سیستم های فازی و کاربردها
سال هفتم، شماره اول (بهار و تابستان ۱۴۰۳)
صص. ۱۰۹ تا ۱۲۳
شاپای چاپی: ۲۷۱۷ـ۴۴۰۹
شاپای الکترونیکی: ۲۷۱۷ـ۳۹۸۴
DOI:10.22034/JFSA.2024.420390.1187

بررسی تاثیر بکارگیری توابع زیان مختلف بر عملکرد مدل خوشه‌بندی فازی برای داده‌های فازی در صورت وجود داده‌های دورافتاده

الهام اسکندری* و علیرضا خواستان

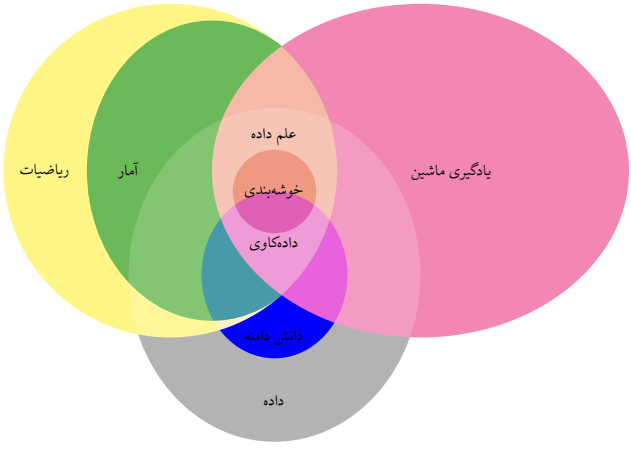
دانشکده ریاضی، دانشگاه تحصیلات تکمیلی علوم پایه، زنجان، ایران

تاریخ دریافت: ۱۴۰۲/۰۷/۲۰ تاریخ پذیرش: ۱۴۰۳/۰۵/۲۸

نوع مقاله: علمی‌ـ پژوهشی

چکیده. در این مقاله، استفاده از یک تابع زیان جدید در مدل خوشه‌بندی فازی برای داده‌های فازی را پیشنهاد و تاثیر استفاده از این تابع زیان و توابع زیان مربعی، هابر، خطی، سیگموییدی و لگاریتمی بر عملکرد مدل را در صورت وجود داده‌های دورافتاده در مجموعه داده‌ها، در شبیه‌سازی صورت پذیرفته بررسی کرده‌ایم. مجموعه داده‌های مورد استفاده، از نظر تعداد ویژگی‌ها (۲ و ۳)، تعداد کلاس‌ها (۳ و ۴) و پخش و تعداد داده‌های دورافتاده (۲۰ و ۳۸ و ۴)، دارای تنوع مناسبی هستند. نتایج شبیه‌سازی موید آن است که تابع زیان هابر و تابع زیان جدید، نسبت به وجود داده‌های دورافتاده مقاوم هستند.

2020 *Mathematics Subject Classification.* 62H30 ; 03E72
* Corresponding author
E-mails: eskandari.e@iasbs.ac.ir and khastan@iasbs.ac.ir.
عبارات و کلمات کلیدی. تابع زیان، خوشه‌بندی فازی مقاوم، داده‌های فازی L-R، داده‌های دورافتاده.



شکل ۱. نمودار وِن نشان‌دهنده‌ی ارتباط میان علم داده، یادگیری ماشین، آمار و داده‌کاوی. چنان‌چه آمار در فرایند خوشه‌بندی به‌کار برده شود، خوشه‌بندی، زیرمجموعه‌ای از علم داده و داده‌کاوی خواهد بود [۱۴].

۱. سرآغاز

خوشه‌بندی یکی از الگوریتم‌های مورد استفاده در یادگیری بدون نظارت[۱] است. یادگیری بدون نظارت، نوعی یادگیری ماشینی[۲] است و یادگیری ماشین، ارتباط در هم تنیده‌ای با داده‌کاوی دارد. از طرفی، یادگیری ماشین (به‌جز روش شبکه عصبی مصنوعی[۳])، همان مباحث مربوط به آمار و یادگیری آماری[۴] است. آمار به‌عنوان شاخه‌ای از ریاضیات، در بسیاری از تحلیل‌های مربوط به علم داده[۵] به‌کار می‌رود. بنابراین، میان این حوزه‌ها تا حدودی هم‌پوشانی وجود دارد (شکل ۱) که اگر از آن‌ها به‌درستی در کنار یک‌دیگر استفاده کنیم، نتایج رضایت‌بخشی را استخراج می‌کنیم.

روش خوشه‌بندی فازی سی‌ـمیانگین[۶] که در سال ۱۹۷۴ به طور مستقل توسط دان[۷] و بزدک[۸] معرفی و در سال ۱۹۸۱ توسط بزدک [۱] تعمیم داده شد، شناخته شده‌ترین و

[1] Unsupervised learning
[2] Machine learning
[3] Artificial neural network
[4] Statistical learning
[5] Data science
[6] Fuzzy C-Means (FCM)
[7] Dunn
[8] Bezdek

کاربردی‌ترین روش خوشه‌بندی فازی است. دراین روش تابع زیان

$$J(\mathbf{U},\mathbf{H})=\sum_{i=۱}^{N}\sum_{g=۱}^{c}(u_{ig})^{m}d^{۲}(\mathbf{x}_i,\mathbf{h}_g),$$

باید با توجه به قیود

$$\sum_{g=۱}^{c}u_{ig}=۱,\ i=۱,\cdots,N,$$

و

$$u_{ig}\geq ۰,\ i=۱,\cdots,N,\ g=۱,\cdots,c,$$

به حداقل برسد که در آن $\mathbf{x}_i$ نمونه‌ی iام، $\mathbf{h}_g$ مرکزخوشه‌ی gام، u_{ig} درجه عضویت نمونه‌ی iام در خوشه‌ی gام، $m\in[۱,\infty)$ پارامتر فازی کننده و

$$d^{۲}(\mathbf{x}_i,\mathbf{x}_{i'})=||\mathbf{x}_i-\mathbf{x}_{i'}||^{۲},$$

مربع فاصله‌ی اقلیدسی است.

برخی پژوهشگران توجه خود را به توسعه‌ی روش‌های مقاوم[۱] در مقابل داده‌های دورافتاده[۲] معطوف کرده‌اند؛ بعضی از مهم‌ترین رویکردها عبارت‌اند از انتخاب واسط[۳] داده‌ها به عنوان سرخوشه[۴] به جای انتخاب میانگین آن‌ها (شکل ۲ را ببینید)، استفاده از یک معیار عدم تشابه مقاوم، در نظر گرفتن یک خوشه‌ی اضافی به نام خوشه‌ی نویز، حذف کسری از داده‌های دورافتاده، وزن‌دهی کم به تاثیر داده‌های دورافتاده و احتمالاتی[۵].

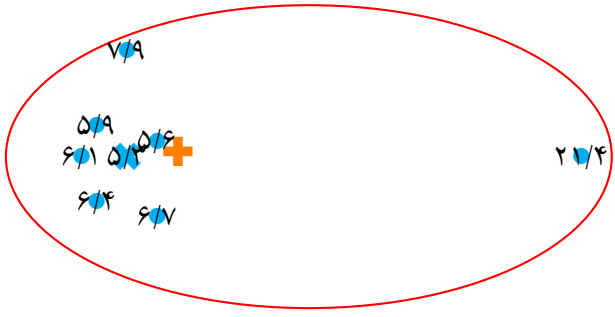
رویکردی دیگر، استفاده از توابع زیان مقاوم است که با مقاوم بودن معیار فاصله مرتبط است. دو تابع زیان بسیار معمول، توابع زیان مربعی[۶] یعنی

$$\mathcal{L}_{\mathrm{SQR}}(e)=e^{۲},$$

و خطی[۷] یعنی

$$\mathcal{L}_{\mathrm{LIN}}(e)=|e|,$$

[1] Robust
[2] Outlier data
[3] Medoid
[4] Prototype
[5] Possibilistic
[6] Squared
[7] Linear



شکل ۲. تعداد ۸ نمونه که میانگین آن‌ها با + و واسط آن‌ها با × نشان داده شده است. عدد نوشته شده روی هر نمونه، مجموع فواصل آن نمونه از سایر نمونه‌ها است. واسط، یکی از اعضای مجموعه داده است که کمترین فاصله با سایر نمونه‌ها را دارد. میانگین، به شدت تحت تاثیر داده‌های دورافتاده قرار می‌گیرد و بنابراین نمی‌تواند نماینده خوبی برای اعضای مجموعه داده باشد [۱۱].

هستند. اگر مقدار خطا کمتر از ۱ باشد، زیان مربعی، کوچک و اگر بیشتر از ۱ باشد، زیان مربعی بزرگ می‌شود. در صورت وجود داده‌های دورافتاده که موجب افزایش خطا می‌شوند؛ تابع زیان مربعی، زیان زیادی را محاسبه می‌کند. اِشکال اصلی تابع زیان خطی، بزرگ بودن مقدار گرادیان آن برای مقدارهای کوچک خطاست.

یک روش متداول برای مدل‌سازی عدم قطعیت، استفاده از اعداد فازی است. اعداد فازی، مفهوم اعداد حقیقی کلاسیک را تعمیم می‌دهند. ساتو[۱] و ساتو[۲] [۱۲] نخستین کسانی بودند که در سال ۱۹۹۵ یک مدل کاملا فازی یعنی یک مدل خوشه‌بندی فازی برای داده‌های فازی ارایه کردند.

هدف ما از مطالعه‌ی این موضوع، آزمودن توابع زیان مختلف در مدل خوشه‌بندی فازی برای داده‌های فازی و یافتن توابع زیان مقاوم نسبت به وجود داده‌های دورافتاده است. در این مقاله، استفاده از تابع زیان مقاوم لگاریتم کسینوس هذلولوی[۳] پیشنهاد شده است. این تابع زیان خاص به دلیل دارا بودن ویژگی‌های جالبی انتخاب شده است که در بخش ۴ به آن خواهیم پرداخت.

[1] Sato
[2] Sato
[3] Logarithm of hyperbolic cosine

این مقاله شامل شش بخش است. در بخش ۲ توضیحاتی مقدماتی ایراد می‌شود. در بخش ۳ به تاریخچه‌ی موضوع مورد مطالعه اشاره می‌شود. ویژگی‌های تابع زیان پیشنهادی، در بخش ۴ تبیین می‌شود. در بخش ۵ شبیه‌سازی صورت گرفته معرفی و نتایج ارایه می‌شود. بخش ۶ نیز به تفسیر نتایج دست‌یافته اختصاص دارد.

۲. پیش‌نیازها

مجموعه داده‌ی فازی را می‌توان در یک ماتریس $N \times J$ ذخیره کرد، به‌طوری‌که N تعداد نمونه و J تعداد ویژگی است.

تعریف ۱.۲. فرض کنید $L, R : [۰,۱] \to [۰,۱]$ دو تابع پیوسته و نزولی هستند که

$$L(۰) = R(۰) = ۱,$$

و

$$L(۱) = R(۱) = ۰.$$

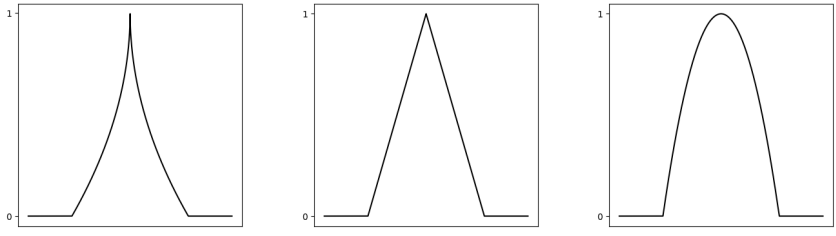
مجموعه‌ی فازی $x_i : \mathbb{R} \to [۰,۱],\ i = ۱,۲,\cdots,N$ که

$$x_i(u) = \begin{cases} ۰, & u \le c_{۱i} - l_i, \\ L\left(\dfrac{c_{۱i} - u}{l_i}\right), & c_{۱i} - l_i \le u \le c_{۱i}, \\ ۱, & c_{۱i} \le u \le c_{۲i}, \\ R\left(\dfrac{u - c_{۲i}}{r_i}\right), & c_{۲i} \le u \le c_{۲i} + r_i, \\ ۰, & u \ge c_{۲i} + r_i, \end{cases}$$

یک عدد فازی $L - R$ است که در این مقاله، آن را با $x_i = (c_{۱i}, c_{۲i}, l_i, r_i)$ نشان می‌دهیم به‌طوری‌که $c_{۱i}$ و $c_{۲i}$ $(c_{۱i} \le c_{۲i})$ به ترتیب، مرکز[1]های چپ و راست x_i و l_i و r_i به ترتیب، گستره[2]های چپ و راست x_{ij} نامیده می‌شوند. وقتی

$$L(x) = R(x) = ۱ - x,$$

[1] Center
[2] Spread



(ج) (ب) (آ)

شکل ۳. تابع عضویت از نوع (آ) سهموی، (ب) مثلثی و (ج) ریشه‌ی دوم.

تابع عضویت[1] از نوع ذوزنقه‌ای[2] است. چنان‌چه توابع L و R خطی باشند و داشته باشیم $c_{1i} = c_{2i} = c_i$، آن‌گاه x_i را عدد فازی مثلثی[3] می‌نامیم.

۳. مروری بر کارهای پیشین

در سال ۱۹۹۶، یانگ[4] و کو[5] [۱۵] فاصله‌ی

$$d^2(x_i, x_{i'}) = (c_i - c_{i'})^2 + ((c_i - \lambda l_i) - (c_{i'} - \lambda l_{i'}))^2 + ((c_i + \rho r_i) - (c_{i'} + \rho r_{i'}))^2, \quad (۱.۳)$$

را بین دو عدد فازی $L - R$ تک‌متغیره پیشنهاد کردند که در آن $\lambda = \int_0^1 L^{-1}(\omega)d\omega$ و $\rho = \int_0^1 R^{-1}(\omega)d\omega$ پارامترهایی هستند که شکل دُم[6] چپ و راست تابع عضویت را بیان می‌کنند؛ برای مثال وقتی $\lambda = \rho = \frac{2}{3}$، تابع عضویت از نوع سهموی[7] (شکل ۳آ)، وقتی $\lambda = \rho = \frac{1}{2}$، از نوع مثلثی (شکل ۳ب) و وقتی $\lambda = \rho = \frac{1}{3}$، از نوع ریشه‌ی دوم[8] (شکل ۳ج) است [۴]. مدل پیشنهادی یانگ و کو به شدت تحت تاثیر داده‌های دورافتاده است.

---

[1] Membership function
[2] Trapezoidal
[3] Triangular
[4] Yang
[5] Ko
[6] Tail
[7] Parabolic
[8] Square root

طی سه دهه‌ی اخیر، مدل‌های خوشه‌بندی فازی مقاومی از جمله مدل‌های یانگ و لیو[1] [۱۶]، بوتکیویچ[2] [۲]، هونگ[3] و یانگ [۱۰]، زرندی[4] و رضایی[5] [۱۷] و دورسو[6] و جیووانی[7] [۵] برای داده‌های فازی ارایه شدند.

در سال ۲۰۲۰، دورسو و لسکی[8] [۶] با در نظر گرفتن نسخه‌ی چندمتغیره فاصله‌ی پیشنهاد شده توسط یانگ و کو (۱.۳)، فاصله‌ی

$$\begin{aligned}\mathcal{D}(\mathbf{x}_i, \mathbf{h}_g) =& \mathcal{L}(\mathbf{c}_{1i} - \mathbf{c}_{1g}) + \mathcal{L}(\mathbf{c}_{2i} - c_{2g}) + \\ & \mathcal{L}[(\mathbf{c}_{1i} - \lambda \mathbf{l}_i) - (\mathbf{c}_{1g} - \lambda \mathbf{l}_g)] + \\ & \mathcal{L}[(\mathbf{c}_{2i} + \rho \mathbf{r}_i) - (\mathbf{c}_{2g} + \rho \mathbf{r}_g)],\end{aligned}$$

را ارایه کردند که در آن $\mathcal{L}$ یکی از توابع زیان مربعی، خطی، هابر[9] یعنی

$$\mathcal{L}_{\mathrm{HUB}}(e) = \begin{cases} \dfrac{e^2}{\delta^2}, & |e| \le \delta, \\ \dfrac{|e|}{\delta}, & |e| > \delta, \end{cases}$$

که $\delta > 0$، سیگموییدی[10] یعنی

$$\mathcal{L}_{\mathrm{SIG}}(e) = \frac{1}{1 + \exp(-\alpha(|e| - \beta))},$$

که $\alpha, \beta > 0$ و لگاریتمی[11] یعنی

$$\mathcal{L}_{\mathrm{LOG}}(e) = \log(1 + e^2),$$

است. آن‌ها مدل

$$J(\mathbf{U}, \mathbf{H}) = \sum_{i=1}^{N} \sum_{g=1}^{c} \beta_i (u_{ig})^m \mathcal{D}(\mathbf{x}_i, \mathbf{h}_g), \qquad (۲.۳)$$

[1] Liu
[2] Butkiewicz
[3] Hung
[4] Zarandi
[5] Razaee
[6] Durso
[7] Giovanni
[8] Leski
[9] Huber
[10] Sigmoidal
[11] Logarithmic

را ارایه کردند که در آن $\beta_i \in [۰, ۱]$ یک پارامتر مناسب است که تاثیر داده‌های دورافتاده را کاهش می‌دهد.

چنان‌چه $\mathcal{L}$ تابع زیان مربعی باشد، فاصله‌ی پیشنهاد شده توسط دورسو و لسکی همان فاصله‌ی پیشنهاد شده توسط یانگ و کو است. تابع زیان هابر، برای مقدارهای کوچک $|e|$، مربعی و برای مقدارهای بزرگ آن، خطی است. درنتیجه نسبت به تابع زیان مربعی، کمتر تحت تاثیر داده‌های دورافتاده است. همچنین، بر خلاف تابع زیان خطی، مشتق‌پذیر بوده و کمینه‌سازی آن شُدنی است. بنابراین تابع زیان هابر از مزایای هر دو تابع زیان مربعی و خطی بهره‌مند و فاقد معایب آن‌هاست.

وجه تمایز نتایج به دست آمده در این مقاله با نتایج دورسو و لسکی آن است که ما تاثیر استفاده از توابع زیان مختلف را با وزن‌دهی یکنواخت به داده‌ها ($\beta_i = ۱$) مورد مطالعه قرار می‌دهیم.

۴. تابع زیان پیشنهادی

در این بخش، ویژگی‌های تابع زیان پیشنهاد شده برای بکارگیری در مدل خوشه‌بندی فازی برای داده‌های فازی را تبیین می‌کنیم.

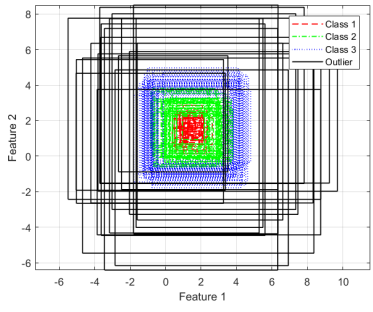
۱.۴. تابع زیان لگاریتم کسینوس هذلولوی. تابع زیان لگاریتم کسینوس هذلولوی یعنی

$$\mathcal{L}_{\text{Log-cosh}}(e) = \log(\cosh(e)),$$

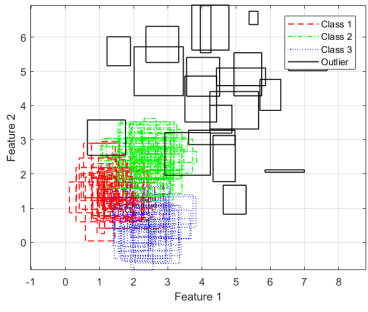
برای مقدارهای کوچک e، تقریبا برابر $\frac{e^۲}{۲}$ و برای مقدارهای بزرگ آن، تقریبا برابر $|e|-\log(۲)$ است. درنتیجه تحت تاثیر داده‌های دورافتاده قرار نمی‌گیرد. وجود مشتق دوم، مزیت دیگر این تابع زیان است که در تابع زیان هابر از آن بی‌بهره بودیم. بنابراین با استفاده از تابع زیان لگاریتم کسینوس هذلولوی، ضمن برخورداری از مزایای تابع زیان هابر، استفاده از روش‌هایی که برای بهینه‌سازی نیاز به مشتق دوم دارند نیز میسر است.

۵. مطالعه‌ی شبیه‌سازی

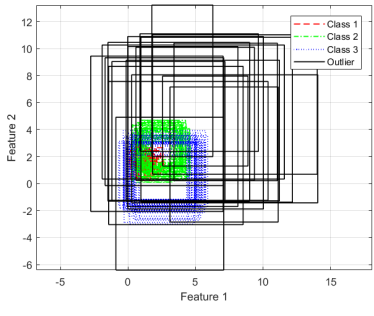
بزدک بر اساس تجربه $m \in [۱/۱, ۵]$ را پیشنهاد کرد [۱]. مدل (۲.۳) با $\beta_i = ۱$ و $m = ۲$ و در نظر گرفتن توابع زیان مربعی، هابر با $\delta = ۵$، خطی، سیگموییدی با $\alpha = \beta = ۲$، لگاریتمی و لگاریتم کسینوس هذلولوی، در MATLAB R2021a پیاده‌سازی شده است.



(ب)



(آ)



(ج)

شکل ۴. مجموعه داده‌های فازی مثلثی تولید شده طبق سناریوهای (آ) مرکز دورافتاده، (ب) گستره‌های دورافتاده و (ج) مرکز و گستره‌های دورافتاده.

همه‌ی آزمایش‌ها بر روی یک ماشین (سیستم عامل: macOS، ظرفیت حافظه RAM: 8GB، پردازنده: 1.1GHz dual-core Intel Core i3) انجام شده است.

۱.۵. مجموعه داده‌های تولید شده. برای تولید مجموعه داده‌های فازی مثلثی، سه سناریوی مرکز دورافتاده مثل شکل ۴آ، گستره‌های دورافتاده مثل شکل ۴ب و مراکز و گستره‌های دورافتاده مثل شکل ۴ج در نظر گرفته شده است [۵، ۷، ۸]. هر مجموعه داده در $\mathbb{R}^2$ شامل ۱۲۰ نمونه در ۳ کلاس دارای هم‌پوشانی با تعداد اعضای برابر و یک گروه ۲۰ عضوی از داده‌های دورافتاده است. جدول ۱، سازوکار تولید مجموعه داده‌ها را بیان می‌کند.

۲.۵. مجموعه داده‌های معیار. مجموعه داده‌ی مصنوعی Zelnik4 [۱۸] در $\mathbb{R}^2$ که در شکل ۵ نشان داده شده است، شامل ۴۸۴ نمونه در ۴ کلاس و یک گروه ۱۳۸ عضوی از

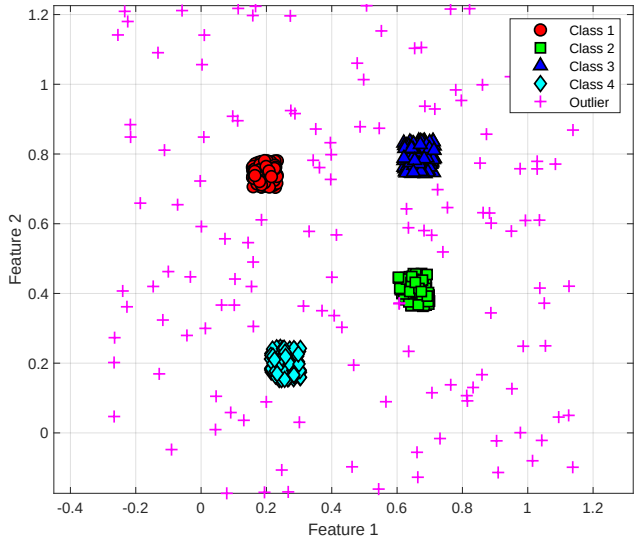
جدول ۱. سازوکار تولید مجموعه داده‌های فازی مثلثی طبق سناریوهای مرکز دورافتاده، گستره‌های دورافتاده و مرکز و گستره‌های دورافتاده.

سناریو		بُعد	مرکز	گستره‌ها
مرکز	کلاس ۱	۱ و ۲	$U[۱,۲]$	$U[۰,۱]$
	کلاس ۲	۱ و ۲	$U[۲,۳]$	$U[۰,۱]$
	کلاس ۳	۱	$U[۲,۳]$	$U[۰,۱]$
		۲	$U[۰,۱]$	
	داده‌های دورافتاده	۱ و ۲	$\mathcal{N}(۵,۲)$	$U[۰,۱]$
گستره‌ها	کلاس ۱	۱ و ۲	$U[۱,۲]$	$U[۰,۱]$
	کلاس ۲	۱ و ۲	$U[۱,۲]$	$U[۱,۲]$
	کلاس ۳	۱ و ۲	$U[۱,۲]$	$U[۲,۳]$
	داده‌های دورافتاده	۱ و ۲	$U[۱,۲]$	$\mathcal{N}(۵,۲)$
مرکز و گستره‌ها	کلاس ۱	۱ و ۲	$U[۱,۲]$	$U[۰,۱]$
	کلاس ۲	۱ و ۲	$U[۲,۳]$	$U[۱,۲]$
	کلاس ۳	۱	$U[۲,۳]$	$U[۲,۳]$
		۲	$U[۰,۱]$	
	داده‌های دورافتاده	۱ و ۲	$\mathcal{N}(۵,۲)$	$\mathcal{N}(۵,۲)$

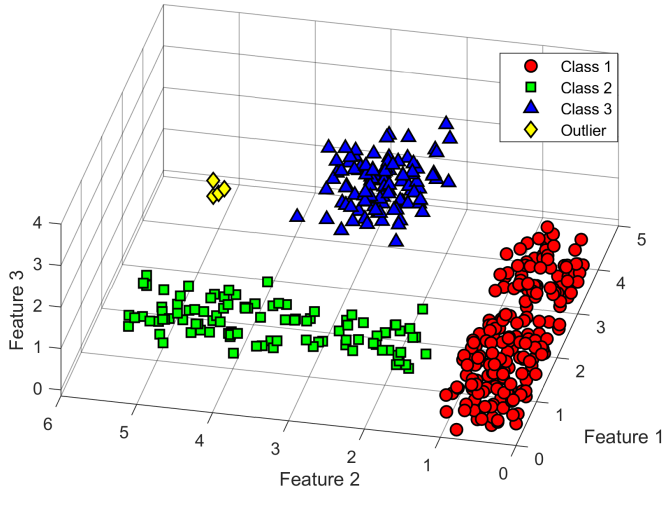
داده‌های دورافتاده است؛ تعداد ۱۱۱ نمونه با برچسب یک، ۱۱۴ نمونه با برچسب دو، ۱۵۰ نمونه با برچسب سه و ۱۰۹ نمونه با برچسب چهار.

مجموعه داده‌ی مصنوعی Lsun3D [۱۳] در $\mathbb{R}^۳$ که در شکل ۶ نشان داده شده است، شامل ۴۰۰ نمونه در ۳ کلاس جدا از هم با اشکال هندسی متفاوت و یک گروه کوچک ۴ عضوی از داده‌های دورافتاده است؛ تعداد ۲۰۰ نمونه با برچسب یک، ۱۰۰ نمونه با برچسب دو و ۱۰۰ نمونه با برچسب سه.

برای فازی‌سازی مجموعه داده‌های Zelnik4 و Lsun3D، هر دو مجموعه داده را به عنوان مرکز داده‌های فازی مثلثی در نظر گرفته‌ایم و فرض کرده‌ایم که گستره‌های این داده‌ها نیز از توزیع آماری $\mathcal{N}(۱, ۰/۳)$ پیروی می‌کنند.



شکل ۵. مجموعه داده‌ی Zelnik4.



شکل ۶. مجموعه داده‌ی Lsun3D.

جدول ۲. نتایج بکارگیری توابع زیان مختلف در مدل خوشه‌بندی فازی برای داده‌های فازی.

مجموعه داده‌ها		مرکز		گستره‌ها		مرکز و گستره‌ها		Zelnik4		Lsun3D		میانگین
تابع زیان	شاخص	میانگین	95% CI	میانگین	95% CI	میانگین	95% CI	میانگین	95% CI	میانگین	95% CI	
SQR	Frand	۰/۷۲	[۰/۷۰, ۰/۷۳]	۰/۵۸	[۰/۵۸, ۰/۵۹]	۰/۶۶	[۰/۶۳, ۰/۶۸]	۰/۷۷	[۰/۷۵, ۰/۷۹]	۰/۷۰	[۰/۶۸, ۰/۷۲]	۰/۶۸۶
	HUL	۰/۵۵	[۰/۵۳, ۰/۵۸]	۰/۳۰	[۰/۲۸, ۰/۳۲]	۰/۴۷	[۰/۴۴, ۰/۵۱]	۰/۵۳	[۰/۵۰, ۰/۵۶]	۰/۵۳	[۰/۵۱, ۰/۵۵]	۰/۴۷۶
	Runtime	۲/۰۷		۱/۱۴		۱/۴۴		۶۲/۲۶		۲۰/۷۶		۱۷/۵۳
HUB	Frand	۰/۷۴	[۰/۷۲, ۰/۷۶]	۰/۵۹	[۰/۵۸, ۰/۶۰]	۰/۶۹	[۰/۶۶, ۰/۷۱]	۰/۷۷	[۰/۷۵, ۰/۷۹]	۰/۷۰	[۰/۶۸, ۰/۷۱]	۰/۶۹۸
	HUL	۰/۵۸	[۰/۵۶, ۰/۶۱]	۰/۳۱	[۰/۲۹, ۰/۳۳]	۰/۵۲	[۰/۴۹, ۰/۵۵]	۰/۵۴	[۰/۵۲, ۰/۵۷]	۰/۵۲	[۰/۵۰, ۰/۵۵]	۰/۴۹۴
	Runtime	۲/۱۲		۱/۴۱		۱/۵۳		۶۳/۹۴		۲۲/۷۵		۱۸/۳۵
LIN	Frand	۰/۶۴	[۰/۶۲, ۰/۶۵]	۰/۵۳	[۰/۵۳, ۰/۵۴]	۰/۶۳	[۰/۶۱, ۰/۶۴]	۰/۶۵	[۰/۶۳, ۰/۶۶]	۰/۶۲	[۰/۶۱, ۰/۶۴]	۰/۶۱۴
	HUL	۰/۳۶	[۰, ۳۴, ۰/۳۸]	۰/۱۷	[۰/۱۶, ۰/۱۸]	۰/۳۵	[۰/۳۳, ۰/۳۷]	۰/۲۹	[۰/۲۷, ۰/۳۰]	۰/۳۷	[۰/۳۵, ۰/۳۹]	۰/۳۰۸
	Runtime	۱/۷۵		۱/۱۴		۱/۴۵		۴۱/۱۸		۱۳/۸۳		۱۱/۸۷
SIG	Frand	۰/۶۶	[۰/۶۵, ۰/۶۷]	۰/۵۳	[۰/۵۲, ۰/۵۳]	۰/۶۶	[۰/۶۴, ۰/۶۸]	۰/۵۳	[۰/۵۳, ۰/۵۴]	۰/۶۵	[۰/۶۳, ۰/۶۶]	۰/۶۰۶
	HUL	۰/۴۴	[۰/۴۲, ۰/۴۵]	۰/۱۵	[۰/۱۴, ۰/۱۶]	۰/۴۱	[۰/۳۹, ۰/۴۴]	۰/۰۳	[۰/۰۳, ۰/۰۳]	۰/۴۳	[۰/۴۱, ۰/۴۵]	۰/۲۹۲
	Runtime	۲/۳۶		۱/۴۲		۱/۵۰		۴۵/۰۸		۱۸/۰۸		۱۳/۶۹
LOG	Frand	۰/۷۶	[۰/۷۴, ۰/۷۸]	۰/۵۹	[۰/۵۸, ۰/۶۰]	۰/۷۱	[۰/۶۹, ۰/۷۳]	۰/۷۷	[۰/۷۵, ۰/۷۹]	۰/۶۴	[۰/۶۲, ۰/۶۵]	۰/۶۹۴
	HUL	۰/۵۸	[۰/۵۵, ۰/۶۰]	۰/۳۰	[۰/۲۸, ۰/۳۱]	۰/۵۱	[۰/۴۹, ۰/۵۴]	۰/۵۳	[۰/۵۱, ۰/۵۶]	۰/۴۲	[۰/۴۰, ۰/۴۴]	۰/۴۶۸
	Runtime	۲/۹۴		۱/۳۴		۱/۹۰		۷۳/۰۵		۱۸/۰۸		۱۹/۴۶
Log-cosh	Frand	۰/۷۷	[۰/۷۵, ۰/۷۹]	۰/۵۹	[۰/۵۸, ۰/۶۰]	۰/۶۹	[۰/۶۷, ۰/۷۱]	۰/۷۶	[۰/۷۴, ۰/۷۷]	۰/۶۷	[۰/۶۵, ۰/۶۸]	۰/۶۹۶
	HUL	۰/۶۰	[۰/۵۸, ۰/۶۳]	۰/۳۰	[۰/۲۹, ۰/۳۲]	۰/۵۱	[۰/۴۸, ۰/۵۳]	۰/۵۲	[۰/۴۹, ۰/۵۴]	۰/۴۷	[۰/۴۵, ۰/۴۹]	۰/۴۸۰
	Runtime	۲/۹۴		۱/۴۵		۱/۸۱		۶۹/۸۴		۲۲/۲۱		۱۹/۶۵

۳.۵. نتایج. میانگین شاخص رَند فازی[1] Frand [۳] و میانگین شاخص هولرمایر[2] HUL [۹] و بازه اطمینان آن‌ها و همچنین میانگین زمان اجرا برای ۱۰۰ بار تکرار آزمایش در جدول ۲ گزارش شده است. هرچه مقدار به دست آمده برای شاخص‌های Frand و HUL به یک نزدیک‌تر و زمان اجرا کمتر باشد، نتیجه‌ی خوشه‌بندی، رضایت‌بخش‌تر خواهد بود.

۶. نتیجه‌گیری

بکارگیری تابع زیان لگاریتم کسینوس هذلولوی در مدل خوشه‌بندی فازی برای داده‌های فازی را پیشنهاد و تاثیر استفاده از این تابع و نیز توابع زیان مربعی، هابر، خطی، سیگموییدی و لگاریتمی بر عملکرد مدل را در صورت وجود داده‌های دورافتاده در مجموعه داده‌ها، بررسی کرده‌ایم.

شکل ۷ نمودارهای توابع زیان مربعی، هابر، خطی، سیگموییدی، لگاریتمی و لگاریتم کسینوس هذلولوی را مقایسه می‌کند. به طور کلی، آن دسته از توابع زیان که

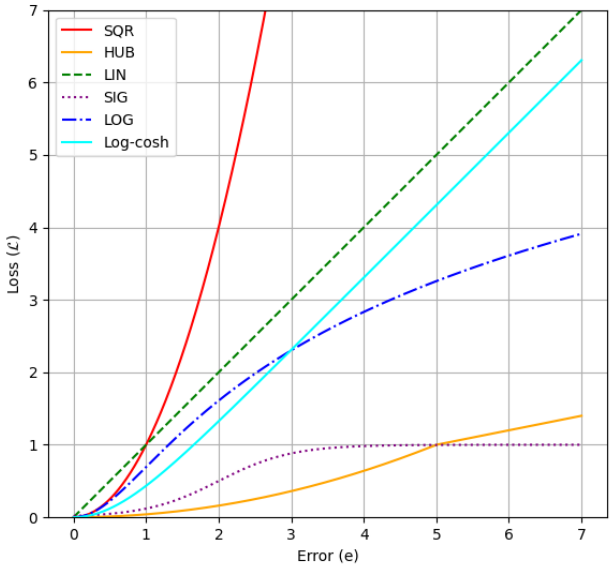
(۱) در نزدیکی $e = ۰$ شیب کمی داشته باشند؛

(۲) با افزایش خطا و دور شدن از نقطه‌ی $e = ۰$، شیب تابع با آهنگ ملایمی افزایش یابد؛

(۳) در صورت وجود داده‌های دورافتاده و افزایش چشمگیر مقدار خطا، زیان مربوطه را با بزرگ‌نمایی محاسبه نکنند؛

کارایی بهتری خواهند داشت. نتایج شبیه‌سازی نیز موید آن است که بهترین عملکرد روش خوشه‌بندی، برای توابع زیان هابر و لگاریتم کسینوس هذلولوی به دست آمد. اِشکال تابع زیان خطی، شیب زیاد آن در نزدیکی $e = ۰$ است. تابع زیان مربعی نیز برای $e > ۱$، زیان زیادی را محاسبه می‌کند و به همین دلیل، نسبت به وجود داده‌های دورافتاده مقاوم نیست. کمترین زمان اجرای روش خوشه‌بندی، برای توابع زیان خطی و سیگموییدی به دست آمد.

[1] Fuzzy rand
[2] Hullermeier



شکل ۷. مقایسه‌ی نمودارهای توابع زیان مربعی، هابر با $\delta = ۵$، خطی، سیگموییدی با $\alpha = \beta = ۲$، لگاریتمی و لگاریتم کسینوس هذلولوی.

مراجع

[1] Bezdek, J. (1981) Pattern recognition with fuzzy objective function algorithms. *Plenum Press, New York*.

[2] Butkiewicz, B. (2005) Robust fuzzy clustering with fuzzy data. *Advances in Web Intelligence, Third International Atlantic Web Intelligence Conference, AWIC 2005, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg*, **3528**, 76–82.

[3] Campello, R. (2007) A fuzzy extension of the Rand index and other related indexes for clustering and classification assessment. *Pattern Recognition Letters*, **28** (7), 833–841.

[4] D'Urso, P. (2007) Fuzzy clustering of fuzzy data. *Advances in Fuzzy Clustering and Its Applications*, 155–192.

[5] D'Urso, P., De Giovanni, L. (2014) Robust clustering of imprecise data. *Chemometrics and Intelligent Laboratory Systems*, **136**, 58–80.

[6] D'Urso, P., Leski, J. (2020) Fuzzy clustering of fuzzy data based on robust loss functions and ordered weighted averaging. *Fuzzy Sets and Systems*, **389**, 1–28.

[7] Eskandari, E., Khastan, A. (2023) A robust fuzzy clustering model for fuzzy data based on an adaptive weighted L۱ norm. *Iranian Journal of Fuzzy Systems*, **20** (6), 1-20.

[8] Eskandari, E., Khastan, A., Tomasiello, S. (2022) Improved determination of the weights in a clustering approach based on a weighted dissimilarity measure between fuzzy data. *2022 IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE), Padua, Italy*, 1-6.

[9] Hullermeier, E., Rifqi, M., Henzgen, S., Senge, R. (2012) Comparing fuzzy partitions: a generalization of the Rand index and related measures. *IEEE Transactions on Fuzzy Systems*, **20** (3), 546–556.

[10] Hung, W., Yang, M. (2005) Fuzzy clustering on LR-type fuzzy numbers with an application in Taiwanese tea evaluation. *Fuzzy Sets and Systems*, **150** (3), 561–577.

[11] Jin, X., Han, J. (2010) K-medoids clustering. *Encyclopedia of Machine Learning*, 697-700.

[12] Sato, M., Sato, Y. (1995) Fuzzy clustering model for fuzzy data. *Proceedings of 1995 IEEE International Conference on Fuzzy Systems*, **4**, 2123–2128.

[13] Thrun, M. (2018) Projection-based clustering through self organization and swarm intelligence. *Springer Vieweg, Wiesbaden*.

[14] Urbanowicz, R.J. [@DocUrbs]. (2018, Jun 15). *New proposed field/term Venn diagram for an upcoming talk. My take on illustrating the relationship between #DataScience, #MachineLearning, #ArtificialIntelligence, #Statistics* [Tweet]. Twitter. https://twitter.com/DocUrbs/status/1007375834347376642

[15] Yang, M., Ko, C. (1996) On a class of fuzzy c-numbers clustering procedures for fuzzy data. *Fuzzy Sets and Systems*, **84** (1), 49–60.

[16] Yang, M., Liu, H. (1999) Fuzzy clustering procedures for conical fuzzy vector data. *Fuzzy Sets and Systems*, **106** (2), 189–200.

[17] Zarandi, M., Razaee, Z. (2010) A fuzzy clustering model for fuzzy data with outliers. *International Journal of Fuzzy System Applications (IJFSA)*, **1** (2), 29–42.

[18] Zelnik-Manor, L., Perona, P. (2004) Self-tuning spectral clustering. *Proceedings of the 17th International Conference on Neural Information Processing Systems*, **17**, 1601–1608.