مجلـه‌ی علمی پژوهش در علوم رایانه ، شماره ۳ ، پاییز ۱۳۹۵، ص ۳۳-۲۶

ISSN: 2745-3673
http://www.Rcsj.ir

مجلـه‌ی علمی پژوهش در علوم رایانه (سال دوم)
شماره ۳ / پاییز ۱۳۹۵ / ص ۳۳-۲۶

# ارائه یک روش مبتنی بر تئوری فازی برای خوشه بندی داده های نامتعادل

**رضا گلی پور[۱]، حمید متقی گلشن[۲]**

[۱] کارشناسی ارشد کامپیوتر، نرم افزار، گروه کامپیوتر، آشتیان، ایران

[۲] دانشگاه آزاد اسلامی، واحد آشتیان، گروه کامپیوتر، آشتیان، ایران

نام و نشانی ایمیل نویسنده مسئول:

**رضا گلی پور**

***reza_gp@yahoo.com***

## چکیده

خوشه بندییکی از شاخه‌هاییادگیری بدون نظارت می باشد و فرایند خودکاری است که در طی آن، نمونه ها به دسته‌هایی که اعضای آن مشابه یکدیگر می‌باشند تقسیم می‌شوند که به این دسته‌ها خوشه گفته می‌شود. یکی از انواع چالش بر انگیز داده ها، مجموعه های داده ای نامتعادل است. به یک مجموعه داده نامتعادل گفته می شود اگر بسیاری از نمونه های موجود در آن مجموعه داده متعلق به یک کلاس و تعداد کمی از نمونه ها متعلق به دسته های دیگر باشد. در نتیجه ارائه یک روش خوشه بندی برای داده های نامتعادل که مشکلات روش های پیشین را حل کند از چالش‌های عمده در زمینه خوشه‌بندی داده‌ها به شمار می‌رود. در اینمقاله تلاش شده است که با ارائه یک الگوریتم خوشه بندی مبتنی بر الگوریتم فازی و استفاده از روش‌های وزن‌دهی ویژگی‌ها و همچنین انتخاب ویژگی باعث بهبود عملکرد الگوریتم خوشه‌بندی فازی در داده های نامتعادل شود. نتایج شبیه سازی عملکرد مناسب روش پیشنهادی را در مقایسه با کارهای پیشین نشان دادند.

**واژگان کلیدی:** تئوری فازی، خوشه‌بندی، خوشه‌بندی فازی، داده های نامتعادل.

مجلـه‌ی علمی پژوهش در علوم رایانه ، شماره ۳ ، پاییز ۱۳۹۵، ص ۳۳-۲۶

ISSN: 2745-3673
http://www.Rcsj.ir

## مقدمه

پردازش داده، یکی از شاخص‌های بسیار مهم در دنیای اطلاعات است. در پردازش داده، داده‌ها کاراکترها و اعداد هستند که بیانگر اندازه‌ها از دیدگاه پدیده‌های قابل مشاهده‌اند. یک داده اولیه تنها یک اندازه از پدیده قابل مشاهده است. اطلاعات اندازه‌گیری شده سپس به‌صورت الگوریتمی مشتق می‌شود و به‌صورت منطقی نتیجه‌گیری می‌شود و یا به‌صورت آماری از چندین داده محاسبه می‌شود. اطلاعات، یا به‌صورت پاسخ به یک درخواست تعریف می‌شود و یا پاسخ به یک محرک که می‌تواند درخواست‌های بعدی را در پی داشته باشد[۱, ۲].

پردازش داده در حال حاضر جزء مهم‌ترین ابزارها جهت بهره‌برداری مؤثر از حجم انبوه داده‌ها می‌باشد و اهمیت وجود آن هر روز افزایش می‌یابد. به عبارتی داده‌کاوی علمی نسبتاً جدید است که از انجام تحقیقات در رشته‌های آمار، یادگیری ماشین و علوم کامپیوتر مخصوصاً مدیریت پایگاه داده‌ها شکل گرفته است. به‌طور کلی مواردی از قبیل؛ استخراج یا کاوش دانش از میان حجم عظیم داده‌ها، استخاج اطلاعات و مدل کردن الگوهای پنهانی در میان انبوه داده‌ها، استخراج اطلاعات غیر منتظره، استخراج اطلاعات یا الگوهای مفید و جالب از داده‌ها در پایگاه داده‌های بزرگ، در حوزه داده‌کاوی قرار می‌گیرد.

یکی از انواع چالش بر انگیز داده ها، مجموعه های داده ای نامتعادل است. به یک مجموعه داده نامتعادل گفته می شود اگر بسیاری از نمونه های موجود در آن مجموعه داده متعلق به یک کلاس و تعداد کمی از نمونه ها متعلق به دسته های دیگر باشد. به عبارتی دیگر یک مجموعه را می توان مجموعه داده ای نا متعادل نامید در حالی که حداقل یک کلاس در آن وجود داشته باشد که تعداد نمونه های آموزش مربوط به آن کلاس بسیار کم باشد[۳].

خوشه‌بندییکی از بهترین روش‌هایی است که برای کار با داده‌ها ارائه شده است. قابلیت آن در ورود به فضای داده و تشخیص ساختار آنها، خوشه‌بندی را یکی از ایده ال ترین مکانیزم‌ها برای کار با دنیای عظیم داده‌ها کرده است[۴]. خوشه‌بندییکی از شاخه‌هاییادگیری بدون نظارت می‌باشد و فرایند خودکاری است که در طی ان، نمونه‌ها به دسته‌هایی که اعضای ان مشابه یکدیگرمی‌باشند تقسیممی‌شوند که به این دسته‌ها خوشه گفته می‌شود. در واقع تحلیل خوشه بندی به دنبال سازما ن دهی مجموعه‌ای از داده‌ها در یک سری خوشه است به طوری که داده‌ها در هر خوشه بالاترین درجه شباهت را دارا بوده و داده‌های متعلق به خوشه‌های مختلف دارای حداکثر درجه عدم شباهت هستند.

روش‌های خوشه‌بندی را می‌توان از چندین جنبه تقسیم‌بندی کرد:

- خوشه‌بندی انحصاری[1]در مقابل خوشه‌بندی با هم‌پوشی[2]:در روش خوشه‌بندی انحصاری پس از خوشه‌بندی هر داده دقیقاً به یک خوشه تعلق می‌گیرد. اما در خوشه‌بندی با همپوشی پس از خوشه‌بندی به هر داده یک درجه تعلق به ازای هر خوشه نسبت داده می‌شود. به‌عبارت‌دیگر، یک داده می‌تواند با نسبت‌های متفاوتی به چندین خوشه تعلق داشته باشد. خوشه‌بندی k-means را می‌توان از انواع خوشه‌بندی انحصاری و خوشه‌بندی C-means را انواع خوشه‌بندی با هم‌پوشی در نظر گرفت[۵].
- خوشه‌بندی سلسله مراتبی[3]در مقابل خوشه‌بندی مسطح[4]: در روش خوشه‌بندی سلسله مراتبی، به خوشه‌های نهایی بر اساس میزان عمومیت آن‌ها ساختاری سلسله مراتبی نسبت داده می‌شود. مانند روش Single Link[۶]. ولی در خوشه‌بندی مسطح تمامی خوشه‌های نهایی دارای یک‌میزان عمومیت هستند مانند K-means. به ساختار سلسله مراتبی حاصل از روش‌های خوشه‌بندی سلسله مراتبی، دندوگرام[5] گفته می‌شود.با توجه به اینکه روش‌های خوشه‌بندی سلسله مراتبی اطلاعات بیشتر و دقیق‌تری تولید می‌کنند برای تحلیل داده‌های با جزئیات پیشنهاد می‌شوند ولی از طرفی چون پیچیدگی محاسباتی بالایی دارند برای مجموعه داده‌های با ابعاد بالا روش‌های خوشه‌بندی مسطح پیشنهاد می‌شوند[۷].

در میان روش‌های مختلف در خوشه‌بندی  داده‌ها، روش فازیبه طور گسترده در زمینه‌های مختلف مورد استفاده قرار می‌گیرد. .الگویتمfuzzy c-meansبه طور گسترده در زمینه‌های تحقیقات فازی برای مسائل خوشه‌بندی مورد توجه قرار گرفته است. همچنین، این الگوریتمخوشه‌بندی به صورت موفقیت آمیز در زمینه‌هایی مانند سنجش از راه دور، خوشه‌بندی سری‌های زمانی و قطعه بندی تصاویر رنگی مورد استفاده قرار گرفته شده است. بسیاری از مفاهیم بنیادی تئوری فازی به‌وسیله زاده در اواخر دهه ۶۰ و اوایل دهه ۷۰ مطرح گردید.

---

[1] Exclusive clustering
[2] Overlapping clustering
[3]Hierarchical clustering
4 Flat clustering
[5] Dendogram

ISSN: 2745-3673
http://www.Rcsj.ir

پس از معرفی مجموعه‌ای فازی در سال ۱۹۶۵ او مفاهیم الگوریتم‌های فازی در سال ۱۹۶۸ تصمیم‌گیری فازی در سال ۱۹۷۰ و ترتیب فازی را در سال ۱۹۷۱ مطرح نمود. در سال ۱۹۷۳ او مقاله دیگری را منتشر کرد به نام: (طرح یک راه‌حل جدید برای تجزیه تحلیل سیستم‌های پیچیده و فرایندهای تصمیم‌گیری) این مقاله اساس کنترل فازی را بنا کرد، او در این مفهوم متغیرهای زبانی و استفاده از قواعد اگر آنگاه را برای فرموله کردن دانش بشری معرفی نمود.

در اکثر الگوریتم‌ها پیش تر ذکر شده برایخوشه‌بندی فازیc-means و همچنین نسخه‌های بهبود یافته آنها، در هنگام خوشه‌بندی داده‌ها، ارزش تمام ویژگی‌ها برابر در نظر گرفته شده است. در حالی که در مجموعه داده‌های واقعی ممکن است بعضی از ویژگی‌ها نسبت به سایر ویژگی‌ها دارای اهمیت بیشتری باشند. به این منظور، در زمان خوشه‌بندی باید آن دسته از ویژگی‌ها که دارای ارزش و قدرت بیشتری هستند، با اهمیت بیشتری در نظر گرفته شوند. برای حل این مشکل روش‌هایی ارائه شده است که در آنها هر ویژگی بر اساس وزنی که دارد مورد ارزیابی قرار می‌گیرد. در سال‌های اخیر الگوریتم مختلف خوشه‌بندی فازی با وزن دهی ویژگی‌ها ارائه شده است. در این روش‌ها اکثرا الگوریتمخوشه‌بندی در دو مرحله مختلف اجرا می‌شود: در مرحله اول وزن هر ویژگی مشخص می‌شود و در مرحله بعد بر اساس وزن‌های به دست آمده در مرحله قبل، الگوریتمخوشه‌بندی فازی اجرا می‌شود. متاسفانه در این روش‌ها در مرحله دوم وزن هر ویژگی ثابت در نظر گرفته می‌شود و وزن‌ها دچار تغییر نمی‌شوند، به همین خاطر ممکن است که وزن دهی ویژگی نتواند به خوبی اهمیت هر ویژگی را در طی فرایندخوشه‌بندی نشان دهد. برای حل مشکلات موجود در روش‌هایخوشه‌بندی فازی وزن دهی شده، در این مقاله یک روش بهبود یافتهخوشه‌بندی فازی ارائه شده است. از نوآوری‌های موجود در این پایان نامه می‌توان به موراد زیر اشاره کرد:

۱- فرایند وزن دهی به ویژگی‌ها و نیرخوشه‌بندی به صورت دینامیک و همزمان انجام می‌شود
۲- خوشه‌بندی و وزن دهی ویژگی‌ها بر اساس یک تابع هدف از قبل مشخص شده صورت می‌گیرد
۳- ارائه یک معیار جدید برای وزن دهی به ویژگی‌ها

در ادامه این مقاله در بخش دوم، روش استاندارد خوشه‌بندی فازیc-meansبه صورت مختصر توصیح داده می‌شود. در بخش سوم الگوریتم بهبود یافتهخوشه‌بندی فازی با وزن دهی ویژگی‌ها شرح داده می‌شود. در بخش چهارم نتایج و آزمایش‌ها مورد بررسی قرار می‌گیرند و در نهایت در بخشپنجمیک نتیجه گیری کلی از روش ارائه شده و همچنین پیشنهادهایی برای کارهای آینده ارائه می‌شود.

## ۱. خوشه‌بندی فازی

خوشه بندی از ابزارهای متداول داده کاوی بوده که به استخراج دسته‌هایی با حداکثر شباهت بین عناصر داخل دسته و حداقل شباهت با عناصر سایر دسته‌هامیپردازد. این تشابه یا عدم تشابه براساس معیارهای اندازه گیری فاصله تعریف می‌شود. در واقع خوشه‌بندی یک کلاس بندی بدون نظارت است که در آن کلاس‌ها از پیش تعریف نشده اند. در خوشه‌بندی کلاسیک، هر نمونه ورودی متعلق به یک و فقط یک خوشه است و نمی‌تواند عضو دو خوشه و یا بیشتر باشد. به عبارتی خوشه‌ها همپوشانی ندارند، در حالی که در خوشه‌بندی فازی یک نمونه می‌تواند متعلق به بیش از یک خوشه باشد. خوشه‌بندی فازی به کشف مدل‌هایفازیازداده‌هامی‌پردازد.تئوری مجموعه فازی به کار گرفته شده در تحلیل خوشه یابی عمدتاً بر روی خوشه یابی فازی بر پایه روابط فازی و توابع هدف تمرکز دارد[۸].

ایده بنیادین در خوشه‌بندی فازی به این ترتیب است که فرض کنیم هر خوشه مجموعه ای از عناصر است،سپس با تغییر در تعریف عضویت عناصر در این مجموعه از حالتی که یک عنصر فقط بتواند عضو یک خوشه باشد(حالت افرازی)، به حالتی که هر عنصر می‌تواند با درجه عضویت‌های مختلف داخل چندین خوشه قرار بگیرد، دسته بندی‌هایی که انطباق بیشتری با واقعیت دارند ارائه کنیم. یکی از اولین روش‌های وشهبندی فازی که بر مبنای تابع هدف و استفاده از فاصله اقلیدسی بنا شده بود درسال ۱۹۷۴توسط دان ارائه داده شد.

برای یک مجموعه داده $D = \{X_j\}_{j=1}^{N}$ که $X_j = (x_{j1}, x_{j2}, \dots, x\_jd) \in \mathcal{R}^d$، الگوریتم خوشه‌بندی فازی c-meansسعی در کمینه کردن تابع هدف زیر دارد[۹]:

$$J(U,V;D) = \sum_{i=1}^{C}\sum_{j=1}^{N} \mu_{ij}^{m}\, d_{ij}^{2} = \sum_{i=1}^{C}\sum_{j=1}^{N} \mu_{ij}^{m} \|X_j - V_i\|^2 \qquad (1)$$

که در اینجا $U = (\mu_{ij})_{C\times N}$ ماتریس عضویت فازی است که در $\mu_{ij}$ میزان عضویت داده $X_j$ به خوشه $i$ ام را نشان می‌دهد، $V = (V_1, V_2, \dots, V_C)^T = (v_{iq})_{C\times d}$ ماتریس مراکز است که مرکز $C$ خوشه را نشان می‌دهد، $m$ یک عدد حقیق بزرگ از یک است که توان فازی سازی را نشان می‌دهد، همچنین، $\|\cdot\|$ بیانگر نرم اقلیدسی است. لازم به ذکر است که مقدار تابع عضویت برای هر داده به هر

ISSN: 2745-3673
http://www.Rcsj.ir

خوشه یک عدد حقیقی بین صفر و یک است($\mu_{ij}\epsilon[0,1]$) که باید شرط $\sum_{i=1}^{c}\mu_{ij}=1$ را ارضا کند. روابط برزو رسانی توابع تعلق و همچنین مراکز خوشه به ترتیب به صورت زیر می‌باشد:

$$\mu_{ij}=\frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}^2}{d_{kj}^2}\right)^{\frac{1}{m-1}}} \quad (2)$$

$$V_i=\frac{\sum_{j=1}^{N}\mu_{ij}^m X_j}{\sum_{j=1}^{N}\mu_{ij}^m} \quad (3)$$

مراحل خوشه‌بندی فازی c-meansرا می‌توان در ۴ مرحله زیر بیان کرد:

۱. مقدار دهی اولیه برای $C$، $m$ و $U$.
۲. بروز رسانی مراکز خوشه‌ها با استفاده از رابطه ۳.
۳. بروز رسانی توابع تعلق داده‌ها با استفاده از رابطه ۲.
۴. بررسی شرط خاتمه و در صورت عدم خاتمه الگوریتم تکرار مرحله دوم.

## ۲. روش پیشنهادی

در الگوریتم استاندارد خوشه‌بندی فازی c-meansارزش تمام ویژگی‌ها برابر در نظر گرفته می‌شود، این در حالی است که در مسائل واقعی ممکن است بعضی از ویژگی‌ها نسبت به سایر ویژگی دارای ارزش بیشتری باشند. در نتیجه اگر ارزش تمام ویژگی‌ها برابر در نظر گرفته شود ممکن دقت خوشه‌بندی پایین بیاید. برای حل این مشکل در روش پیشنهادی در این مقاله از وزن‌دهی ویژگی‌ها برای بهبود عملکرد خوشه‌بندی فازی استفاده شده است. در ادامه این بخش رابطه‌های مربوط الگوریتم خوشه‌بندی فازی وزن دهی شده شرح داده می‌شود.

برای مجموعه داده ای $D=\{X_j\}_{j=1}^{N}$ که $X_j=(x_{j1},x_{j2},\dots,x\_jd)\in\mathcal{R}^d$، الگوریتم خوشه‌بندی فازیوزن دهی شده،سعی در کمینه کردن تابع هدف زیر دارد:

$$J(U,V;D)=\sum_{i=1}^{C}\sum_{j=1}^{N}\mu_{ij}^m\left[d_{ij}^{(w)}\right]^2 \quad (4)$$

در اینجا $d_{ij}^{(w)}=\left\|diag(w)(X_j-V_i)\right\|$ که $w=(w_1,w_2,\dots,w_d)$ بردار وزن ویژگی‌ها است و

$$diag(w)=\begin{pmatrix} w_1 & 0 & \dots & 0\\ 0 & 0 & \dots & 0\\ \dots & \dots & \dots & \dots\\ 0 & 0 & \dots & w_d\end{pmatrix}$$

همچنین، عناصر $\mu_{ij}$ در ماتریس تعلق $U$ و بردار $w_q$ به ترتیب باید شروط زیر را ارضا کنند:

$$\sum_{i=1}^{c}\mu_{ij}=1 \quad (5)$$

$$\sum_{q=1}^{a}w_q=1 \quad (6)$$

در نهایت روابط بروزسانی برای توابع تعلق و مراکز خوشه‌ها و همچنین وزن ویژگی‌ها به ترتیب در فرمول‌های ۷ تا ۹نشان داده شده است:

$$\mu_{ij}=\frac{1}{\sum_{k=1}^{c}\left(\frac{\left[d_{ij}^{(w)}\right]}{\left[d_{kj}^{(w)}\right]}\right)^{\frac{1}{m-1}}} \quad (7)$$

ISSN: 2745-3673
http://www.Rcsj.ir

$$V_i = \frac{\sum_{j=1}^{N} \mu_{ij}^{m} X_j}{\sum_{j=1}^{N} \mu_{ij}^{m}} \quad (8)$$

$$w_q^t = \frac{1}{\sum_{l=1}^{d} \left( \frac{\sum_{i=1}^{C} \sum_{j=1}^{N} [\mu_{ij}^t]^m (x_{jq} - v_{iq})^2}{\sum_{i=1}^{C} \sum_{j=1}^{N} [\mu_{ij}^t]^m (x_{jl} - v_{il})^2} \right)} \quad (9)$$

لازم به ذکر است که ماتریس تعلق اولیه و مراکز خوشه‌های اولیه به صورت تصادفی انتخاب می‌شود در حالی که بردار وزن ویژگی‌های اولیه با استفاده از معیار امتیاز لاپلاسین [۱۰] که یکی از معیارهای پر استفاده در انتخاب ویژگی در حالت بدون ناظر است،مقداردهی اولیه می‌شوند.

امتیاز لاپلاسین۶ (LS) [۱۰] یک روش مبتنی بر تئوری گراف است که در هر دو حالت با ناظر و بدون ناظر عمل می‌کند. امتیاز لاپلاسین فضای داده‌ها را به یک گراف مدل می‌کند و مبتنی بر این عقیده است که اگر دونقطه داده‌ای نزدیک به یکدیگر باشند احتمالاً به یک کلاس تعلق دارند. درواقع این روش از ساختار محلی فضای داده برای انتخاب ویژگی استفاده می‌کند. امتیاز لاپلاسین برای ویژگی $A$ و با توجه به مجموعه الگوهای $S$ با استفاده از رابطه (۱۰) محاسبه می‌شود.

$$LS(S,A) = \frac{\sum_{i,j} (A(i) - A(j)) S_{ij}}{\sum_{i} (A(i) - \bar{A}) D_{ii}} \quad (10)$$

که در این رابطه $A(i)$ مقدار ویژگی $A$ در الگوی $i$ ام در نشان می‌دهد، $\bar{A}$ میانگین ویژگی $A$ را مشخص می‌کند، $D$ یک ماتریس قطری است که $\sum_j S_{ij}$ همچنین $S_{ij}$ رابطه همسایگی بین الگوها را نشان می‌دهد که به‌صورت رابطه (۱۱) محاسبه می‌شود.

$$S_{ij} = \begin{cases} e^{\frac{x_i - x_j}{t}}, & if\ x_i\ and\ x_j\ are\ neighbors \\ 0, & Otherwise \end{cases} \quad (11)$$

که در اینجا $t$ یک ضریب ثابت، و دو الگوی $x_i$ و $x_j$ همسایه در نظر گرفته می‌شوند اگر یکی از آن‌ها جزء $K$-نزدیک‌ترین همسایه‌های دیگری باشد.

این کار از دو جهت سبب بهبود عملکرد روش پیشنهادی می‌شود. از یک طرف وزن‌های نسبت داده شده به ویژگی‌ها با ارزش ویژگی‌ها متناسب است و از طرف دیگر زمان همگرایی الگوریتم خوشه‌بندی سریع تر می‌شود. در الگوریتم ۱ شبه کد روش پیشنهادی با عنوان خوشه‌بندی فازی با استفاده از وزن دهی ویژگی‌ها نشان داده شده است.

## ۳. نتایج آزمایش‌ها

در این بخش، عملکرد روشهای پیشنهادی برای مسئله خوشه‌بندی داده‌ها مورد ارزیابی قرار می‌گیرد. به این منظور، روش پیشنهادی با جدیدترین و شناخته شده ترین روشهایخوشه‌بندی فازی، مقایسهمی‌شود. در این مقاله، از چندین مجموعه دادهای، با مشخصات مختلف جهت ارزیابی روشهای پیشنهادی و مقایسه عملکرد آن با سایر روشهایخوشه‌بندی فازی استفاده شده است. این مجموعه‌های داده ای شامل Ionosphere، Pima و Vehicle1می‌باشد که به طور گسترده در مسائل خوشه‌بندی مورد استفاده قرار می‌گیرند. مشخصاتکلیاینمجموعه‌هایدادهایدرجدول ۱ نشانداده شدهاست. لازم به ذکر است برای تمام مجموعه‌های داده ای تعداد کلاس‌ها آن مجموعه داده ای را به عنوان تعداد خوشه‌ها در نظر گرفته شده است. همچنین، مقدار $m$ را برابر ۱،۵ قرار داده شده است.

[6] Laplacian score

ISSN: 2745-3673
http://www.Rcsj.ir

**جدول ۱. مشخصات مجموعه‌های داده ای**

| مجموعه داده ای | تعداد نمونه‌ها | تعداد ویژگی‌ها | ضریب نامتعادلی |
|---|---|---|---|
| Ionosphere | 351 | 34 | 0.57 |
| Pima | 768 | 8 | 0.5 |
| Vehicle1 | 4230 | 18 | 0.35 |

جدول ۲ تا ۴ نرخ خطای خوشه‌بندی و همچنین وزن ویژگی‌های مختلف را برای روش‌های مختلف نشان می‌دهد و به ترتیب برای مجموعه‌های داده ای Ionosphere، Pima و Vehicle1نشان می‌دهد. همان طور که داده‌های این جدول نشان می‌دهند در هر سه مجمومه داده ای مختلف روش پیشنهادی با خطای کمتری خوشه‌بندی را انجام داده است.

**جدول۲. مقایسه خطای روش‌های مختلف خوشه‌بندی بر روی مجموعه داده ای Ionosphere**

| روش خوشه‌بندی | نرخ خطا (درصد) |
|---|---|
| خوشه‌بندی فازی c-means استاندارد | 26.65 |
| وانگ و همکارانش [۱۱] | 17.36 |
| هونگ و هکارانش[۱۲] | 9.81 |
| روش پیشنهادی | 4.97 |

**جدول۳. مقایسه خطای روش‌های مختلف خوشه‌بندی بر روی مجموعه داده ای Pima**

| روش خوشه‌بندی | نرخ خطا (درصد) |
|---|---|
| خوشه‌بندی فازی c-means استاندارد | 28.39 |
| وانگ و همکارانش [۱۱] | 17.91 |
| هونگ و هکارانش[۱۲] | 19.61 |
| روش پیشنهادی | 12.78 |

**جدول٤. مقایسه خطای روش‌های مختلف خوشه‌بندی بر روی مجموعه داده ای Vehicle1**

| روش خوشه‌بندی | نرخ خطا (درصد) |
|---|---|
| خوشه‌بندی فازی c-means استاندارد | 35.66 |
| وانگ و همکارانش [۱۱] | 18.76 |
| هونگ و هکارانش[۱۲] | 15.98 |
| روش پیشنهادی | 7.31 |

ISSN: 2745-3673
http://www.Rcsj.ir

جدول ۳ زمان اجرا را برای روش‌های مختلف خوشه‌بندی در مجموعه‌های داده ای مختلف نشان می‌دهد. همان طور که در این جدول دیده می‌شود روش پیشنهادی فقط نسبت به روش خوشه‌بندی فازی c-meansدر حالت استاندارد زمان بیشتری را صرف کرده. این در حالی است که روش پیشنهادی نسبت به خوشه‌بندی فازی c-means در حال استاندارد دارای خطای بسیار کمتری است.

**جدول۵. مقایسه زمان اجرا در روش‌های مختلف خوشه‌بندی (بر حسب میلی ثانیه)**

| | خوشه‌بندی فازی استاندارد | وانگ و همکارنس[۱۱] | هونگ و همارانش[۱۲] | روش پیشنهادی |
|---|---|---|---|---|
| Ionosphere | **35** | **2378** | **69** | **51** |
| Pima | **79** | **6189** | **198** | **105** |
| Vehicle1 | **489** | **231891** | **911** | **819** |

## ۴. نتیجه گیری

مسئله خوشه بندی یکی از مسائل مهمی‌است که از دیرباز توسط محققان مورد توجه بوده است. خوشه بندی فازی از تلفیق رویکرد فازی در بحث خوشه بندی برای کاربردی تر نمودن آن و انطباق بیشتر با دنیای واقعی حاصل شده است. یکی از انواع چالش بر انگیز داده ها، مجموعه های داده ای نامتعادل است. به یک مجموعه داده نامتعادل گفته می شود اگر بسیاری از نمونه های موجود در آن مجموعه داده متعلق به یک کلاس و تعداد کمی از نمونه ها متعلق به دسته های دیگر باشد. به عبارتی دیگر یک مجموعه را می توان مجموعه داده ای نا متعادل نامید در حالی که حداقل یک کلاس در آن وجود داشته باشد که تعداد نمونه های آموزش مربوط به آن کلاس بسیار کم باشد. به این کلاس که تعداد نمونه های کمی در مجموعه آموزشی دارد کلاس اقلیت گفته می شود و سایر کلاس هایی که دارای نمونه های آموزش زیادی باشند کلاس اکثریت نامیده می شوند. در این حالت طبقه بندی برای داده هایی که عضو کلاس اکثریت باشند به خوبی عمل می کند و همچنین برای داده های کلاس اقلیت بسیار ضعیف و همراه با خطا می باشد. مطالعه و یادگیری مجموعه داده های نامتعادل از موضوع های مهم و چالش برانگیزی است که در حوزه یادگیری ماشین مطرح شده است. در این مقاله برای خوشه بندی داده های نامتعادل یک روش جدید مبتنی بر تئوری فازی ارائه شده است. یکی از الگوریتم‌های پایه ای خوشه‌بندی فازی، الگوریتم خوشه‌بندی فازیc-means است. یکی از مشکلات این الگوریتم این است که ارزش تمام ویژگی‌ها برابر در نظر گرفته می‌شود. برای حل این مشکل در این مقاله از یک راهکار خوشه‌بندی فازی وزن دهی شده استفاده شده است که در آن وزن ویژگی‌ها و همچنین مراکز خوشه‌ها به صورت همزمان بروزرسانیمی‌شوند.. آزمایش‌های انجام شده بر روی سه مجموعه داده واقعی نشان داد که روش پیشنهادی در مقایسه با سایر الگوریتم‌های پیش تر ارائه شده با دقت بهتری قادر خواهد بود که خوشه بندی داده‌ها را انجام دهد.

.

ISSN: 2745-3673
http://www.Rcsj.ir

**منابع و مراجع**

[1]. Nekooeimehr, I. and S.K. Lai-Yuen, *Adaptive semi-unsupervised weighted oversampling (A-SUWO) for imbalanced datasets.* Expert Systems with Applications, 2016. **46**: p. 405-416.

[2]. Zhang, T., L. Chen, and F. Ma, *A modified rough c-means clustering algorithm based on hybrid imbalanced measure of distance and density.* International Journal of Approximate Reasoning, 2014. **55**(8): p. 1805-1818.

[3]. Fan, J., et al., *Probability model selection and parameter evolutionary estimation for clustering imbalanced data without sampling.* Neurocomputing, 2016. **211**: p. 172-181.

[4]. Dervis Karaboga and C. Ozturk, *A novel clustering approach: Artificial Bee Colony (ABC) algorithm.* Applied Soft Computing, 2011. **11**: p. pp. 652-657.

[5]. Xing, H.-J. and M.-H. Ha, *Further improvements in Feature-Weighted Fuzzy C-Means.* Information Sciences, 2014. **267**(0): p. 1-15.

[6]. Patra, B. and S. Nandi, *Fast Single-Link Clustering Method Based on Tolerance Rough Set Model*, in *Rough Sets, Fuzzy Sets, Data Mining and Granular Computing*, H. Sakai, et al., Editors. 2009, Springer Berlin Heidelberg. p. 414-422.

[7]. Murtagh, F., *Hierarchical Clustering*, in *International Encyclopedia of Statistical Science*, M. Lovric, Editor. 2011, Springer Berlin Heidelberg. p. 633-635.

[8]. Wang, L.-X., *A Course in Fuzzy Systems and Control.* Prentice-Hall International, Inc, 1997.

[9]. Bezdek, J.C., *Pattern Recognition with Fuzzy Objective Function Algorithm.* Plenum Press, New York, 1981.

[10]. Xiaofei He, Deng Cai, and P. Niyogi1, *Laplacian Score for Feature Selection.* Adv. Neural Inf. Process. Syst, 2005. **18**: p. 507-514.

[11]. X.Z. Wang, Y.D.W., L.J. Wang, *Improving fuzzy c-means clustering based on feature-weight learning.* pattern Recognition Letters, 2004. **25**: p. pp. 1123–1132.

[12]. W.L. Hung, M.S.Y., D.H. Chen, *Bootstrapping approach to feature-weight selection in fuzzy c-means algorithms with an application in color image segmentation.* pattern Recognition Letters, 2004. **29**: p. pp. 1317-1325.