \text{(رابطه ۱۱)}$$

که R_i در آن به صورت زیر محاسبه می شود:

$$R_i = \max_{j=1,\ldots,n_c, i \neq j} (R_{ij}), \quad i = 1, \ldots, n_c$$

این شاخص در واقع میانگین شباهت بین هر خوشه با شبیه‌ترین خوشه به آن را محاسبه می‌کند. می‌توان دریافت که هرچه مقدار این شاخص بیشتر باشد، خوشه‌های بهتری تولید شده است. بسیاری از شاخص‌های دیگر نیز در سال‌های اخیر طراحی شده‌اند که هر یک سعی کرده‌اند نواقص شاخص‌های قبلی را برطرف کنند که به علت طولانی بودن روش محاسبه از بیان آنها خودداری می شود [۲۴] [۲۵].

قبل از اینکه به توضیح شاخص معرفی شده بپردازیم، دو شاخص مرتبط با آن را مرور می‌کنیم. اولین آنها ECAS است که توسط فاضل و همکاران ارائه شده و به صورت زیر تعریف شده است [۱].

$$V_{ECAS} = ECAS(c) = \frac{EC_{Comp}(c)}{\max_c(EC_{Comp}(c))} - \frac{ES_{Sep}(c)}{\max_c(ES_{Sep}(c))} \qquad \text{(رابطه ۱۲)}$$

همان‌طور که مشاهده می‌شود، این شاخص فشردگی و جدایش نرمال شده را اندازه

1- Davis-Bouldin Index

می‌گیرد که با روابط زیر محاسبه می‌شود.

$$EC_{Comp}(c)=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}\exp\left(-\left(\frac{\|x_i-v_j\|^2}{\beta_{Comp}}+\frac{1}{c+1}\right)\right)$$

$$ES_{Sep}(c)=\sum_{i=1}^{c}\exp\left(-\min_{i\neq k}\left\{\frac{(c-1)\|v_i-v_k\|^2}{\beta_{Sep}}\right\}\right)$$

معیار جدایش این شاخص کوچکترین فاصله اقلیدسی بین مراکز خوشه‌ها را در نظر می‌گیرد. ولی هم پوشانی آنها را در نظر نمی‌گیرد.

V_{OS} شاخص اعتبار خوشه‌بندی دیگری است که توسط کیم و همکاران معرفی شده و تلاش می‌کند هم پوشانی و جدایش خوشه‌ها را اندازه گیری کند [۱۷]. معیار جدایش آن دقیقا مانند معیار جدایش V_W است، اما معیار هم پوشانی آن درجه هم پوشانی بین خوشه‌های فازی را با محاسبه هم پوشانی بین خوشه‌ها اندازه‌گیری می‌کند. این شاخص وقتی از درجات مختلف فازی بودن در الگوریتم FCM استفاده شود، پایدار است. اما این شاخص فشردگی و شکل هندسی خوشه‌ها را در نظر نمی گیرد.

همان‌طور که در بخش قبل توضیح داده شد، معیار جدایش V_W و V_{OS} فقط یک نقطه یعنی ماکزیمم اشتراک دو خوشه را برای محاسبه فاصله بین خوشه‌ها در نظر می‌گیرد. در حالی که معیار جدایش V_{ECAS} فقط فاصله اقلیدسی بین مراکز خوشه‌ها را در نظر می‌گیرد. در این مطالعه شاخص ECAS طوری اصلاح شده است که معیار جدایش آن بتواند شکل هندسی خوشه‌ها و هم پوشانی آنها را در نظر بگیرد. فاصله جاکارد به دلیل درنظر گرفتن شکل خوشه‌ها و هم پوشانی آنها برای محاسبه جدایش خوشه‌ها به کار می‌رود. فاصله جاکارد که عدم شباهت بین مجموعه‌های داده را اندازه می‌گیرد، با کم کردن ضریب جاکارد از ۱، یا به طور معادل با تقسیم تفاضل اندازه اجتماع از اندازه اشتراک، بر اندازه اجتماع دو مجموعه فازی حاصل می‌شود. بنابراین این فاصله که با رابطه (۱۳) تعریف می‌شود، اطلاعات بیشتری در مورد وضعیت خوشه‌ها به دست می‌دهد.

$$J_\delta(A,B)=1-J(A,B)=\frac{|A\cup B|-|A\cap B|}{|A\cup B|} \qquad \text{(رابطه ۱۳)}$$

شاخص ارائه شده جدید

شاخص ارائه شده در این مقاله دارای دو بخش است که فشردگی و جدایش خوشه‌ها را اندازه‌گیری می‌کنند که در بخش فشردگی همانند شاخص ECAS بوده و به صورت زیر تعریف زیر تعریف می‌شود.

$$Comp(c) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{m} \exp\left(-\left(\frac{\| x_i - v_j \|}{\beta_{Comp}} + \frac{1}{c+1}\right)\right) \qquad \text{(رابطه ۱۴)}$$

که، β_{Comp} به صورت کواریانس ماتریس i تعریف شده است. یعنی:

$$\beta_{Comp} = \frac{\sum_{k=1}^{n} \| x_k - \bar{v} \|^2}{n(i)} \quad \text{و} \quad \bar{v} = \frac{\sum_{j=1}^{n} x_j}{n}$$

که در آن، $n(i)$ تعداد داده‌ها در خوشه iام است، که می‌تواند با روش‌های مربوطه اندازه‌گیری شود، (یعنی با استفاده از هر t–نرم در منطق فازی بین درجات عضویت هر داده در c خوشه). توان EC_{Comp} از دو عبارت تشکیل شده است: $||x_i-v_j||^2/\beta_{Comp}$ و $1/c+1$. عبارت اول مربوط به فشردگی بین خوشه‌ها است و عبارت دوم فقط برای تنظیم مقدار معیار فشردگی استفاده شده است. اگر مقدار این معیار فشردگی برای یک مقدار از c بزرگ باشد، نشان دهنده فشرده بودن خوشه‌ها در آن خوشه بندی است [۱].

معیار جدید جدایش ارائه شده

نکته اصلی در معیار جدایش ارائه شده این است که هر خوشه به صورت یک مجموعه فازی، $F_i = \{(x_j, \mu_{F_i}(x_j)) \mid x_j \in X\}$، در نظر گرفته می‌شود که می‌تواند به صورت زیر نشان داده شود:

$$F_i = \sum_{j=1}^{n} \mu_{F_i}(x_j)/x_j = \mu_{F_i}(x_1)/x_1 + \mu_{F_i}(x_2)/x_2 + \ldots + \mu_{F_i}(x_n)/x_n$$

با استفاده از این رویکرد، شاخص ارائه شده از ساختار هندسی کلی بین خوشه‌ها استفاده می‌کند و نشان‌دهنده ساختار داده است. تعریف زیر به عنوان معیار جدایش جدید، ارائه شده است:

Exponential overlap and Separation with Jaccard distance (c) = $EOASJ(c) = \exp\left(-\min_{i\neq k} J_\delta(F_i, F_k)\right) = \exp\left(-\min_{i\neq k}\left(1-\frac{|F_i \cap F_k|}{|F_i \cup F_k|}\right)\right)$

اگر مقدار EOASJ (c) کم باشد، به این معنی است که در این خوشه‌بندی، خوشه‌ها همپوشانی کمی باهم داشته و از هم جدا هستند. این معیار جدایش مشخصات زیر را دارا می‌باشد:

مشخصه ۱: این معیار جدایش از اندازه اجتماع و اشتراک برای اندازه‌گیری فاصله بین دو مجموعه فازی استفاده می‌کند.

مشخصه ۲: معیار جدایش کران‌دار می‌باشد، $0 < EOASJ\ (c) \leq 1$ و دارای مزیت استفاده از تابع نمایی است که در حوزه تحلیل خوشه‌بندی شناخته شده است.

مشخصه ۳: اگر $F_p = F_q$ آنگاه $EOASWJ\ (c) = 0$.

شاخص ارائه شده جدید

شاخص ارائه شده با نام " Compactness and Separation with Jaccard distance (ECASJ)" به صورت زیر تعریف می‌شود:

$$ECASJ\ (c) = Comp^N(c) - \mathrm{Sep}^N(c) = \frac{Comp(c)}{Comp_{\max}} - \frac{EOASJ\ (c)}{Sep_{\max}}$$

به طوری که $Comp_{max} = \max_c Comp(c)$ و $Sep_{max} = \max_c EOASJ\ (c)$. یک مقدار بزرگ برای شاخص ارائه شده، $ECASJ\ (c)$، نشان دهنده خوشه‌بندی مناسب، یعنی خوشه‌های جدا ازهم، فشرده تر و با همپوشانی کمتر است.

توضیحات

(الف) همان‌طور که می دانیم، $(\forall\ z \in R^+\ ,\ 0 < \exp(-z) \leq 1)$ ، بنابراین، $0 \leq Comp(c) \leq \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m$ و $0 < EOASJ\ (c) \leq 1$. در یک مجموعه داده بزرگ، مقدار $Comp(c)$ بسیار بزرگ است و مقدار $Sep(c)$ کوچک است. این مقادیر متفاوت برای اینکه قابل مقایسه باشند باید نرمال شوند. بنابراین ما $Comp(c)$ و $Sep\ (c)$ را به ترتیب بر $Comp_{max}(c)$ و $Sep_{max}(c)$ تقسیم می کنیم.

$$\text{for } c = 2,\ldots,c_{max}\ Comp_{\max}(c) = \max_c(Comp(c))$$

$$\text{for } c = 2,\ldots,c_{max}\ Sep_{\max}(c) = \max_c(EOASJ(c))$$

(ب) هنگامی که c به تعداد داده‌ها نزدیک می‌شود، آنگاه $||x_i - x_j|| \to 0$ ، $n(i) \to 1$ و $\frac{1}{c+1}$

کاهش می‌یابد. بنابراین، توان $Comp(c)$ به سمت صفر نزدیک می‌شود و در نتیجه $Comp(c)$ افزایش می‌یابد. از سوی دیگر، وقتی که $c \to n$، آنگاه هر خوشه فقط شامل یک داده از مجموعه داده‌ها است ، بنابراین خوشه‌ها به هم نزدیک می شوند و اشتراک مجموعه فازی خوشه‌ها افزایش می یابد، بنابراین فاصله جاکارد آنها کم می‌شود و $EOASJ(c)$ افزایش می‌یابد. این موضوع از افزایش شاخص هنگام افزایش تعداد خوشه‌ها جلوگیری می کند.

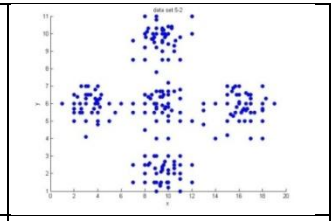
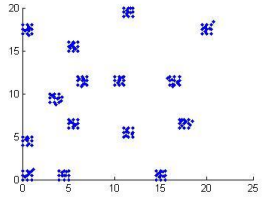
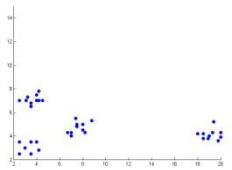
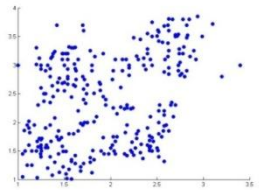
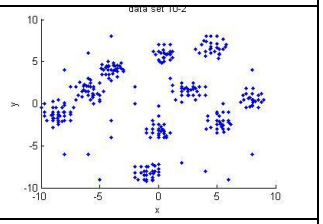
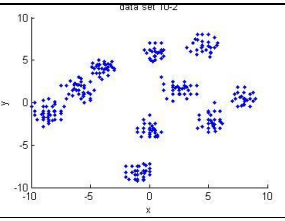
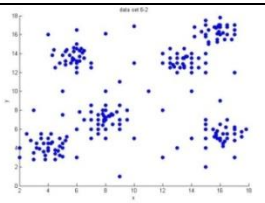
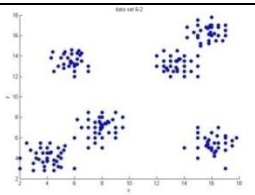
(ج) هر دوی $Comp(c)$ و $Sep(c)$ از مزیت تابع نمایی برای اندازه‌گیری فاصله‌ها استفاده می‌کنند. علت استفاده و مزایای استفاده از تابع نمایی این است که تابع نمایی در برخورد با انتروپی کلاسیک شانون [۱۸][۱۹] و تحلیل خوشه‌بندی [۶][۱۶][۲۰] بسیار خوب عمل می‌کند. مخصوصا وو و یانگ ادعا کرده‌اند که بر اساس تحلیل تابع تاثیر، یک فاصله از نوع نمایی مشخصات پایدارتری را به دست می‌دهد [۱].

(د) با توجه به رابطه (۱۲) مشاهده می‌شود که معیار فشردگی شاخص ECAS با مقدار m رابطه دارد. هر چه m بیشتر شود u_{ij}^m کوچکتر می‌شود زیرا $0<u_{ij}<1$ و $m>1$. اما $\beta_{Comp} = \frac{\sum_{k=1}^{n} \| x_k - \overline{v} \|^2}{n(i)}$ و با افزایش m خوشه‌ها گستردگی بیشتری پیدا می‌کنند بنابراین $n(i)$ بیشتر می‌شود. در نتیجه β_{Comp} کاهش یافته $\exp\left(-\left(\frac{\| x_i - v_j \|}{\beta_{Comp}} + \frac{1}{c+1}\right)\right)$ افزایش می‌یابد. بنابراین افزایش یا کاهش معیار فشردگی که به این دو مقدار وابسته است، بستگی به ساختار داده دارد. اما معیار جدایش در شاخص ECAS به مقدار m وابسته نیست. بنابراین با تغییر مقدار m اندازه معیار فشردگی تغییر می کند ولی اندازه معیار جدایش ثابت است. این شاخص با مقدار $m=2$ همانطور که در مطالعه فاضل و همکاران [۱] بررسی شده، نتایج مناسبی را فراهم می‌کند. این در حالی است که در شاخص معرفی شده در این مقاله، معیار جدایش نیز به مقدار m مرتبط است. زیرا در آن از فاصله جاکارد استفاده شده است که آن نیز به اندازه اشتراک و اجتماع مجموعه‌های فازی خوشه‌ها بستگی دارد. همان‌طور که گفته شد با تغییر m گستردگی خوشه‌ها نیز تغییر می‌کند و کوچکترین فاصله جاکارد بین خوشه‌ها نیز تغییر می‌کند. بنابراین در با تغییر مقدار m هر دو معیار فشردگی و جدایش تغییر می کنند که در نهایت باعث می-شود این شاخص جدید نسبت به تغییر مقدار m پایدار باشد.

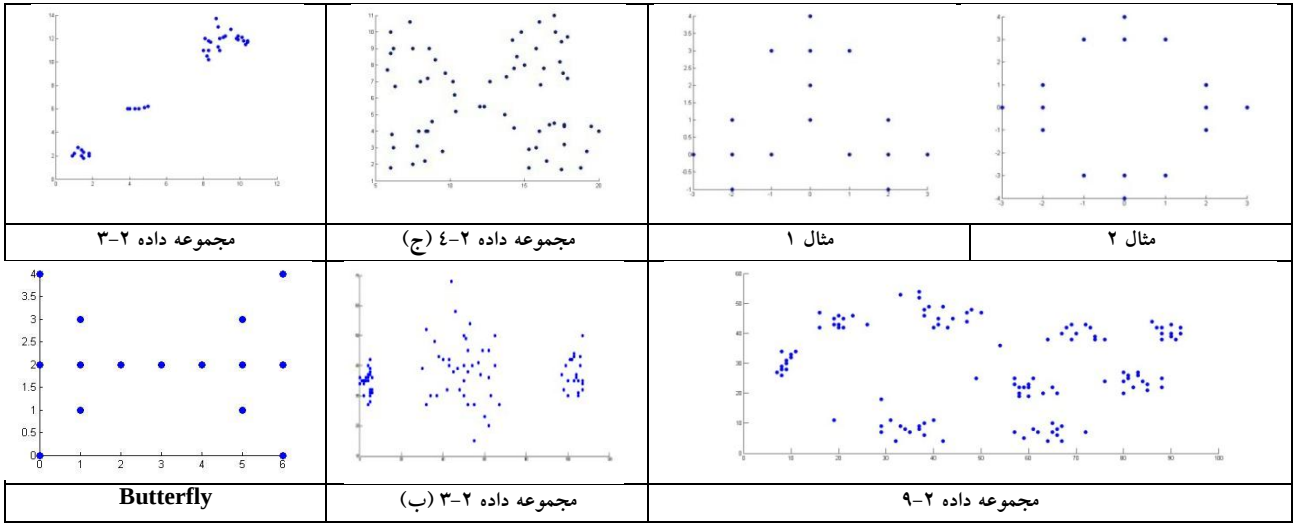
رویه پیدا کردن تعداد بهینه خوشه‌ها به صورت زیر است:
گام اول: مقادیر پارامترهای مربوط به خوشه بندی FCM و شاخص اعتبار را مشخص کنید: $c_{min}=2, c_{max}=\sqrt{n}$ (بر اساس پیشنهاد بزدک[۲۱]) و $\varepsilon=0.00001$.
گام دوم: FCM را مقادیر متفاوت $2\le c\le\sqrt{n}$ و $1.1\le m\le 4$ اجرا کنید.
گام سوم: شاخص را برای هر زوج از c و m محاسبه کنید.
گام چهارم: زوجی را که شاخص در آن بیشترین مقدار را دارد انتخاب کنید.

آزمایش‌ها و نتایج

در این بخش، شاخص ارائه شده روی ۳ مجموعه داده استاندارد و ۱۰ مجموعه داده مصنوعی آزمایش شده است. مجموعه داده‌های مصنوعی قبلا توسط زانگ و همکاران [۶]، فاضل زرندی و همکاران [۱]، کیم و همکاران [۱۷]، وو و یانگ [۱۶] به کار رفته‌اند و عبارتند از مجموعه داده‌های ۲-۴، ۲-۱۵ ، ۲-۵، ۲-۶، ۲-۶ با اغتشاش، ۲-۱۰ و ۲-۱۰ با اغتشاش [۶]، مجموعه داده ۲-۴ (ب) و ۲-۴ (ج) [۱۶]، مجموعه داده ۲-۳ [۱۷] و مجموعه داده ۲-۳ (ب) و ۲-۹ که مشبه مجموعه داده های استفاده شده در مقاله کیم و همکاران [۱۷]. عدد سمت چپ در نام مجموعه داده‌ها نشان دهنده تعداد بهینه خوشه‌ها در آن مجموعه داده و عدد سمت راست ابعاد مجموعه داده است. مجموعه داده‌های مرسوم که در بسیاری از مطالعات خوشه‌بندی به کار رفته‌اند عبارتند از Butterfly، مثال ۱ و مثال ۲ [۱]. تعداد بهینه خوشه‌ها در آن‌ها همان طور که در شکل (۲) دیده می‌شود، ۲، ۳ و ۴ است. شکل (۲) مجموعه داده‌هایی را که در این مطالعه آزمایش می‌شوند، نشان می‌دهد. مجموعه‌های داده قبل از خوشه‌بندی نرمال می‌شوند. نتیجه استفاده از شاخص معرفی شده در این مطالعه با ۸ شاخصی که قبلا طراحی شده‌اند، مقایسه می‌شود.

			
مجموعه داده ۲-۵	مجموعه داده ۲-۴ (ب)	مجموعه داده ۲-۱۵	مجموعه داده ۲-۴(الف)
			
مجموعه داده ۲-۱۰ با اغتشاش	مجموعه داده ۲-۱۰	مجموعه داده ۲- ۶ با اغتشاش	مجموعه داده ۲-۶

شکل ۲. مجموعه داده‌های آزمایش شده در این مطالعه



ادامه شکل ۲. مجموعه داده‌های آزمایش شده در این مطالعه

فاضل زرندی و همکاران [۱] در مطالعه‌ای که انجام دادند، موثر بودن شاخص ECAS را نسبت به شاخص‌های قبلی نشان دادند. اما بسیاری از شاخص‌های قبلی و از جمله ECAS برای تمام مقادیر m (درجه فازی بودن خوشه بندی) مطالعه نشده‌اند. برای مثال V_{PBMF} فقط با m=1.5 جواب بهینه را به دست می‌آورد [۲۲]. شاخص ارائه شده در این مطالعه، در حالی که مقادیر مختلف m به کار رفته‌اند، محاسبه شده و نتایج نشان می‌دهد که نتیجه استفاده از شاخص جدید، در مورد تغییر m پایدار است. جدول (۱) نشان دهنده مقادیر بهینه خوشه‌ها برای مجموعه‌های داده نامبرده است که توسط شاخص‌های مختلف انتخاب شده‌اند. ستون سمت راست این جدول مجموعه‌های داده مورد آزمایش را نشان می‌دهد. ستون دوم تعداد بهینه خوشه‌ها در هر مجموعه داده را مشخص می‌کند. ستون‌های بعدی هر یک نشان دهنده تعداد خوشه‌ها است که توسط شاخص‌انتخاب شده‌اند. دو ستون آخر مربوط به شاخص ECAS و ECASJ است.

جدول ۱. تعداد بهینه خوشه‌های انتخاب شده توسط شاخص‌های مختلف برای مجموعه داده‌های مورد آزمایش

Dataset	c^*	V_{PC}	V_{PE}	V_{FS}	V_{XB}	V_K	V_{PCAES}	V_W	V_{DB}	V_{ECAS}		V_{ECASWJ}
										m=2	m=1.6+t×0.1 t=1,2,…,8	
مجموعه داده ۲-٤(الف)	۴	۲	۲	۱۶	۴	۴	۴	۴	۴	۴	۲	۴
مجموعه داده ۲-۱۵	۱۵	۱۰	۱۷	۱۵	۱۵	۱۵	۱۵	۱۵	۱۱	۱۵	۱۴	۱۵
مجموعه داده ۲-٤(ب)	۴	۲	۲	۴	۲	۴	۴	۴	۳	۴	۴	۴
مجموعه داده ۲-۵	۵	۵	۲	۵	۴	۴	۴	۴	۴	۵	۴	۵
مجموعه داده ۲-٦	۶	۶	۶	۶	۴	۴	۴	۶	۵	۶	۶	۶
مجموعه داده ۲-۱۰	۱۰	۱۰	۲	۱۰	۱۰	۱۰	۴	۱۰	۱۰	۱۰	۸	۱۰
مجموعه داده ۲-۳ (الف)	۳	۳	۳	۴	۴	۲	۴	۴	۳	۳	۳	۳
مجموعه داده ۲-٤ (ج)	۴	۴	۲	۴	۴	۲	۴	۴	۲	۶	۴	۴
مجموعه داده ۲-٦ با وجود اغتشاش	۶	۶	۲	۶	۴	۴	۴	۶	۵	۶	۶	۶
مجموعه داده ۲-۱۰ با وجود اغتشاش	۱۰	۲	۲	۱۰	۷	۷	۴	۱۰	۱۰	۱۰	۷	۱۰
Butterfly	۲	۲	۲	۲	۲	۲	۲	۲	۲	۲	۲	۲
مثال ۱	۳	۳	۳	۳	۳	۳	۳	۳	۴	۳	۳	۳
مثال ۲	۴	۴	۴	۴	۴	۴	۴	۴	۴	۴	۴	۴
مجموعه داده ۲-۹	۹	۹	۲	۱۱	۵	۹		۹	۸	۹	۹	۹
مجموعه داده ۲-۳ (ب)	۳	۳	۳	۷	۳	۳	۳	۳	۳	۳	۳	۳

همان‌طور که مشاهده می‌شود شاخص جدید در تمام مجموعه داده‌ها تعداد بهینه خوشه‌ها را پیدا کرده است. شاخص‌های PC، PE و PCAES به m بستگی ندارند ولی شاخص‌های FS ، XB و ECAS به m بستگی دارند. این شاخص‌ها هر یک توانسته‌اند برای برخی از مجموعه داده‌ها تعداد بهینه خوشه‌ها را پیدا کنند ولی هیچ یک در تمامی موارد موفق نبوده است. برای بررسی رفتار شاخص ECAS در مقابل m (که قبلا کارایی آن نسبت به شاخص‌های قبلی نشان داده شده است[۱])، این شاخص با الگوریتم گفته شده در بالا یعنی برای تمام زوج‌های m و c و نیز فقط با $m=2$ آزمایش شده است. همان‌طور که مشاهده می‌شود این شاخص برای $m=2$ در اکثر مجموعه داده‌ها توانسته است تعداد بهینه خوشه‌ها را پیدا کند ولی وقتی تمام زوج‌های m و c در نظر گرفته می‌شوند، در برخی موارد نمی‌تواند پاسخ صحیح را پیدا کند. از جمله در مجموعه داده‌های ۲-۴، ۲-۱۵، ۲-۵، ۲-۱۰، ۲-۱۰ (با وجود اغتشاش)، وقتی شاخص برای تمام زوج‌های m و c محاسبه شده است، مقدار شاخص ECAS در تعداد بهینه خوشه‌ها ماکزیمم نمی‌شود. زیرا همان‌طور که در بالا استدلال شد، این شاخص نسبت به تغییرات m پایدار نیست. همچنین در مجموعه داده ۲-۴(ج) برای $m=2$ نیز تعداد خوشه بهینه را پیدا نکرده است. در این آزمایش‌ها، به مجموعه داده ۲-۶ و ۲-۱۰ داده‌هایی به صورت اغتشاش اضافه شده‌اند، شاخص جدید در این موارد نیز توانسته است تعداد بهینه خوشه‌ها را پیدا کند که نشان دهنده پایداری این شاخص نسبت به وجود اغتشاش است.

به منظور بررسی تاثیر درجه فازی بودن خوشه‌بندی روی مقدار شاخص، مقادیر شاخص برای مجموعه داده‌های ۲-۵، ۲-۴(الف) و ۲-۴(ج) برای زوج‌های m و c مختلف در جداول (۲)، (۳) و (۴) آورده شده است. این جداول نشان می‌دهند که چگونه مقدار شاخص با تغییر مقدار m همچنان برای تعداد بهینه خوشه، ماکزیمم باقی می‌ماند.

جدول۲. مقادیر شاخص جدید ECASJ برای مجموعه داده ۲-۵ و برای مقادیر مختلف m و c

m \ c	2	3	4	5	6	7	8	9	10	11	12	13	14
1.7	0.658288	0.789158	0.928261		0.989857	0.977061	0.963631	0.946662	0.938473	0.942079	0.93721	0.935813	0.932336
1.8	0.655608	0.786636	0.929705	1	0.96925	0.95609	0.935436	0.910952	0.909655	0.899464	0.888965	0.88783	0.88446
1.9	0.656792	0.78876	0.934163	1	0.955128	0.936132	0.904469	0.881218	0.864186	0.855332	0.84324	0.838839	0.813663
2	0.661267	0.795559	0.941825	1	0.960029	0.9037	0.855482	0.829169	0.828223	0.801321	0.793829	0.791912	0.766136
2.1	0.669853	0.806825	0.951599	1	0.924662	0.880234	0.837264	0.812286	0.773228	0.756706	0.746243	0.731917	0.724591
2.2	0.683023	0.822598	0.963941	1	0.908721	0.848545	0.812726	0.754117	0.745646	0.715388	0.701575	0.681954	0.68559

جدول۳.. مقادیر شاخص جدید ECASJ برای مجموعه داده ۲-٤(الف) و برای مقادیر مختلف m و c

m \ c	2	3	4	5	6	7	8	9	10	11	12	13	14
1.7	0.724599	0.86406	0.966369	0.961227	0.954606	0.977531	0.987171	0.977126	0	-0.0057	0.987789	0.98502	0.988268
1.8	0.763662	0.897627	1	0.981531	0.977043	0.981883	0.989278	0.993483	0.989707	0.961336	0.976431	0.969886	-0.04225
1.9	0.781497	0.902499	1	0.968162	0.949355	0.948847	0.954411	0.952244	0.943247	0.930245	0.912575	0.919902	0.892966
2	0.802916	0.908464	1	0.953911	0.941416	0.928854	0.917117	0.900987	0.90302	0.88498	0.865107	0.843954	-0.16151
2.1	0.826723	0.916385	1	0.938824	0.91769	0.899885	0.878949	0.858707	-0.15102	0.823654	0.806539	0.808908	0.791706
2.2	0.855682	0.924082	1	0.922583	0.874752	0.858926	0.857355	0.820522	0.795967	0.78183	0.771278	0.748514	0.772457

جدول٤. مقادیر شاخص جدید ECASJ برای مجموعه داده ۲-٤(ج) و برای مقادیر مختلف m و c

c \ m	2	3	4	5	6	7	8
1.7	0.385236	0.391565	0.431724	0.424045	0.425859	0.41696	0.42417
1.8	0.370267	0.386036	0.400126	0.400058	0.390406	0.384861	0.387764
1.9	0.354123	0.362237	0.373729	0.366913	0.355979	0.349869	-0.65561
2	0.337091	0.315837	0.34594	0.333102	0.316466	0.314945	0.315481
2.1	0.320017	0.31192	0.317583	0.299832	0.280913	-0.71896	0.277415
2.2	0.303084	0.286895	0.289538	0.267822	0.248216	0.236575	0.001529

نتیجه‌گیری و پیشنهاد

در این مقاله شاخص اعتبار خوشه‌بندی ECAS طوری تصحیح شده است که بتواند همپوشانی بین خوشه‌ها را نیز در نظر بگیرد. این تصحیح با استفاده از فاصله جاکارد در معیار جدایش انجام شده و شاخص جدید ECASJ ارائه شده است که همچنین از مزایای تابع نمایی نیز استفاده می‌کند. این شاخص تمام مزایای شاخص ECAS را دارا، و علاوه بر آن بنا به دلیل گفته شده، نسبت به تغییر درجه فازی بودن خوشه‌بندی پایدار است. نتایج آزمایش شاخص جدید روی مجموعه داده‌های متعدد مصنوعی و مرسوم، نشان می‌دهد که شاخص ECASJ نیز مانند ECAS تعداد بهینه خوشه‌ها را در تمام مجموعه‌های داده پیدا کرده و نیز در مقابل اغتشاش و تغییرات m پایدار است. به علت پیچیدگی محاسبه فاصله جاکارد در مجموعه‌های داده با ابعاد بالاتر در این‌جا مجموعه‌های داده ۲ بعدی مورد آزمایش قرار گرفته‌اند. آزمایش شاخص جدید بر روی مجموعه‌های داده با ابعاد بالاتر نیز در مطالعات بعدی مطلوب می‌باشد.

منابع

1. M. H. F. Zarandi, M. R. Faraji, and M. Karbasian, **An Exponential Cluster Validity Index for Fuzzy Clustering with Crisp and Fuzzy Data**, *Scientia Iranica*, vol. 17, no. 2, pp. 95–110, 2010.
2. L. Zadeh, **Fuzzy sets**, *Inf.control*, no. 8, pp. 338–353, 1965.
3. J. C. Dunn, **A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters**, *Journal of Cybernetics*, vol. 3, pp. 32–57, 1974.
4. J. C. Bezdek, ***Pattern recognition with fuzzy objective function algorithms***. New York: Plenum Press, 1981.
5. J. V. De Oliveira, W. Pedrycz, and others, ***Advances in fuzzy clustering and its applications***. Wiley Online Library, 2007.
6. Y. Zhang, W. Wang, X. Zhang, and Y. Li, **A cluster validity index for fuzzy clustering**, *Information Sciences*, vol. 178, no. 4, pp. 1205–1218, Feb. 2008.
7. J. C. Bezdek, **Cluster validity with fuzzy sets**, *Journal of Cybernetics*, vol. 3, no. 3, pp. 58–73, 1973.
8. J. C. Bezdek, **Numerical taxonomy with fuzzy sets**, *Journal of Mathematical Biology*, vol. 1, no. 1, pp. 57–71, 1974.
9. Y. Fukuyama and M. Sugeno, **A new method of choosing the number of clusters for the fuzzy c-means method**, in *Proc. 5th Fuzzy Syst. Symp*, 1989, vol. 247.
10. X. L. Xie and G. Beni, **A validity measure for fuzzy clustering**, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 13, no. 8, pp. 841–847, 1991.
11. S. H. Kwon, **Cluster validity index for fuzzy clustering**, *Electronics Letters*, vol. 34, no. 22, pp. 2176–2177, 1998.
12. W. Wang and Y. Zhang, **On fuzzy cluster validity indices**, *Fuzzy Sets and Systems*, vol. 158, no. 19, pp. 2095–2117, 2007.
13. K. Rizman Žalik, **Cluster validity index for estimation of fuzzy clusters of different sizes and densities**, *Pattern Recognition*, vol. 43, no. 10, pp. 3374–3390, Oct. 2010.
14. E. Trauwaert, **On the meaning of Dunn's partition coefficient for fuzzy clusters**, *Fuzzy Sets and Systems*, vol. 25, no. 2, pp. 217–242, 1988.
15. R. N. Dave, **Validating fuzzy partitions obtained through c-shells clustering**, *Pattern Recognition Letters*, vol. 17, no. 6, pp. 613–623, 1996.
16. K.-L. Wu and M.-S. Yang, **A cluster validity index for fuzzy clustering**, *Pattern Recognition Letters*, vol. 26, no. 9, pp. 1275–1291, Jul. 2005.
17. D.-W. Kim, K. H. Lee, and D. Lee, **On cluster validity index for estimation of the optimal number of fuzzy clusters**, *Pattern Recognition*, vol. 37, no. 10, pp. 2009–2025, Oct. 2004.
18. N. R. Pal and S. K. Pal, **Entropy: a new definition and its applications**, *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 21, no. 5, pp. 1260–1270, 1991.
19. N. R. Pal and S. K. Pal, **Some properties of the exponential entropy**, *Information sciences*, vol. 66, no. 1, pp. 119–137, 1992.
20. K. L. Wu and M. S. Yang, **Alternative c-means clustering algorithms**, *Pattern recognition*, vol. 35, no. 10, pp. 2267–2278, 2002.
21. J. C. Bezdek, **Pattern Recognition in handbook of Fuzzy computation**, IOP Publishing Ltd., Boston, NY, 1998.
22. M. K. Pakhira, S. Bandyopadhyay, and U. Maulik, **Validity index for crisp and fuzzy clusters**, *Pattern recognition*, vol. 37, no. 3, pp. 487–501, 2004.
23. D. L. Davies and D W. Bouldin, **A Cluster Separation Measure**. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1 (2): 224–227, 1979.
24. S. Saha and S. Bandyopadhyay, **Some connectivity based cluster validity indices**, *Applied Soft Computing*, vol. 12, no. 5, pp. 1555–1565, May 2012.

25. Y. Hu, C. Zuo, Y. Yang, and F. Qu, **A cluster validity index for fuzzy c-means clustering**, in *System Science, Engineering Design and Manufacturing Informatization (ICSEM), 2011 International Conference on*, 2011, vol. 2, pp. 263–266.