سیستم های فازی و کاربردها
سال هشتم، شماره اول (بهار و تابستان ۱۴۰۴)
صص. ۲۷ تا ۳۸
شاپای چاپی: ۲۷۱۷-۴۴۰۹
شاپای الکترونیکی: ۲۷۱۷-۳۹۸۴
DOI:10.22034/JFSA.2025.516509.1269

ارایه روشی برای انتخاب ویژگی مبتنی بر آنتروپی فازی

زینب رئیسی، دکتر حمیدرضا ملکی* و دکتر رضا اکبری*

دانشکده علوم، دانشگاه صنعتی شیراز، شیراز، ایران
دانشکده علوم، دانشگاه صنعتی شیراز، شیراز، ایران
دانشکده مهندسی فناوری اطلاعات و ارتباطات، دانشگاه صنعتی شیراز، شیراز، ایران

تاریخ دریافت: ۱۴۰۴/۰۱/۲۵ تاریخ پذیرش: ۱۴۰۴/۰۲/۳۱

نوع مقاله: علمی-پژوهشی

چکیده. در عصر دیجیتال داده‌ها به عنوان یک دارایی با ارزش توسعه یافته‌اند و سهم بزرگی درعملکرد روش‌های یادگیری ماشین دارند. با گسترش داده‌هایی با ابعاد بالا، انتخاب ویژگی‌های مرتبط و حذف ویژگی‌های زاید گامی موثر در بهبود کارایی الگوریتم‌های یادگیری بشمار می‌آید. یکی از روش‌های متداول برای انتخاب ویژگی، روش‌های مبتنی بر فیلتر است. در روش‌های مبتنی بر فیلتر معیارهای متفاوتی برای ارزیابی ویژگی‌ها بکار گرفته می‌شود. در این پژوهش معیاری جدید مبتنی بر آنتروپی فازی برای شناسایی ویژگی‌های مرتبط و افزونه پیشنهاد شده است. روش پیشنهادی با کاهش ویژگی‌های انتخابی علاوه بر کاهش فضای مورد نیاز و کاهش زمان پردازش داده، موجب بهبود عملکرد الگوریتم یادگیری شده است. کیفیت روش پیشنهادی با بررسی سه مجموعه‌داده از پایگاه‌های داده UCI و $Kaggle$ ارزیابی شده است. نتایج نشان می‌دهد روش پیشنهادی در مقایسه با روش انتخاب ویژگی بر مبنای آنتروپی، قادر به بهبود دقت طبقه‌بندی است؛ هم‌چنین زمان مورد نیاز برای پردازش داده کاهش می‌یابد.

2010 *Mathematics Subject Classification.* 68T27 ; 03B52
* Corresponding author
E-mails: Z.Raeisi@sutech.ac.ir, maleki@sutech.ac.ir and Akbari@sutech.ac.ir.
عبارات و کلمات کلیدی. انتخاب ویژگی،آنتروپی فازی، طبقه‌بندی، بهینه‌سازی، یادگیری‌ماشین .

۱. مقدمه

ابعاد بسیار بزرگ داده‌ها در دنیای واقعی چالش‌ها و فرصت‌های اساسی را برای بکارگیری روش‌های یادگیری ماشین[۱] و داده‌کاوی بوجود آورده است. کار با داده‌های با ابعاد بالا علاوه بر این‌که پیچیدگی محاسباتی بالایی را در پی دارد، از نظر زمان محاسبات و اشغال حافظه مورد نیاز نیز مقرون‌به‌صرفه نمی‌باشد؛ از این‌رو برای حل این مشکل انتخاب ویژگی (FS[۲]) از اهمیت بالایی برخوردار است. انتخاب ویژگی فرایند کاهش ابعاد داده‌های ورودی با حذف ویژگی‌های غیرمرتبط و افزونه و انتخاب یک زیرمجموعه مرتبط از مجموعه ویژگی‌های اولیه با هدف توسعه یک مدل یادگیری ماشین است؛ هدف انتخاب ویژگی در مسایل یادگیری نظارت شده، که در آن برچسب کلاس نمونه‌ها مشخص است؛ انتخاب زیرمجموعه‌ای از ویژگی‌ها است که قادر به طبقه‌بندی نمونه‌ها در کلاس‌های مختلف باشند؛ به‌طوری‌که بیشترین دقت طبقه‌بندی را داشته باشیم.

تکنیک‌های انتخاب ویژگی به سه دسته فیلتر [۳] بسته‌بندی [۴] و تعبیه‌شده [۵] قابل تقسیم است. در روش‌های مبتنی بر فیلتر، اهمیت ویژگی‌ها براساس مشخصات ذاتی آن‌ها و با استفاده از معیارهای آماری مشخص می‌شود و ویژگی‌هایی با بیشترین ارزش انتخاب می‌شوند. در این روش‌ها الگوریتم‌های یادگیری نقشی ندارند. در روش‌های مبتنی بربسته‌بندی در هر مرحله ابتدا زیرمجموعه‌ای از ویژگی‌ها انتخاب و عملکرد الگوریتم یادگیری برروی این مجموعه مورد بررسی قرار می‌گیرد و از نتیجه اجرای الگوریتم برای ارزیابی زیر مجموعه انتخاب شده استفاده می‌شود. سپس مجموعه انتخابی طبق نتایج بروزرسانی شده و در پایان زیرمجموعه‌ای که دارای بیشترین دقت [۶] است، انتخاب می‌شود. روش‌های فیلتر اگرچه از نظر محاسباتی از روش‌های بسته‌بندی سریع‌تر هستند اما با کمترین دقت همراه است. هم‌چنین تاثیر زیر مجموعه انتخابی در طبقه‌بندی نادیده گرفته می‌شود. روش‌های بسته‌بندی با وجود دقت بالاتر، از نظر محاسباتی کندتر می‌باشند. از این رو برای غلبه بر این مشکل روش‌های جایگزینی معرفی می‌شوند که با حفظ مزایای دو روش قبل کارآیی بیشتری نیز دارند. روش‌های تعبیه، مبادله‌ای بین روش‌های فیلتر و بسته‌بندی هستند؛ در این روش‌ها انتخاب ویژگی در فرایند آموزش مدل انجام می‌شود. روش‌های تعبیه نیاز به ارزیابی مجموعه ویژگی‌ها به صورت تکراری ندارند و بسیار کارآمدتر هستند [۱]. معیارهای متفاوتی برای یافتن ویژگی‌های مناسب وجود دارد مانند: معیارهای مبتنی بر شباهت، معیارهای مبتنی افزونگی، معیارهای مبتنی

[1] Machine Learning
[2] Feature Selection
[3] Filter
[4] Wrapper
[5] Embedded
[6] Accuracy

بر وابستگی و معیارهای مبتنی بر دقت. چهار گام ضروری در فرایند انتخاب ویژگی به صورت زیر است [۱]:

- ایجاد یک زیرمجموعه از ویژگی‌ها؛
- ارزیابی زیرمجموعه ویژگی؛
- بررسی شرایط خاتمه؛
- اعتبارسنجی نتایج.

عدم قطعیت از حوزه‌های کاربردی ریاضیات در زمینه‌های داده‌کاوی و طبقه‌بندی است. پس از معرفی نظریه مجموعه‌های فازی[1] توسط پروفسور زاده، کاربردهای این نظریه در حیطه‌های مختلف علوم کامپیوتر مورد توجه قرار گرفته است. ازجمله کاربردهای نظریه سیستم‌های فازی در علوم کامپیوتر می‌توان به پایگاه داده اشاره کرد، که منجر به ظهور داده‌کاوی فازی شده است. به عنوان یکی از کاربردها می‌توان به حالتی که داده‌ها در پایگاه داده فازی باشند اشاره کرد. برای بررسی این داده‌ها، معیارهای فازی که برای سنجش میزان شباهت و ارتباط بین بردارهای فازی کاربرد دارد، مورد استفاده قرار می‌گیرند.

یکی از معیارهای مبتنی بر احتمال شناخته شده که برای ارزیابی ویژگی‌ها در روش‌های مبتنی بر فیلتر مورد استفاده قرار می‌گیرد، آنتروپی است. برای محاسبه آنتروپی به محاسبه احتمال رخ دادن ویژگی‌ها نیاز است؛ زمانی که ویژگی‌ها از جنس داده‌های عددی پیوسته باشند، محاسبه این احتمال در مقایسه با داده‌های کیفی کار راحتی نیست. برای حل این مشکل ایده استفاده از آنتروپی فازی که بر مبنای نظریه مجموعه‌های فازی می‌باشد کارآمد است. در محاسبه مقدار آنتروپی فازی تنها تعداد ویژگی‌ها درنظر گرفته نمی‌شود؛ بلکه توزیع واقعی ویژگی‌ها، بر اساس تابع عضویت در نظر گرفته می‌شود. بنابراین آنتروپی فازی اطلاعات بیشتری را در توزیع واقعی ویژگی‌ها نسبت به آنتروپی شانون در فضای ویژگی منعکس می‌کند. هم‌چنین برای محاسبه مقدار آنتروپی فازی نیازی به محاسبه احتمال در معنای مرسوم نیست.

در این پژوهش برای محاسبه آنتروپی فازی، با بکارگیری ایده الگوریتم $C-Means$ رویکردی برای محاسبه میزان عضویت هر بردار ویژگی در کلاس هدف را پیشنهاد می‌دهیم. با بکارگیری رویکرد پیشنهادی و با استفاده از آنتروپی فازی به انتخاب ویژگی‌های مطلوب خواهیم پرداخت. در پایان به مقایسه عملکرد روش پیشنهادی با روش انتخاب ویژگی مبتنی بر آنتروپی مرسوم می‌پردازیم.

[1] Fuzzy Set Theory

ساختار کلی ارایه شده در این مقاله بدین شرح است که در بخش ۲، به بیان پیشینه‌ای از پژوهش‌های صورت گرفته در گذشته می‌پردازیم؛ پس از آن در بخش ۳، مساله مورد بررسی بیان خواهد شد. در بخش ۵، به شرح روش پیشنهادی برای حل مساله می‌پردازیم و در پایان این بخش نتایج عددی حاصل از بکارگیری روش پیشنهادی گزارش داده می‌شود. نتیجه کلی حاصل از پژوهش در بخش ۶ بیان می‌شود.

۲. پیشینه پژوهش

در این بخش، به بررسی برخی پژوهش‌های صورت گرفته در گذشته، در زمینه انتخاب ویژگی خواهیم پرداخت.

مساله انتخاب ویژگی از گذشته تاکنون یک موضوع تحقیقاتی و کاربردی در زمینه‌های گوناگون داده‌کاوی، آمار و یادگیری‌ماشین بوده است. نارندرا[1] و همکارش در سال ۱۹۷۷ از الگوریتم شاخه‌وکران برای انتخاب زیرمجموعه‌های m عضوی از n ویژگی استفاده کردند؛ روش پیشنهادی آن‌ها برخلاف جستجوی کامل که از نظر محاسباتی غیر ممکن است، یک روش کارآمد است و بهترین زیر مجموعه ویژگی را بدون جستجوی کامل انتخاب می‌کند [۲]. روش‌های انتخاب ویژگی شامل دو مرحله مهم ارزیابی زیر مجموعه ویژگی انتخابی و جستجو در فضای ویژگی است؛ الگوریتم‌های متفاوتی برای ارزیابی خوب بودن زیرمجموعه ویژگی بکار برده می‌شود؛ برای این منظور دش[2] وهمکارانش در سال ۲۰۰۳ معیار ناسازگاری را مورد استفاده قرار دادند. در حقیقت یک زیر مجموعه ویژگی ناسازگار است اگر حداقل دو نمونه با مقادیر ویژگی یکسان اما با برچسب‌های کلاس متفاوت در آن وجود داشته باشد. آن‌ها معیار ناسازگاری را با سایر معیارها مقایسه کرده و استراتژی‌های جستجوی مختلف مانند جستجوی جامع، جستجوی اکتشافی و تصادفی را که می‌توان بر این معیارها اعمال کرد، مورد مطالعه قرار دادند [۳]. یامادا[3] و همکاران در سال ۲۰۱۴ از معیار وابستگی برای ارزیابی ویژگی‌ها در کار خود بهره بردند [۴]. باهاسین[4] و همکاران در سال ۲۰۲۰ یک روش اصلاح شده برای طبقه‌بندی متن عربی پیشنهاد کردند که از انتخاب ویژگی مبتنی بر Chi-Square برای بهبود عملکرد طبقه‌بندی استفاده می‌کند. آن‌ها برای طبقه‌بندی از الگوریتم SVM[5] و DT[6] استفاده کردند [۵].

[1] Narendra
[2] Dash
[3] Yamada
[4] Bahassine
[5] Support Vector Machine
[6] Decision Tree

از زمان مطرح شدن نظریه مجموعه‌های فازی در سال ۱۹۶۵ توسط پروفسور زاده [۶]، توسیع‌هایی از مجموعه‌های فازی برای اهداف متفاوت ارایه شده است. در سال ۲۰۰۱ لی[1] و همکاران از انتخاب ویژگی بر اساس آنتروپی فازی برای طبقه‌بندی فازی استفاده کردند. در نتیجه این پژوهش، هم پیچیدگی و هم بار محاسباتی طبقه‌بندی کننده کاهش یافته و بنابراین زمان آموزش و زمان طبقه‌بندی کاهش می‌یابد [۷]. در سال ۲۰۱۵ هوک[2] و همکاران روشی برای انتخاب ویژگی برمبنای اطلاعات متقابل فازی با پاسخ نامغلوب ارایه دادند. نتایج کار آنها با بکارگیری داده‌های پایگاه داده UCI نشان دهنده عملکرد بالای روش پیشنهادی آنها است [۸]. رواناسیداپا[3] و همکارش در سال ۲۰۱۸ یک روش انتخاب ویژگی بر اساس آنتروپی فازی شهودی (IFS[4]) برای طبقه‌بندی متن ارایه کردند، آنها برای ارزیابی عملکرد طبقه‌بند کننده از معیار F استفاده کردند. آنها نشان دادند که روش پیشنهادی عملکرد بهتری در مقایسه با سایر روش‌های FS دارد [۹].

۳. بیان مساله

در یادگیری نظارت شده به منظور طبقه‌بندی، فرض کنید داده‌های ورودی به عنوان یک ماتریس $n \times (p+1)$ درنظر گرفته شود که در آن n تعداد نمونه‌ها و $p+1$ بُعد هر یک از نمونه‌ها است؛ بدین صورت که هر نمونه شامل p ویژگی f_1،...f_p است، و مولفه $p+1$ام برچسب کلاسی که حاوی این نمونه است را نمایش دهد. آنگاه مساله‌ای که با آن مواجه هستیم، یافتن یک زیر مجموعه بهینه از مجموعه $\{f_1, f_2, ...f_p\}$، به نام S، است به طوری‌که $|S| \leq p$ و ماتریس کاهش یافته $n \times (|S|+1)$ اطلاعات کافی برای طبقه‌بندی و پیش‌بینی کلاس C را در اختیار ما قرار دهد. برای انتخاب چنین زیر مجموعه‌ای که تعداد حالت‌های ممکن آن به صورت 2^p قابل بیان است، همان‌طور که پیش‌تر اشاره شد، با یک مساله NP-سخت مواجه هستیم. برای رسیدن به چنین هدفی از منظر داده‌کاوی نیاز به حذف برخی از ویژگی‌ها است.

در بیشتر مواقع برخی از ویژگی‌های موجود در پایگاه داده را با بکارگیری معیارهای آماری و ریاضی، در صورت بهبود عملکرد طبقه‌بندی، می‌توان در تصمیم گیری‌ها نادیده گرفت. اگر تابع معیاری که به دنبال بهینه کردن آن هستیم را با $D(X)$ نشان دهیم که در آن $X \in \mathbb{B}^p$، $\mathbb{B}^p$ نشان دهنده فضای باینری p بُعدی و x_j مولفه jام بردار X است؛ آنگاه $x_j = 1$ نشان دهنده انتخاب ویژگی f_j و $x_j = 0$ نشان دهنده عدم انتخاب ویژگی f_j است. بنابراین مساله مورد بررسی را به

[1] Li
[2] Hoque
[3] Revanasiddappa
[4] Intuitionistic Fuzzy Sets

طور کلی به صورت مساله بهینه‌سازی زیر می‌توان در نظر گرفت.

$$\begin{aligned} &\max: \quad D(X), \\ &\text{s.t.} \quad \|X\|_1 \leq p. \end{aligned} \tag{۱.۳}$$

یکی از معیارهایی که برای ارزیابی ویژگی‌ها بکار برده می‌شود، معیار واگرایی کولبک لیبلر است. با بکارگیری این معیار به عنوان یک تابع ارزیاب می‌توان تفاوت بین توزیع ویژگی‌ها را محاسبه کرد. معیار واگرایی کولبک لیبلر میان دو توزیع گسسته p و q با رابطه ۲.۳ قابل محاسبه است.

$$\mathbb{D}_{\mathrm{KL}}(p \parallel q) = \sum_{i=1}^{N} p(x_i) \log \left(\frac{p(x_i)}{q(x_i)} \right) \tag{۲.۳}$$

معیار دیگری که می‌توان برای ارزیابی بکار برد، معیار آنتروپی شانون است که به‌صورت زیر محاسبه می‌شود.

$$H(X) = -\sum_{x \in X} p(x) \log \left(p(x) \right) \tag{۳.۳}$$

بردار ویژگی که بالاترین میزان آنتروپی را داشته باشد بر سایر ویژگی‌ها ترجیح داده می‌شود. در این پژوهش تابع معیار D(X) را آنتروپی فازی بردارهای ویژگی درنظر گرفته‌ایم و به بررسی عملکرد این معیار بر تعداد ویژگی‌های انتخابی و دقت طبقه‌بندی می‌پردازیم. نحوه محاسبه این معیار در بخش ۵ بیان شده است.

۴. آنتروپی فازی

آنتروپی در نظریه مجموعه‌های فازی، به عنوان معیاری برای سنجش میزان «ابهام» یا «عدم قطعیت» در توصیف عناصر یک مجموعه فازی تعریف می‌شود. برخلاف آنتروپی شانون که بر مبنای مفاهیم احتمال بناشده، آنتروپی فازی بر پایه مقدار تابع عضویت استوار است. تعریف این مفهوم برای نخستین بار توسط لوکا[1] و ترمینی[2] در سال ۱۹۹۳ ارایه شد [۱۰].

فرض کنید $X = \{x_1, x_2, \ldots, x_n\}$ یک مجموعه از متغییرهای تصادفی باشد. $\mu_A(x_i)$ درجه عضویت عضو x_i از مجموعه فازی A باشد و F یک نگاشت بصورت $F: G(2^X) \rightarrow [0, 1]$. تابع F یک مجموعه فازی است که بر روی یک مجموعه فازی تعریف شده است. آن‌گاه F یک معیار برای محاسبه آنتروپی فازی است اگر در چهار اصل اساسی زیر (اصول لوکا-ترمینی) صدق کند.

[1] Luca
[2] Termini

- $F(A) = ۰$ اگر و تنها اگر A یک مجموعه کلاسیک باشد. (درجه عضویت تمام عناصر ۰ یا ۱ باشد.)
- $F(A) = ۱$ اگر و تنها اگر برای هر i، $\mu_A(x_i) = ۰/۵$.
- $F(A) \leq F(B)$ اگر،

$\mu_A(x_i) \leq \mu_B(x_i)$ وقتی‌که $\mu_B(x_i) \leq ۰/۵$ و $\mu_A(x_i) \geq \mu_B(x_i)$ وقتی‌که $\mu_B(x_i) \geq ۰/۵$

- $F(A) = F(A^c)$. به این معنا که آنتروپی فازی یک مجوعه برابر با آنتروپی فازی مجموعه مکمل آن است؛ جایی‌که A^c به صورت زیر تعریف می‌شود:

$$A^c = (۱ - \mu_A(x_۱), (۱ - \mu_A(x_۲)), \ldots, (۱ - \mu_A(x_n))$$

لوکا و ترمینی در مقاله خود، تابع زیر که تعمیمی از آنتروپی شانون است و در چهار اصل بالا صدق می‌کند، به‌عنوان آنتروپی فازی معرفی کردند.

$$H(A) = -\frac{۱}{n}\sum_i [\mu_A(x_i) log \mu_A(x_i) + (۱ - \mu_A(x_i)) log(۱ - \mu_A(x_i))]. \quad (۱.۴)$$

۵. روش حل و نتایج عددی

برای محاسبه آنتروپی فازی، اولین گام محاسبه درجه عضویت هر بردار ویژگی در کلاس هدف مساله طبقه‌بندی است. روش‌های متفاوتی برای محاسبه میزان عضویت ویژگی‌ها وجود دارد؛ مانند رویکرد مبتنی بر KNN [۱] و رویکرد مبتنی بر $C - means$. در ادامه از روش ارایه شده توسط خوشابا[۲] و همکاران [۱۱]، برای تخمین این درجه عضویت استفاده کرده‌ایم. پس از محاسبه درجه عضویت ویژگی‌ها، با استفاده از رابطه ۱.۵، آنتروپی فازی هر بردار ویژگی در کلاس‌های هدف را محاسبه می‌کنیم. در گام بعد آنتروپی فازی کل برای هر بردار ویژگی با استفاده از رابطه ۲.۵ محاسبه می‌شود و ویژگی‌ها را بر اساس مقادیر آنتروپی فازی کل رتبه‌بندی می‌کنیم. در انتها با استفاده از طبقه‌بند DT، عملکرد روش پیشنهادی را برای تعداد متفاوت ویژگی ارزیابی خواهیم کرد.

$$H(f_i, C_j) = -[\mu_{ij} log(\mu_{ij}) + (۱ - \mu_{ij}) log(۱ - \mu_{ij})]. \quad (۱.۵)$$

[1] K-Nearest Neighbors
[2] Khushaba

$$FE(f_i) = \frac{1}{n} \sum_j H(f_i, Cj). \quad (۲.۵)$$

برای ارزیابی روش پیشنهادی و مقایسه آن با روش انتخاب ویژگی بر اساس آنتروپی، عملکرد انتخاب ویژگی با استفاده از طبقه‌بندکننده DT برروی داده‌ها بررسی شده است. برای این منظور سه مجموعه داده به شرح جدول ۱ از پایگاه‌های داده UCI و $Kaggle$ انتخاب شده است. دقت طبقه‌بندی معیاریست که برای ارزیابی نهایی مد نظر قرار گرفته است. نتایج حاصل از پیاده‌سازی دو روش در شکل‌های ۱ تا ۳ قابل مشاهده است.

بررسی عددی نتایج عملکرد قابل قبول روش پیشنهادی را تایید می‌کند. نتایج بدست آمده برای مجموعه داده $ionosphere$ نشان می‌دهد، تنها با انتخاب ۱۵ ویژگی به روش آنتروپی فازی دقت طبقه‌بندی ۸۹٫۵۲ است؛ در حالی که با همین تعداد ویژگی با روش آنتروپی، دقت طبقه‌بندی ۷۸٫۱ است. بررسی نتایج برای مجموعه داده‌ی $Hill - valley$ نشان می‌دهد، اگرچه در تعداد ویژگی بالا دو روش عملکرد مشابهی دارند اما، زمانی که تعداد ویژگی انتخابی کمتر از ۵۰ ویژگی باشد، نتایج برتری روش آنتروپی فازی را نشان می‌دهد. با انتخاب یک مجموعه ۴۰ عضوی از ویژگی‌ها دقت طبقه‌بندی به روش آنتروپی فازی ۶۱٫۸۸ و به روش آنتروپی ۵۶٫۹۱ بدست آمده است. در مورد مجموعه داده $new - coords$ نتایج مشابهی بدست آمده است؛ با انتخاب یک مجموعه ۱۰ عضوی از ویژگی‌ها به روش آنتروپی فازی دقت طبقه‌بندی ۹۹٫۲، و برای زیرمجموعه انتخابی به روش آنتروپی دقت طبقه‌بندی ۹۸٫۲۲ است. نتایج تجربی نشان می‌دهد تنها با انتخاب کمتر از ۲/۳ از ویژگی های اولیه می‌توان نتایج قابل قبولی را بدست آورد و این به معنای صرفه جویی در زمان و حافظه مورد نیاز است. برای مقایسه بهتر نتایج عددی، میزان صحت دو روش برروی مجموعه داده‌های آزمایش، بدون حذف ویژگی و با انتخاب ۲/۳ از ویژگی‌های مرتبط در جدول ۲ آمده است. بررسی عددی این نتایج نیز عملکرد قابل قبول روش پیشنهادی را تایید می‌کند.

مزیت اصلی روش پیشنهادی در مقایسه با روش‌های کلاسیک مانند آنتروپی شانون، توانایی بهتر در پردازش داده‌های نادقیق، نویزی و همراه با عدم قطعیت است. در بسیاری از مسایل دنیای واقعی، داده‌ها ممکن است دارای مقادیر پیوسته، گم‌شده، نویز یا تغییرات تصادفی باشند که روش‌های سنتی را با چالش مواجه می‌کند. آنتروپی فازی با مدلسازی درجات تعلق و عدم قطعیت، انعطاف‌پذیری بیشتری در برابر این ناهمگونی‌ها از خود نشان می‌دهد و در نتیجه، برآوردهای پایدارتری از اهمیت ویژگی‌ها ارایه می‌کند. علاوه بر این، روش پیشنهادی با حذف ویژگی‌های زاید و کم‌اهمیت، باعث کاهش ابعاد داده و بهبود کارایی مدل طبقه‌بندی می‌شود. همانطور که در نتایج تجربی نشان داده

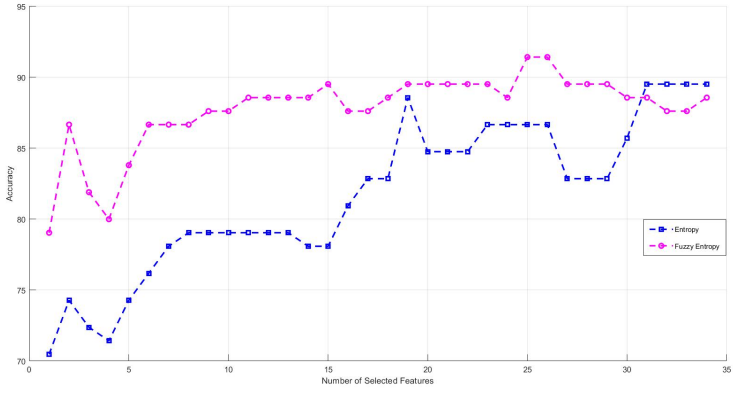
شده است، این رویکرد منجر به دقت بالاتر در مقایسه با روش‌های مبتنی بر آنتروپی کلاسیک شده است. این بهبود عملکرد به‌ویژه در مجموعه‌داده‌های با ابعاد بالا یا داده‌های نامتوازن مشهودتر است، جایی که انتخاب ویژگی‌های بهینه تأثیر مستقیمی بر کیفیت یادگیری ماشین دارد.

جدول ۱: مشخصات مجموعه داده‌های آزمایش.

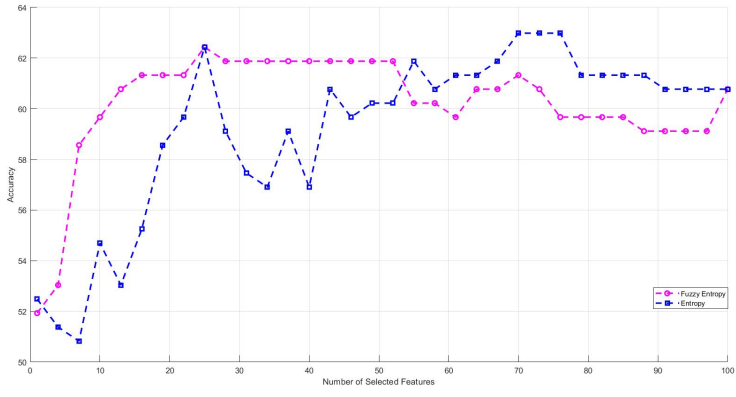
مجموعه داده	تعداد نمونه‌ها	تعداد ویژگی‌ها	تعداد کلاس‌ها
ionosphere	۳۵۶	۳۴	۲
Hill-Valley	۶۰۶	۱۰۰	۲
new-coords	۱۷۰۵۶	۱۳۲	۵

جدول ۲: مقایسه میزان صحت طبقه بندی (%) بدون انتخاب ویژگی با انتخاب ۲/۳ از داده‌ها.

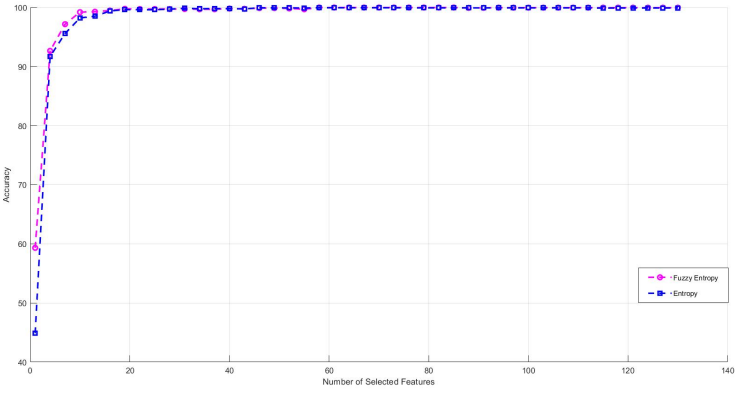
مجموعه داده	تعداد ویژگی‌ها	بدون حذف ویژگی	انتخاب ویژگی با روش آنتروپی	انتخاب ویژگی با روش آنتروپی فازی
ionosphere	۳۴	۸۸٫۵۷	۸۴٫۷۶	۸۹٫۵۲
Hill-Valley	۱۰۰	۶۰٫۷۷	۶۱٫۸۸	۶۰٫۷۷
new-coords	۱۳۲	۹۹٫۹۴	۹۹٫۹۲	۹۹٫۹۲



شکل ۱: دقت طبقه‌بندی بر اساس تعداد ویژگی انتخابی برای مجموعه داده ionosphere .



شکل ۲: دقت طبقه‌بندی بر اساس تعداد ویژگی انتخابی برای مجموعه داده Hill-Valley .



شکل ۳: دقت طبقه‌بندی بر اساس تعداد ویژگی انتخابی برای مجموعه داده New-coords .

۶. نتیجه‌گیری

در این پژوهش با استفاده از روش فیلتر مبتنی بر معیار آنتروپی فازی به انتخاب ویژگی‌های مرتبط و حذف ویژگی‌های زاید پرداختیم. بر اساس نتایج بدست آمده، انتخاب ویژگی براساس آنتروپی فازی در هر سه مجموعه داده مورد آزمایش عملکرد بهتری نسبت به انتخاب ویژگی بر اساس آنتروپی دارد. روش پیشنهادی علاوه بر افزایش عملکرد طبقه‌بندی، با کاهش تعداد ویژگی‌های

انتخابی باعث کاهش فضای مورد نیاز و زمان پردازش داده‌ها می‌شود. از سوی دیگر با وجود کاهش تعداد ویژگی‌ها دقت طبقه‌بندی افزایش یافته است.

روش پیشنهادی دارای کاربردهای عملی گسترده‌ای در حوزه‌های مختلف است. در حوزه پزشکی، این روش به‌ویژه در تحلیل تصاویر تشخیصی مانند ماموگرافی و داده‌های آزمایشگاهی که اغلب با مشکلات نویز و نقص داده مواجه هستند، عملکرد مطلوبی از خود نشان می‌دهد. در زمینه پردازش تصویر، روش پیشنهادی با توانایی انتخاب بهینه ویژگی‌های کلیدی از تصاویر، منجر به توسعه مدل‌های طبقه‌بندی کارآمدتر می‌شود. هم‌چنین در حوزه مالی که با حجم بالایی از داده‌های نامتوازن و پرنویز روبرو هستیم، این روش قادر است ویژگی‌های معنادار برای تشخیص تقلب را با دقت بالاتری شناسایی کند. مزیت اصلی این رویکرد، توانایی ذاتی آن در مدیریت عدم قطعیت و پردازش داده‌های نادقیق است که آن را به انتخابی مناسب برای مسائل دنیای واقعی تبدیل می‌کند.

برای توسعه این پژوهش، پیشنهاد می‌شود مطالعات آینده بر روی ترکیب روش پیشنهادی با سایر روش‌های انتخاب ویژگی مانند روش‌های مبتنی بر نظریه اطلاعات و یادگیری عمیق تمرکز کنند. این ترکیب‌پذیری می‌تواند به بهبود عملکرد مدل بینجامد. همچنین با توجه به ماهیت چندهدفه مساله انتخاب ویژگی، توسعه چارچوبی چندمعیاره با در نظر گرفتن معیارهای مکمل در کنار انتروپی فازی امکان تحلیل جامع‌تر مساله را فراهم می‌سازد. علاوه بر این، بکارگیری الگوریتم‌های فراابتکاری می‌تواند همزمان باعث بهبود سرعت و دقت مدل گردد. این مسیرهای پژوهشی می‌توانند به ارتقای کارایی روش‌های انتخاب ویژگی در مسائل پیچیده دنیای واقعی منجر شوند.

مراجع

[1] Li, J., Cheng, K., Wang, S., Morstatter, F.,Trevino, R, P., Tang, J., Liu, H. (2017) Feature selection: A data perspective. ACM computing surveys (CSUR), 50(6),1–45.

[2] Narendra and Fukunaga. (1977) A branch and bound algorithm for feature subset selection. IEEE Transactions on computers, 100(9), 917–922.

[3] Dash, M and Liu, H. (2003) Consistency-based search in feature selection. Artificial intelligence, 151(1-2), 155–176.

[4] Yamada, M., Jitkrittum, W., Sigal, L., Xing, E. P., Sugiyama, M. (2014) High-dimensional feature selection by feature-wise kernelized lasso. Neural computation, 26(1), 185–207.

[5] Bahassine, S., Madani, A., Al-Sarem, M., Kissi, M. (2020) Feature selection using an improved Chi-square for Arabic text classification. Journal of King Saud University-Computer and Information Sciences, 32(2), 225–231.

[6] Zadeh, L. A. (1965) Fuzzy sets. Information and control, 8(3), 338–353.

[7] Lee, H., Chen, Ch,. Chen, J., Jou, Y. (2001) An efficient fuzzy classifier with feature selection based on fuzzy entropy, IEEE transactions on systems, man, and cybernetics, part B (cybernetics), 32(3), 426–432.

[8] Hoque, N., Ahmed, HA., Bhattacharyya, DK., Kalita, JK. (2016) A fuzzy mutual information-based feature selection method for classification, Fuzzy Information and Engineering, 8(3), 355–384.

[9] Harish, BS and Revanasiddappa, MB. (2018) A new feature selection method based on intuitionistic fuzzy entropy to categorize text documents. International Journal of Interactive Multimedia and Artificial Intelligence ….

[10] De Luca, A and Termini, S. (1993) A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory. Elsevier, 197–202.

[11] Khushaba, R. N., Kodagoda, S., Lal, S., Dissanayake, G. (2010) Driver drowsiness classification using fuzzy wavelet-packet-based feature-extraction algorithm. IEEE transactions on biomedical engineering, 58(1), 121–131.