یک روش جدید برای انتخاب ویژگی مبتنی بر منطق فازی

حسن نصرتی ناهوک[۱] و مهدی افتخاری[۲]

۱- کارشناس ارشد کامپیوتر- هوش مصنوعی، گروه کامپیوتر- دانشگاه آزاد اسلامی واحد سراوان- سراوان- ایران
hsn.nosrati@gmail.com
۲- استادیار، بخش مهندسی کامپیوتر- دانشگاه شهید باهنر - کرمان- ایران
m.eftekhari@uk.ac.ir

چکیده: انتخاب ویژگی یکی از چالش برانگیز ترین و از مهمترین فعالیت‌ها در توسعه یادگیری ماشین و تشخیص الگوست. معیارهای ارزیابی ویژگی نقش بسیار مهمی برای ساخت یک الگوریتم انتخاب ویژگی دارند. در این مقاله یک معیار انتخاب ویژگی اصلاح شده با استفاده از منطق فازی برای انتخاب تعداد ویژگی‌های مورد نیاز ارائه می شود. این معیار به شکل غیر فازی در تحقیقات قبلی استفاده می‌شود، اما در این مقاله با تعریف تعداد ویژگی‌ها به صورت یک عدد فازی و با استفاده از اصل توسعه، شکل فازی معیار مزبور به دست آمد. عملکرد روش پیشنهادی بر روی مجموعه داده‌های منتشر شده از UCI ارزیابی شد و نتایج حاصل نشان دهنده کارایی روش مزبور در مقایسه با نسخه غیر فازی آن است.

واژه‌های کلیدی: انتخاب ویژگی، الگوریتم ژنتیک، منطق فازی، عدد فازی مثلثی.

۱- مقدمه

«ویژگی» و یا « موجودیت» و یا «متغیر» به جنبه ای از داده‌ها اشاره می کند. معمولاً قبل از جمع آوری داده‌ها، ویژگی‌ها مشخص یا انتخاب شده اند. ویژگی‌ها می‌توانند گسسته، پیوسته، یا اسمی باشند به طور کلی، ویژگی‌ها به صورت زیر وصف می‌شوند: مربوط: ویژگی‌هایی وجود دارند که بر خروجی تأثیر دارند و نقش آنها با بقیه نمی‌تواند در نظر گرفته شود.

نامربوط: ویژگی‌های نامربوط به عنوان ویژگی‌هایی تعریف می شوند که بر خروجی تأثیری ندارند، و مقادیری که برای هر مثال تولید می شوند، تصادفی هستند.

زائد: افزونگی وجود دارد، هر زمان که یک ویژگی بتواند نقش دیگری داشته باشد(شاید ساده ترین راه برای مدل افزونگی).

انتخاب ویژگی نقش مهمی را در تعدادی از وظایف یادگیری ماشین و تشخیص الگو بازی می کند[۱]. بسیاری از ویژگی‌های کاندید معمولاً با یک الگوریتم یادگیری برای تولید خصوصیات کامل عمل کلاس بندی تهیه می شوند. با این حال، در اغلب موارد بسیاری از ویژگی‌های کاندید برای کار یادگیری، نامربوط یا زائد هستند، و کارایی به کارگیری الگوریتم یادگیری را خرابتر خواهند کرد و به مشکل برازش[۱] منجر می‌شوند. دقت یادگیری و سرعت آموزش ممکن است به میزان درخور توجهی با این ویژگی‌های زائد بدتر شود [۲-۴]. بنابراین، انتخاب ویژگی‌های مرتبط و ضروری در مرحله پیش پردازش از اهمیتی بنیادین برخوردار است.

بعضی از روش‌ها برای انتخاب ویژگی در دهه گذشته

---

[۱] تاریخ ارسال مقاله : ۱۳۹۱/۰۹/۱۶
تاریخ پذیرش مقاله : ۱۳۹۲/۰۵/۲۰
نام نویسنده مسؤول : حسن نصرتی ناهوک
نشانی نویسنده مسؤول : ایران – سیستان و بلوچستان - سراوان – بلوار معلم - دانشگاه آزاد اسلامی واحد سراوان.

توسعه داده شده اند [۵]. موضوع اصلی در ساخت الگوریتم‌های انتخاب ویژگی ارزیابی کیفیت ویژگی‌های کاندید است [۶و۷]. مسأله انتخاب ویژگی می تواند به عنوان یک مسأله بهینه سازی چند – هدفه فرموله سازی شده باشد. با وجود وسعت تحقیق در حوزه انتخاب ویژگی، بهترین مجموعه از اهداف یا معیارها برای تعریف راه حل بهینه وجود ندارد. بنابراین، جستجو برای معیارهای کلی به طور مؤثر کانون توجه تحقیقات حاضر است. گذشته از این، ما نیاز داریم روش‌های انتخاب ویژگی را با استفاده از دو معیار مختلف تعریف کنیم: حداکثر رساندن دقت روش و حداقل رساندن تعداد ویژگی‌های که استفاده می شوند [۸ و۹]، و یک فرمول سازی چند – شاخصه از مسأله انتخاب ویژگی ارائه شود.

انتخاب ویژگی، فرایند انتخاب بهترین ویژگی از میان تمام ویژگی‌هاست، زیرا تمام ویژگی‌ها در ساخت خوشه‌ها مفید نیستند: بعضی از ویژگی‌ها ممکن است زائد یا نامربوط باشند بنابراین، برای فرایند یادگیری مؤثر نیستند [۱۰]. انتخاب ویژگی(همچنین به عنوان انتخاب زیرمجموعه شناخته می شوند) فرایندی است که معمولاً در یادگیری ماشین استفاده می شود، در جایی که یک زیر مجموعه از ویژگی‌های در دسترس از داده‌ها برای کاربرد یک الگوریتم یادگیری انتخاب شده است. بهترین زیرمجموعه شامل حداقل تعداد ابعاد است که بیشترین مشارکت را در دقیق سازی دارد. ما ابعاد باقیمانده و بی اهمیت را نادیده می گیریم. هدف اصلی انتخاب ویژگی، تعیین زیرمجموعه ویژگی مینیمال از دامنه مسأله با حفظ دقت بالا بطور، مناسب در ارائه ویژگی‌های اصلی است [۱۱]. این مقاله یک روش جدید انتخاب ویژگی مبتنی برمنطق فازی برای یادگیری ماشین ارائه می کند که از یک رویکرد فازی بر روی روش‌های قبلی استفاده می کند. این روش جدید، روش انتخاب ویژگی مبتنی بر منطق فازی (FSFL[۲]) نامیده می‌شود. این روش ساده است، سریع اجرا می شود و به آسانی برای مسائل کلاس پیوسته با به کارگیری معیارهای تشابه مناسب توسعه داده می شود. این روش، تعداد ویژگی‌های انتخابی را به عنوان ورودی یک عدد فازی در نظر می گیرد و قابلیت ویژگی‌ها را پس از فازی زدایی با استفاده از الگوریتم ژنتیک بهینه می کند و تعداد ویژگی‌های مورد نظر را انتخاب می نماید. بخش بعدی روش‌های انتخاب ویژگی را توصیف می کند. بخش ۳، منطق فازی را توصیف می کند. بخش ۴، روش پیشنهادی را شرح می‌دهد. بخش ۵، الگوریتم ژنتیک ونحوه استفاده از آن را در این روش توضیح می دهد. بخش ۶، آزمایش‌ها و نتایج حاصل از روش پیشنهادی را بیان می کند و بخش آخر، خلاصه و نتیجه گیری را بیان می کند.

۲- روش‌های انتخاب ویژگی

۲-۱- طبقه بندی روش‌های انتخاب ویژگی

به منظور ارزیابی ویژگی‌های انتخاب شده، خصوصیاتی از داده‌ها، مفهوم هدف و الگوریتم یادگیری باید در نظر گرفته شود. براساس این اطلاعات، روش‌های انتخاب ویژگی به سه نوع دسته بندی می شوند: روش‌های فیلتر[۳]، روش‌های پنهان[۴] و روش‌های جاسازی شده[۵]. برای بررسی خوب روش‌های موجود برای انتخاب ویژگی، خوانندگان می توانند به [۱۲و۱۳] مراجعه کنند. روش فیلتر، سادهترین و رایجترین روش مورد استفاده در نوشته‌هاست. این روش شامل الگوریتم‌های رتبه بندی ویژگی [۱۴] و الگوریتم‌های جستجوی زیرمجموعه [۱۵] می باشد. برای روش‌های فیلتر، ویژگی‌ها با توجه به دلایل قدرت پیش بینی را نشان می دهند و سپس رتبه بندی می‌کنند و دارای خصوصیات زیر هستند: ۱. ویژگی‌ها مستقل در نظر گرفته می شوند؛ ۲. ویژگی‌های زائد ممکن است در نظر گرفته شوند؛ ۳. بعضی از ویژگی‌ها به عنوان یک گروه قدرت تبعیض بالایی دارند، اما ضعیف هستند، به همین جهت، به عنوان ویژگی‌های منحصر به فرد نادیده گرفته خواهند شد؛ ۴. رویه فیلتر مستقل از روش کلاس بندی است.

روش‌های پنهان از روش‌های تکراری استفاده می کنند. بسیاری از «زیر مجموعه‌های ویژگی» براساس عملکرد کلاس بندی امتیازدهی می‌شوند و بهترین استفاده را دارند. رویکردهای انتخاب زیرمجموعه شامل انتخاب رو به به جلو، انتخاب رو به عقب، ترکیب آنهاست [۱۶]. این روش دارای خصوصیات زیر است: ۱. از نظر محاسباتی برای هر زیرمجوعه ویژگی در نظر گرفته شده که طبقه بند ساخته شده و ارزیابی شود، گران است. ۲. جستجوی جامع غیر ممکن است، تنها جستجوی حریصانه اعمال می شود. استفاده از جستجوی حریصانه ساده است و به سرعت راه حل‌ها را پیدا می کند، اما عیب آن این است که بهینه نیست و نسبت به شروع‌های نادرست حساس است. ۳. در اغلب موارد در این روش‌ها برای برازش کردن آسان است. و سرانجام در روش‌های جاسازی شده، فرآیند انتخاب ویژگی در درون خود الگوریتم‌های استقرایی انجام می‌شود؛ یعنی تلاش تا به طور مشترک یا همزمان هر دوی طبقه بند و زیرمجموعه ویژگی آموزش داده شوند. آنها معمولاً یک تابع هدف را بهینه سازی می‌کنند که به طور مشترک دقت کلاس بندی را امتیاز می دهد و استفاده از ویژگی‌های بیشتر را جریمه می کند. به هر حال، روش‌های فیلتر و پنهان یک سطح انتزاعی درباره روش جاسازی شده تعیین می‌کنند، فرآیند انتخاب ویژگی برای مدل نهایی مجزا از انتخاب جاسازی شده با خود الگوریتم‌های یادگیری انجام می شود [۱۷و۱۸].

۲-۲- همبستگی مبتنی بر انتخاب ویژگی[۶]

در مرکز الگوریتم CFS، هیوریستیکی برای ارزیابی ارزش یا شایستگی یک زیرمجموعه ویژگی وجود دارد. این هیوریستیک سودمندی ویژگی‌های منحصر به فرد را برای پیش بینی برچسب کلاس همراه با سطحی از همبستگی متقابل در میان آنها به حساب می آورد. این فرضیه که در آن هیوریستیکی براساس: زیرمجموعه ویژگی‌های خوب دارای ویژگی‌هایی با همبستگی بالا با کلاس هستند، در عین حال با یکدیگر ناهمبسته اند.

در آزمون تئوری [۱۸]، همین اصل است که برای طراحی یک آزمون مرکب(مجموع یا متوسط آزمون‌های منحصر به فرد) برای پیش بینی متغیرهای خارجی مورد نظر استفاده می شود. در این وضعیت، ویژگی‌ها، آزمون‌های منحصر به فردی هستند که صفات مربوط به متغیر مورد نظر(کلاس) را اندازه می گیرند. معادله (۱) [۱۹] هیوریستیک را فرمول بندی می کند:

$$Merits = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \qquad (۱)$$

که Merits هیوریستیک « شایستگی» یک زیرمجموعه ویژگی S شامل k ویژگی، $\overline{r_{cf}}$ میانگین همبستگی ویژگی – کلاس، و $\overline{r_{ff}}$ میانگین همبستگی متقابل ویژگی – ویژگی است. معادله (۱)، در واقع، همبستگی Pearson's است، جایی که در آن تمام متغیرها استاندارد شده اند. به صورت کسر می توان به عنوان یک نشانه داده شده فکر کرد که چگونه یک گروه از ویژگی‌ها را پیش بینی می کند و از مخرج کسر این را که چه مقدار افزونگی در میان آنها وجود دارد. هیوریستیک ویژگی‌های نامربوط را که به عنوان پیش‌بینی کننده‌های ضعیف از کلاس خواهند بود، کنترل می کند.

۲-۳- جستجوی فضای زیرمجموعه ویژگی‌ها

استراتژی‌های جستجوی اکتشافی مختلفی مانند تپه نوردی و اول بهترین [۱۹] در اغلب موارد برای جستجوی فضای زیرمجموعه ویژگی‌ها در مدت زمان قابل قبول به کار برده می شوند. ابتدا ماتریس همبستگی ویژگی – کلاس و ویژگی – ویژگی از مجموعه داده‌های آموزشی محاسبه شده و سپس فضای زیرمجموعه ویژگی با استفاده از جستجوی اول – بهترین جستجو می شود. در این مقاله از الگوریتم ژنتیک برای جستجوی فضای زیرمجموعه ویژگی استفاده

می شود.

۳– منطق فازی

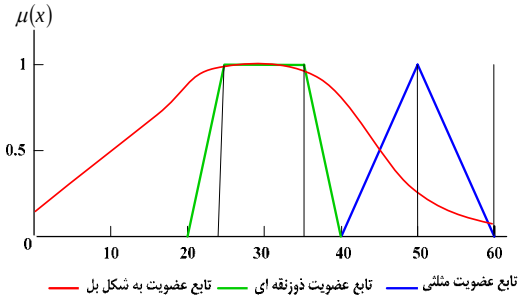
۳–۱– مجموعه‌های فازی

یک مجموعه فازی، مجموعه ای است که اجازه می‌دهد اعضای آن، درجه عضویت متفاوتی در بازه [۰ ,۱] داشته باشند. در منطق کلاسیک، عضویت یک عضو از یک مجموعه با صفر (۰) نمایش داده می شود اگر به مجموعه تعلق نداشته باشد؛ با یک (۱) نشان داده می شود اگر به مجموعه تعلق داشته باشد. یعنی به صورت مجموعه {۰،۱} نشان داده می شود، ولی در منطق فازی این مجموعه به صورت بازه [۰ ,۱] توسعه داده شده است [۲۰و ۲۱]. یک مجموعه فازی توسعه ای از مجموعه کلاسیک است. اگر X، جهان مورد بحث باشد و اعضای آن با x نشان داده شوند، آنگاه مجموعه فازی A از X با زوج مرتب با رابطه (۲) تعریف می شود: $\mu_A(x)$ تابع عضویت، x در A است.

$$A = \{(x, \mu_A) \mid x \in X\} \qquad (۲)$$

اعداد فازی روشی برای توصیف عدم دقت و ابهام داده هستند. یک عدد فازی در مفهوم توسعه ای از یک عدد منظم است که به یک مقدار منفرد اشاره نمی‌کند، اما تا حدودی با مجموعه مقادیر ممکن ارتباط برقرار می کند. این مقدار برای خودش یک وزن بین '0' و '1' دارد. این وزن تابع عضویت[۷] نامیده می شود [۲۲]. یک عدد فازی می‌تواند یکی از سه نوع زیر باشد:

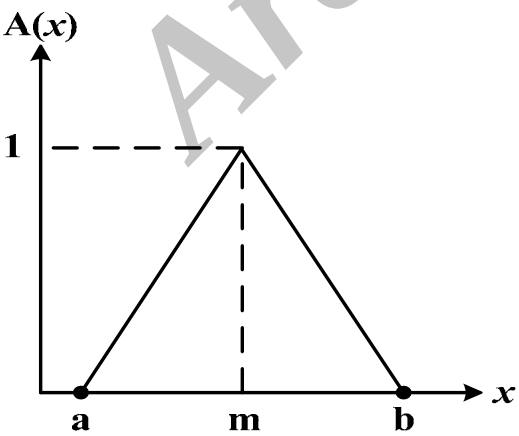
۱) عدد فازی مثلثی؛ ۲) عدد فازی ذوزنقه ای؛ ۳) عدد فازی به شکل بِل، که در شکل (۱)، نشان داده شده اند.



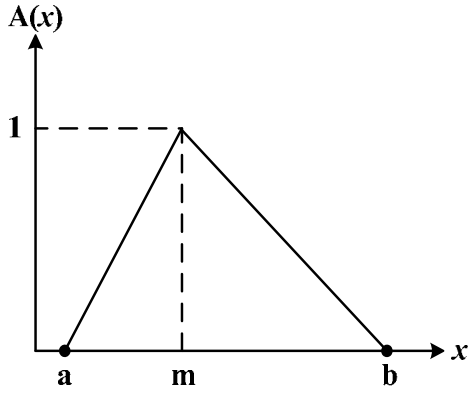
شکل (۱): توابع عضویت.

روش‌های مختلفی برای شکل MFها وجود دارد. در این مقاله، فقط عدد فازی مثلثی را برای روش پیشنهادی در انتخاب ویژگی بیان می کنیم. عدد فازی مثلثی متقارن و نامتقارن با تابع عضویت زیر در شکل‌های (۲) و (۳) نشان داده شده است. با استفاده از کران پایین a و کران بالا b و مقدار میانی m تعریف می شود، که $a < m < b$ است. مقدار $b - m$ را حاشیه می نامند وقتی که با مقدار $m - a$ مساوی باشد. عدد فازی مثلثی در معادله ی (۳) آمده است.

$$A(x) = \begin{cases} 0 & x \le a \quad or \quad x \ge b \\ \dfrac{(x-a)}{(m-a)} & x \in (a,m) \\ \dfrac{(b-x)}{(b-m)} & x \in (m,b) \end{cases} \qquad (۳)$$



شکل (۲): عدد فازی مثلثی متقارن.



شکل (۳): عدد فازی مثلثی نامتقارن.

۳-۳- فازی کردن[۸]

فازی کردن تابع عضویت مثلثی(عدد فازی مثلثی[۹]) A(x)=TFN(α, m, β) با استفاده از معادله ی (۴) تعریف می شود [۲۳]:

$$TFN(F) = \frac{\beta - \alpha}{2m}, 0 < F < 1 \qquad (۴)$$

۳-۴- اصل توسعه[۱۰]

اصل توسعه یک مفهوم اساسی تئوری مجموعه‌های فازی است که یک رویه کلی برای گسترش دامنه‌های قطعی عبارات ریاضی به دامنه‌های فازی فراهم می‌کند. این رویه نگاشت نقطه به نقطه متداول تابع f(.) را به نگاشت بین مجموعه‌های فازی تعمیم می‌دهد. به طور خاص، فرض کنید f یک تابع از X به Y و A یک مجموعه فازی بر روی X با معادله (۵) تعریف شده باشد:

$$A = \frac{\mu_A(x1)}{x1} + \frac{\mu_A(x2)}{x2} + \ldots \frac{\mu_A(xn)}{xn} \qquad (۵)$$

سپس اصل توسعه بیان می کند که تصویر مجموعه فازی A تحت نگاشت f(.) می تواند به عنوان مجموعه فازی B با معادله (۶) بیان شود:

$$B = f(A) = \frac{\mu_A(x1)}{y1} + \frac{\mu_A(x2)}{y2} + \ldots \frac{\mu_A(xn)}{yn} \qquad (۶)$$

به طور کلی، اصل توسعه در معادله (۷) بیان شده است:

$$\mu_B(y) = \max_{x=f^{-1}(y)} \mu_A(x) \qquad (۷)$$

در حالت کلی، اصل توسعه را از یک فضای n بعدی به یک فضای یک بعدی به صورت زیر تعریف می کنیم: فرض کنید که تابع f یک نگاشت از فضای n بعدی ضرب دکارتزین $X_1 \times X_2 \times \ldots \times X_n$ به جهان یک بعدی Y به طوری که $y=f(x_1, x_2, \ldots, x_n)$ باشد و فرض کنید $A_1, A_2, \ldots, A_n$ به ترتیب n مجموعه فازی برروی $X_1, X_2, \ldots, X_n$ هستند. آنگاه اصل توسعه اثبات می کند که مجموعه فازی B استنتاج شده توسط نگاشت f با استفاده معادله (۸) تعریف می شود[۲۴].

$$\mu_B(y) = \begin{cases} \max\left[\min_i \mu_{A_i}(x_i)\right] & f^{-1}(y) \neq 0 \\ (x1,\ldots xn),(x1,\ldots xn)=f^{-1}(y) & \\ 0 \quad f^{-1}(y) = 0 & \end{cases} \qquad (۸)$$

۴ – روش پیشنهادی

در این روش از اصل توسعه که در بخش قبلی بیان شد، در معادله (۱) استفاده می کنیم. در این حالت T یک عدد فازی است که با تابع عضویت T(k) توصیف می شود که درجه عضویت k را با عدد فازی مثلثی T به صورت زیر بیان می کند:

به عنوان ورودی متغیر (تعداد ویژگی‌ها =k)یک عدد فازی(مجموعه فازی) است. تابع f که در اصل توسعه توضیح داده شد، در اینجا همان معادله (۱) می‌باشد. بنابراین، تعیین تابع عضویت شایستگی مبتنی بر اصل توسعه است. اصل توسعه در این روش با معادله (۹) تعریف شده است.

$$\mu(M) = \max[T(k)]$$
$$k \epsilon z^+ : M = \frac{krcf}{\sqrt{k + k(k-1)\overline{rff}}} \qquad (۹)$$

وقتی که T(k) و **μ**(M) توابع عضویت مربوط به تعداد ویژگی‌ها را مشخص می کنند. اصل توسعه تحت معادله (۱) عدد فازی مربوط به k را به عدد فازی مربوط به M با

استفاده از معادله (۱۰) نگاشت می کند.

$$\mu(M) = \max\left[T(k)=T\left(\frac{M^2\left(1-\overline{r_{ff}}\right)}{\overline{r_{cf}}^2 - M^2\overline{r_{ff}}}\right)\right] \qquad (۱۰)$$
$$k\in Z^+; M=\frac{k\overline{r_{cf}}}{\sqrt{k+k(k-1)\overline{r_{ff}}}}$$

معادله (۱۰) می تواند به طور مستقیم برای محاسبه مقدار شایستگی از تعداد ویژگی‌ها استفاده شود. در ادامه نحوه استفاده از معادله (۱۰) در این روش شرح داده می‌شود.

(۱) فازی سازی

در این روش عدد فازی مثلثی T(k) با استفاده از معادله (۱۱) تعریف می شود:

$$T(k)=\begin{cases}0 & k\le\alpha \quad or \quad k\ge\beta\\ \dfrac{(k-\alpha)}{(m-\alpha)} & k\in(\alpha,m)\\ \dfrac{(\beta-k)}{(\beta-m)} & k\in(m,\beta)\end{cases} \qquad (۱۱)$$

مثلث شکل(۳) را به نسبت $P:1$ تقسیم می کنیم مطابق با معادله (۱۲): ($P\in R^+$)

$$\frac{P}{1}=\frac{m-\alpha}{\beta-m}\Rightarrow m=\frac{\alpha+P\times\beta}{P+1} \qquad (۱۲)$$

از (۴) و (۱۲) معادلات (۱۳) و (۱۴) نتیجه می شوند:

$$\alpha=\left(1-\frac{2PF}{P+1}\right)m \qquad (۱۳)$$

$$\beta=\left(1+\frac{2F}{P+1}\right)m \qquad (۱۴)$$

عدد فازی مثلثی M، با استفاده از معادله (۱۱) و اصل توسعه (معادله (۱۰)) با معادله (۱۵) تعریف می‌شود:

$$\mu(M)=\begin{cases}0 \;;M\le M1 \quad or \quad M\ge M2\\ \dfrac{M^2\left(1-\overline{r_{ff}}\right)}{\overline{r_{cf}}^2 - M^2\overline{r_{ff}}}-\alpha \;;M1\le M\le M3\\ \beta-\dfrac{M^2\left(1-\overline{r_{ff}}\right)}{\overline{r_{cf}}^2 - M^2\overline{r_{ff}}} \;;M3\le M\le M2\end{cases} \qquad (۱۵)$$

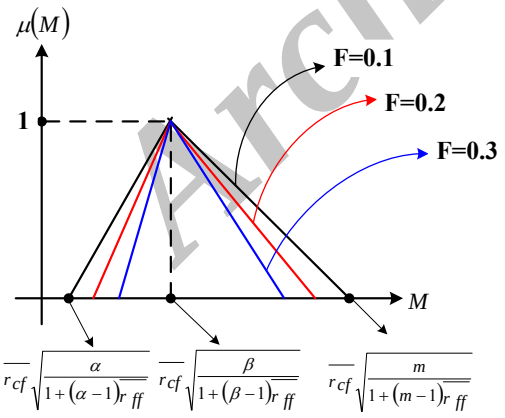
در معادله (۱۵)، مقدار متغیرهای $M1$، $M2$ و $M3$ با استفاده از معادلات (۱۶)، (۱۷) و (۱۸) محاسبه می شوند.

$$M1=\overline{r_{cf}}\sqrt{\frac{\alpha}{1+(\alpha-1)\overline{r_{ff}}}} \qquad (۱۶)$$

$$M2=\overline{r_{cf}}\sqrt{\frac{\beta}{1+(\beta-1)\overline{r_{ff}}}} \qquad (۱۷)$$

$$M3=\overline{r_{cf}}\sqrt{\frac{m}{1+(m-1)\overline{r_{ff}}}} \qquad (۱۸)$$

شکل (۴)، مقادیر α و β را برای مقادیر مختلف F با استفاده از معادله‌های (۱۳) و (۱۴) نشان می دهد.



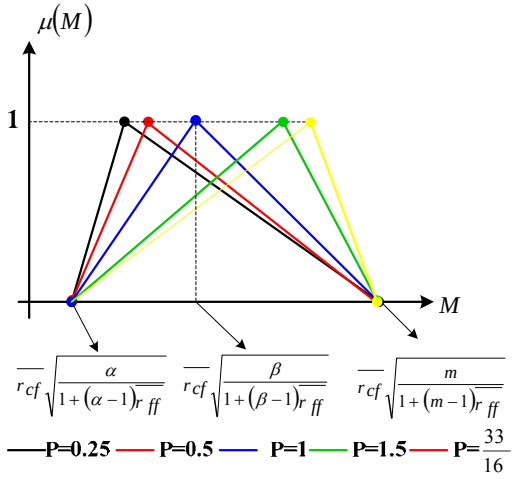
شکل (۴): نمایشی از μ(M) با استفاده از F.

جدول (۱)، مقادیری از α و β را برای مقادیر مختلف F=0.1, 0.2, 0.3 و P نشان می‌دهد، جایی که m تخمینی از تعداد ویژگی‌های انتخابی روش پیشنهادی است.

جدول (۱): مقادیر α و β برای مقادیر مختلف P و F.

F \ P	0.5	1
0.1	$\alpha=\frac{14}{15}m$ $\beta=\frac{17}{15}m$	$\alpha=\frac{9}{10}m$ $\beta=\frac{11}{10}m$
0.2	$\alpha=\frac{13}{15}m$ $\beta=\frac{19}{15}m$	$\alpha=\frac{8}{10}m$ $\beta=\frac{12}{10}m$
0.3	$\alpha=\frac{12}{15}m$ $\beta=\frac{21}{15}m$	$\alpha=\frac{7}{10}m$ $\beta=\frac{13}{10}m$

یا می توان m (کنترل تعداد ویژگی‌ها (k)) را با استفاده از معادله (۱۲) به دست آورد. شکل (۵)، مقادیر مختلفی از m را با استفاده از معادله (۱۲) نشان می دهد.



شکل (۵): نمایشی از μ(M) با استفاده از P.

فرض کنید یک مجموعه داده آموزشی شامل ۲۰۰ مثال و ۹۹ ویژگی داریم، پس α=۱ و β=۹۹ که با تغییر $P\in R^{+}$ می‌توان m را کنترل کرد. جدول (۲)، مقادیری از m را برای مقادیر مختلف P از مجموعه داده آموزشی بالا نشان می دهد.

جدول (۲): محاسبه مقدار m برای مقادیر مختلف P.

P	m
$\frac{33}{16}$	67
1.5	60
1	50
0.5	33
0.25	20

در این روش، ما برای اینکه هر دفعه تعداد ویژگی‌های متفاوتی را انتخاب کنیم، مرکز عدد فازی مثلثی مربوط به تعداد ویژگی‌ها(k) را تغییر می‌دهیم تا مقدار شایستگی تغییر کند. سپس برای انتخاب ویژگی‌های مورد نظر از الگوریتم ژنتیک استفاده می کنیم. هدف به دست آوردن تعداد ویژگی‌های کمتری نسبت به روش‌های معمولی است. یا می‌توان از دو پارامتر P و F استفاده کرد، m تعداد ویژگی‌های انتخابی طوری انتخاب می شوند که α و β به ترتیب کمترین و بیشترین تعداد ویژگی باشند، ولی ما در این روش فقط از معادله (۱۲) یعنی از پارامتر P استفاده کرده ایم.

(۲) فازی زدایی

اکنون می توان خروجی؛ یعنی M را به عنوان میانگین وزنی با تخمین‌های خوش بینانه ($M_1=\overline{rcf}\sqrt{\frac{\alpha}{1+(\alpha-1)\overline{rff}}}$)، محتمل ترین ($M_2=\overline{rcf}\sqrt{\frac{\beta}{1+(\beta-1)\overline{rff}}}$) و بدبینانه ($M_3=\overline{rcf}\sqrt{\frac{m}{1+(m-1)\overline{rff}}}$) با معادله (۱۹) محاسبه کرد[۲۵].

$$Merits(M)=\frac{w_1(M_1)+w_2(M_3)+w_3(M_2)}{w_1+w_2+w_3} \qquad (۱۹)$$

جایی که w1، w2 و w3 به ترتیب وزن‌های مربوطه هستند. ماکزیمم وزن‌ها باید برای پذیرش بهترین M داده شوند. w1، w2 و w3، P و F ثابت‌های اختیاری هستند که با انتخاب w1=1، w2=4 و w3=1 به معادله (۲۰) می‌رسیم.

$$Merits(M)=\frac{M1+4M3+M2}{6} \qquad (۲۰)$$

۵ – الگوریتم ژنتیک

ما در روش پیشنهادی از الگوریتم ژنتیک برای انتخاب تعداد ویژگی‌های بهینه استفاده کردیم، که نحوه استفاده از آن را به صورت بیان می کنیم:

ابتدا ماتریس همبستگی[11] را با استفاده از معادله (۲۰) به صورت زیر محاسبه می کنیم: برای هر مجموعه داده α و β را به ترتیب حد پایین و بالای تعداد ویژگی‌های موجود در آن در نظر می گیریم که بهترین حالت به صورت α=۱ و ماکزیمم تعداد ویژگی = β است. سپس تعداد ویژگی‌ها (k) را (با استفاده از اصل توسعه) به عنوان یک عدد فازی مثلثی در نظر گرفته، با استفاده از پارامتر P تعداد ویژگی‌های تخمینی(مرکز عدد فازی) را تغییر می‌دهیم و متغیرهای M1، M2 و M3 را محاسبه کرده، در آخر با فازی زدایی مقدار شایستگی (M) را به دست می آوریم(در واقع ما هر دفعه با تغییر دادن مرکز عدد فازی مثلثی مربوط به تعداد ویژگی‌ها، M متفاوتی به دست می آوریم). و در نهایت ماتریس همبستگی با توجه به مقدار شایستگی و تابع پیش فرض ضریب همبستگی خطی Pearson's محاسبه می‌شود.

سپس با استفاده از الگوریتم ژنتیک به صورت زیر تعداد ویژگی‌های بهینه انتخاب می شوند: با به کارگیری m، تعداد ویژگی‌های تخمینی و معادله (۲۰) تابعی برای ارزیابی همبستگی ویژگی‌ها به کار برده‌ایم(البته برای اینکه تابع ارزیابی[12] کمینه شود، باید M محاسبه شده را از ۱۰۰ کم کنیم)؛ یعنی تابع ارزیابی استفاده شده در الگوریتم ژنتیک همان معادله (۲۰) است. با این تابع ارزیابی و جعبه ابزار بهینه سازی ژنتیک تعداد ویژگی‌های انتخابی را از مجموعه داده‌ها به دست می آوریم.

۶ – نتایج آزمایش

در این مقاله از شش مجموعه داده، (D1): Arrhythmia، (D2): Dbworld، (D3): Dbworld_bodies_stemmed، (D4): Isolet، (D5): Madelon و (D6): Multiple Features (mfeat) برای تست روش جدید استفاده شده است. مجموعه داده‌ها از منبع داده‌های یادگیری ماشین UCI گرفته شده اند[۲۶].

خلاصه ای از خصوصیات مجموعه داده‌ها در جدول (۳) آمده اند. ما روش پیشنهادی را بر روی مجموعه داده‌هایی که ذکر شدند، پیاده سازی و چهار نوع مختلف از طبقه بندهای موجود در Weka را بر روی هر یک از این مجموعه داده‌ها برای محاسبه دقت کلاس بندی در این آزمایش استفاده کرده ایم. در این مقاله، روش FSFL با استفاده از الگوریتم ژنتیک برای انتخاب تعداد ویژگی‌ها پیاده سازی شده است و با روش‌های CFS و انتخاب ویژگی مبتنی بر مجموعه‌های سخت فازی[13] مقایسه می‌کنیم[۲۷]. روش معمولی برای هر مجموعه داده آموزشی تعداد ویژگی‌هایی را که انتخاب می کند ثابت هستند، اما روش پیشنهادی می‌تواند برای هر مجموعه داده با تغییر دادن مرکز عدد فازی مثلثی استفاده شده برای تعداد ویژگی‌ها، تعداد ویژگی متفاوتی را انتخاب کند. استفاده از نظریه مجموعه فازی در روش CFS باعث شد تا برای هر مجموعه داده استاندارد بتوانیم هر دفعه تعداد ویژگی‌های متفاوتی انتخاب کنیم. در واقع ما می توانیم با روش جدید پیشنهادی تعداد ویژگی‌های کمتری نسبت به روش CFS، FRFS یا روش‌های انتخاب ویژگی قبلی انجام شده، انتخاب کنیم. تعداد ویژگی‌های انتخابی روش جدید بر روی شش مجموعه داده در مقایسه با روش‌های CFS و FRFS در جدول (۴) آمده اند.

(در این جدول فقط تعداد ویژگی‌های انتخابی کمتر نسبت به روش معمولی آمده‌اند در صورتی که می توان با تغییر پارامترها ویژگی‌های مختلفی را انتخاب کرد، اما تعداد ویژگی‌هایی را که روش جدید انتخاب کرده و دقت کلاس بندی نزدیک یا بیشتر از روش‌های CFS و FRFS داشته باشند، آورده‌ایم).

جدول (۳): خلاصه ای از داده‌های آزمایش

مجموعه داده‌ها	تعداد ویژگی‌ها	تعداد مثال‌ها
Arrhythmia	۲۷۹	۴۵۲
Dbworld	۴۷۰۲	۶۴
Dbworld_bodies_stemmed	۳۷۲۱	۶۴
Isolet	۶۱۷	۷۷۹۷
Madelon	۵۰۰	۱۸۰۰
Multiple Features (mfeat)	۲۱۶	۲۰۰۰

سپس ما دقت کلاس بندی روش جدید را در مقایسه با روش‌های CFS، FRFS و کل مجموعه داده بر روی چهار طبقه بند مختلف برای هر یک از مجموعه داده‌ها محاسبه کرده و در جدول‌های (۵)، (۶)، (۷)، (۸)، (۹) و (۱۰) آورده‌ایم.

جدول (۴): تعداد ویژگی‌های انتخابی روش پیشنهادی

مجموعه داده	تعداد کل ویژگی‌ها	تعداد ویژگی‌های انتخابی روش FSFL و روش‌های CFS و FRFS		
		CFS	FSFL	FRFS
(D1)	۲۷۹	۱۱	۸	۱۳۱
(D2)	۴۷۰۲	۲۱۸	۱۹۲	۱۲۸
(D3)	۳۷۲۱	۱۶۱	۱۳۲	۳۰۱
(D4)	۶۱۷	۴	۳	۸۰
(D5)	۵۰۰	۱۸	۱۷	۳۳
(D6)	۲۱۶	۸	۷	۲۷

جدول (۵): دقت کلاس بندی روش جدید در مقایسه با روش‌های قبلی و کل داده‌ها برای مجموعه داده D1 با چهار طبقه بند متفاوت

مجموعه داده	طبقه بندها	دقت کلاس بندی			
		کل داده‌ها	CFS	FSFL	FRFS
(D1)	M5P	44.69٪	43.52٪	45.40٪	44.66٪
	SMOreg	26.45٪	42.54٪	42.58٪	33.83٪
	Bagging	50.9٪	39.25٪	49.72٪	48.19٪
	M5Rules	44.22٪	43.52٪	43.71٪	44.84٪

جدول (۶): دقت کلاس بندی روش جدید در مقایسه با روش‌های قبلی و کل داده‌ها برای مجموعه داده D2 با چهار طبقه بند متفاوت

مجموعه داده	طبقه بندها	دقت کلاس بندی			
		کل داده‌ها	CFS	FSFL	FRFS
(D1)	NaiveBayes	75٪	98.43٪	98.43٪	84.375٪
	SMO	87.5٪	98.43٪	98.43٪	95.312٪
	AdaBoostM1	82.81٪	98.43٪	98.43٪	82.812٪
	LMT	81.25٪	98.43٪	98.43٪	82.812٪

جدول (۷): دقت کلاس بندی روش جدید در مقایسه با روش‌های قبلی و کل داده‌ها برای مجموعه داده D3 با چهار طبقه بند متفاوت

مجموعه داده	طبقه بندها	دقت کلاس بندی			
		کل داده‌ها	CFS	FSFL	FRFS
(D3)	NaiveBayes	76.56٪	96.75٪	96.90٪	90.625٪
	SMO	89.06٪	97.70٪	97.70٪	98.437٪
	AdaBoostM1	79.68٪	92.57٪	96.87٪	92.187٪
	LMT	79.68٪	98.43٪	98.43٪	84.375٪

جدول (۸): دقت کلاس بندی روش جدید در مقایسه با روش‌های قبلی و کل داده‌ها برای مجموعه داده D4 با چهار طبقه بند متفاوت

مجموعه داده	طبقه بندها	دقت کلاس بندی			
		کل داده‌ها	CFS	FSFL	FRFS
(D4)	M5P	75.83٪	33.30٪	33.10٪	34.40٪
	SMOreg	72.51٪	25.66٪	24.82٪	27.76٪
	Bagging	80.42٪	31.38٪	31.38٪	32.19٪
	M5Rules	71.94٪	32.30٪	31.97٪	34.15٪

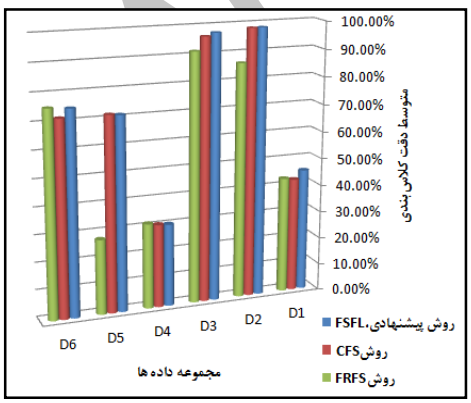
جدول (۹): دقت کلاس بندی روش جدید در مقایسه با روش‌های قبلی و کل داده‌ها برای مجموعه داده D5 با چهار طبقه بند متفاوت

مجموعه داده	طبقه بندها	دقت کلاس بندی			
		کل داده‌ها	CFS	FSFL	FRFS
(D5)	M5P	29.19٪	75.32٪	75.16٪	11.32٪
	SMOreg	42.77٪	61.32٪	60.67٪	11.71٪
	Bagging	76.82٪	75.47٪	75.62٪	75.99٪
	M5Rules	29.19٪	74.38٪	73.56٪	11.32٪

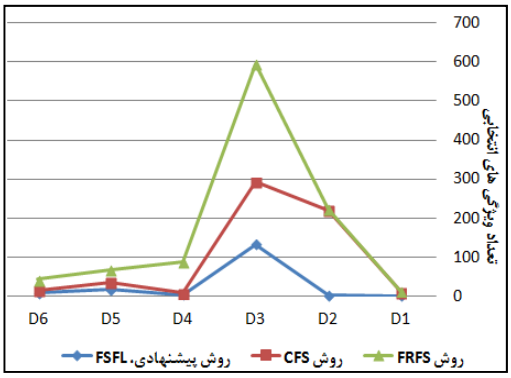
جدول (۱۰): دقت کلاس بندی روش جدید در مقایسه با روش‌های قبلی و کل داده‌ها برای مجموعه داده D6 با چهار طبقه بند متفاوت

مجموعه داده	طبقه بندها	دقت کلاس بندی			
		کل داده‌ها	CFS	FSFL	FRFS
(D6)	M5P	99.84٪	75.32٪	78.95٪	78.99٪
	SMOreg	99.55٪	61.32٪	61.32٪	62.58٪
	Bagging	99.98٪	75.47٪	84.62٪	84.89٪
	M5Rules	99.83٪	74.38٪	74.13٪	74.91٪

همان طور که در جداول (۵) تا (۱۰) مشاهده می شود، عملکرد روش پیشنهادی (FSFL) نسبت به روش معمولی آن (CFS) و FRFS بهتر است. در روش جدید ما توانستیم با استفاده از مقادیر مختلف متغیر P و انتخاب بهینه ترین مقدار از بین آنها، تعداد ویژگی‌های کمتر با متوسط دقت کلاس بندی بیشتر یا نزدیک به آن نسبت به روش‌های CFS و FRFS انتخاب کنیم؛ همان طور که در نمودار شکل‌های (۶) و (۷) آمده است.



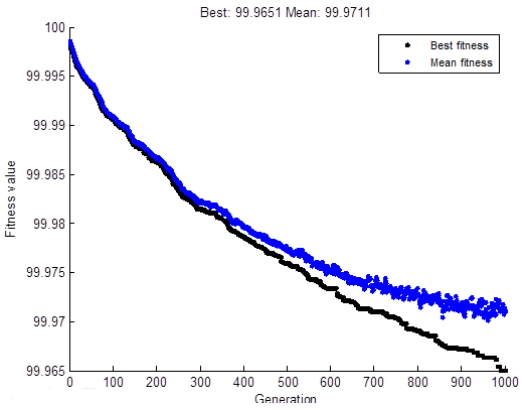
شکل (۶): نمودار مقایسه ای روش FSFL با روش‌های دیگر.



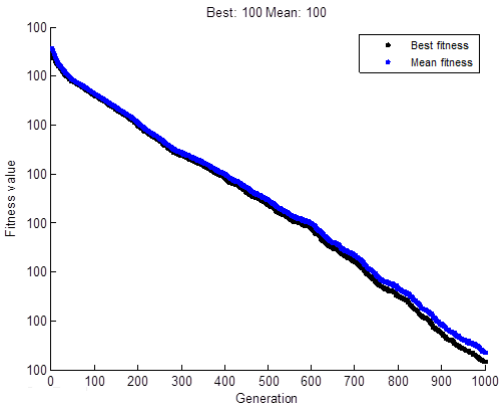
شکل (۷): تعداد ویژگی‌های انتخابی روش FSFL در مقایسه با روش‌های دیگر.

اگر چه روش‌های قبلی در بعضی از مجموعه داده‌ها متوسط دقت کلاس بندی بیشتری دارند، ولی این روش‌ها تعداد ویژگی‌های بیشتری انتخاب می کنند که در یادگیری ماشین هدف حداقل کردن تعداد ویژگی‌ها با متوسط دقت کلاس بندی بالاست. مقایسه نتایج آزمایش‌ها برتری روش پیشنهادی را به طور قابل محسوسی نشان می‌دهد.

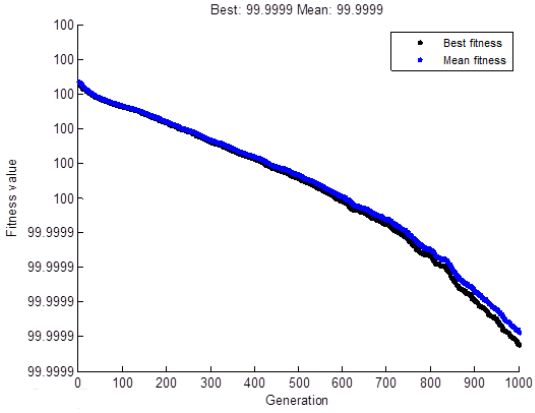
شکل‌های (۸)، (۹)، (۱۰)، (۱۱)، (۱۲) و (۱۳)، نمودارهای بهترین سازگاری[۱۴] (همگرایی) حاصل از الگوریتم ژنتیک در روش پیشنهادی را برای مجموعه داده‌های آمده در جدول (۳) نشان می‌دهند.



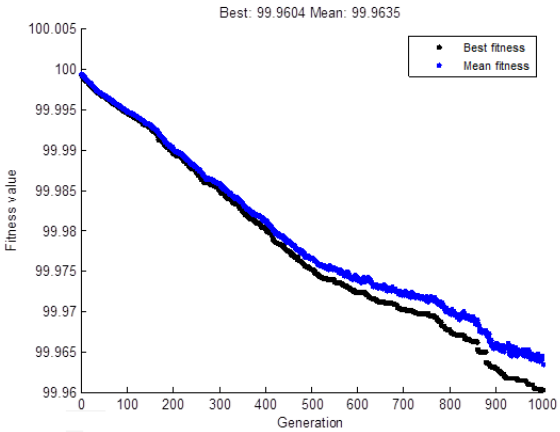
شکل (۸): نمودار همگرایی‌الگوریتم ژنتیک مجموعه داده اول



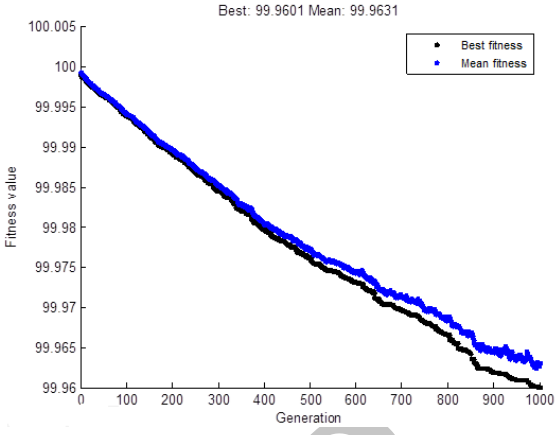
شکل (۹): نمودار همگرایی‌الگوریتم ژنتیک مجموعه داده دوم



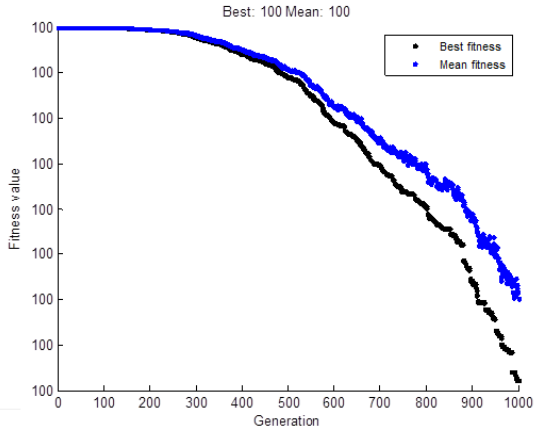
شکل (۱۰): نمودار همگرایی‌الگوریتم ژنتیک مجموعه داده سوم



شکل (۱۱): نمودار همگرایی الگوریتم ژنتیک مجموعه داده چهارم



شکل (۱۲): نمودار همگرایی الگوریتم ژنتیک مجموعه داده پنجم



شکل (۱۳): نمودار همگرایی الگوریتم ژنتیک مجموعه داده ششم

۷ – نتیجه گیری

در این مقاله، روشی برای بهبود روش همبستگی مبتنی بر انتخاب ویژگی(CFS) در یادگیری ماشین ارائه شد. متغیر تعداد ویژگی فازی سازی شد و سپس با استفاده از اصل توسعه معیار CFS فازی سازی و مقدار آن محاسبه شد. همان طور که در آزمایش‌های نشان داده شد، در این روش تعداد ویژگی‌های کمتری با دقت طبقه بندی نزدیک یا بیشتر نسبت به روش قبلی برای مجموعه داده‌های مختلف بر روی طبقه بندهای متفاوت به دست آمد. همچنین، در روش پیشنهادی قادر خواهیم بود تعداد ویژگی‌های انتخابی را با تنظیم پارامترهای عدد فازی مثلثی تعریف شده روی متغیر تعداد ویژگی کنترل کنیم.

مراجع

[1] H. Liu et al., Boosting feature selection using information metric for classification, Neurocomputing, Vol. 73, No. 1, pp. 295 – 303, 2009.

[2] L. Yu, H. Liu., Efficient feature selection via analysis of relevance and redundancy, Journal of Machine Learning Research, Vol. 5, No. 2, pp. 1205 – 1224, 2004.

[3] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, Journal of Machine Learning Research, Vol. 3, No. 1, pp. 1157 – 1182, 2003.

[4] H. Liu, L. Yu, toward integrating feature selection algorithms for classification and clustering, IEEE Transactions on Knowledge and Data Engineering, Vol. 17, No. 4, pp. 491 – 502, 2005.

[5] J. Neumann, C. Schnorr, G. Steidl, Combined SVM-based feature selection and classification, Machine Learning, Vol. 61, No. 3, pp. 129 – 150, 2005.

[6] Q.H. Hu, D. Yu, J.F. Liu, C. Wu, Neighborhood rough set based heterogeneous feature subset selection, Information Sciences, Vol. 178, No. 4, pp. 3577 – 3594, 2008.

[7] M.A. Hall, Correlation-based feature selection for discrete and numeric class machine learning, in: Proceedings of 17 th International Conference on Machine Learning, pp. 359 – 368, 2000.

[8] S.M. Vieira, J.M.C. Sousa, U. Kaymak, Feature selection using fuzzy objective functions, in: Proceedings of the IFSA/EUSFLAT International Fuzzy Systems Association World Congress and 6th EUSFLAT Conference, pp. 1673–1678, 2009.

[9] Susana M.Vieira, JoãoM.C.Sousa, UzayKaymak, Fuzzy criteria for feature selection, Fuzzy Sets and Systems, Vol. 189, No. 1, pp. 1–18, 2012.

[10] Shailendra Singh., Sanjay Silakari., An ensemble approach for feature selection of Cyber Attack Dataset., (IJCSIS) International Journal of Computer Science and Information Security, Vol. 6, No. 2, pp. 297–302, 2009.

[11] L.Ladha, Research Scholar, T.Deepa, Lecturer, Feature Selection Methods and

Algorithms, International Journal on Computer Science and Engineering (IJCSE), Vol. 3, No. 5, pp. 1787–1797, 2011.

[12] Guyon, I.Elisseeff, A. An introduction to variable and feature selection. Journal of Machine Learning Research, Vol. 3, pp. 1157–1182, 2003.

[13] J. Huang, Y. Cai, X. Xu, A hybrid genetic algorithm for feature selection wrapper based on mutual information, Pattern Recognition Letters, Vol. 28, pp. 1825–1844, 2007.

[14] Silvia Casado Yusta, Different metaheuristic strategies to solve the feature selection problem, Pattern Recognition Letters, Vol. 30, No. 5, pp. 525–534, 2009.

[15] Marc Sebban, Richard Nock, A hybrid /wrapper approach of feature selection using information theory, Pattern Recognition, Vol. 35, pp. 835–846, 2002.

[16] Hongwen Zheng, Yanxia Zhang, Feature selection for high-dimensional data in astronomy, Advances in Space Research, Vol. 41, No. 2, pp. 1960–1964, 2008.

[17] Yvan saey, in aki inza and pedro larran aga, "A review of feature selection techniques in bioinformatics", Vol. 23, No. 4, pp. 2507–2517, 2007.

[18] Patricia E.N. Lutu, Andries P. Engelbrecht, A decision rule-based method for feature selection in predictive data mining, Expert Systems with Applications, Vol. 37, pp. 602 – 609, 2010.

[19] D. Huang, Zhaohui Gan, Tommy W.S. Chow, Enhanced feature selection models using gradient-based and point injection techniques, Neurocomputing, Vol. 71, No. 5, pp. 3114–3123, 2008.

[20] Crespo, J., Sicilia, M.A, Garcia, E., Cuadrado J.J, "On Aggregating Second-Level Software Estimation Cost Drivers: A Usability Cost Estimation Case Study", Information Processing and Management Of Uncertainty in Knowledge-Based Systems IPMU, pp. 1255–1260, 2004.

[21] A.Mittal, K.Parkash, H.Mittal, Software Cost Estimation Using Fuzzy Logic, ACM SIGSOFT Software Engineering, Vol. 35, No. 1, pp. 1–7, 2010.

[22] Harish Mittal, Pardeep Bhatia," Software Maintainability Assessment based on fuzzy logic Technique "ACM SIGSOFT, Vol. 34, No. 2, pp. 1–5, 2009.

[23] A.BalaKrishna, T.K.Rama Krishna, Fuzzy and Swarm Intelligence for Estimation,Advances in Information Technology and Management (AITM), Vol. 2, No.1, pp. 246 – 250, 2012.

[24] Jyh – shing Roger Jang, Chuen – Tasi Sun, Eiji Mizutani, "Neuro – Fuzzy and Soft Computing" A Computational Approach to Learning and Machine Intelligence, Prentice Hall Upper Saddle River, 07458, pp. 47–50, 1997.

[25] Shi- Jay Chen and ShykMing Chen, A New Method to Measure the Similarity between Fuzzy Numbers, IEEE International Fuzzy Systems Conference, Vol. 3, pp. 1123–1126, 2001.

[26] Newman, D. J, Hettich, S., Blake, C. L., & Merz, C. J. UCI Repository of machine learning databases. Irvine, CA: University of California, Department of Information and ComputerScience,www.archive.ics.uci.edu/ml, 2007.

[27] Richard Jensen, Qiang Shen, New Approaches to Fuzzy-Rough Feature Selection, Ieee Transactions on Fuzzy Systems, Vol. 17, No. 4, pp. 824–838, 2009.

زیرنویس

[1] Overfitting
[2] Feature Selection based on Fuzzy Logic
[3] Filter
[4] Wrapper
[5] Embedded
[6] Correlation based on Feature Selection(CFS)
[7] Membership Function(MF)
[8] Fuzziness
[9] Triangular Fuzzy Number
[10] Extension Principle
[11] Correlation Matrix
[12] Fitness Function
[13] Fuzzy – Rough – Set based on Feature Selection(FRFS)
[14] Best Fitness