سیستم‌های فازی و کاربردها
شاپای چاپی: ۲۷۱۷-۴۴۰۹
سال پنجم، شماره اول (بهار و تابستان ۱۴۰۱)
شاپای الکترونیکی: ۲۷۱۷-۳۹۸۴
صص. ۱۶۱ تا ۱۸۵
DOI:10.22034/JFSA.2022.331660.1116

انتخاب ویژگی برخط بر اساس انتگرال فازی چوکت

امین هاشمی، محمدرضا پژوهان* و محمدباقر دولتشاهی

دانشکده مهندسی کامپیوتر،پردیس فنی و مهندسی،دانشگاه یزد
دانشکده مهندسی کامپیوتر،پردیس فنی و مهندسی،دانشگاه یزد
گروه کامپیوتر،دانشکده فنی و مهندسی،دانشگاه لرستان

تاریخ دریافت: ۱۴۰۰/۱۲/۸ تاریخ پذیرش: ۱۴۰۱/۵/۱۲
نوع مقاله: علمی-پژوهشی

چکیده

انتخاب ویژگی یک فرایند پیش‌پردازش داده‌ها است که برای مجموعه‌داده‌های با ابعاد بالا قبل از اجرای الگوریتم‌های یادگیری ماشین و داده‌کاوی مورداستفاده قرار می‌گیرد. هدف از انتخاب ویژگی، پیداکردن یک زیرمجموعه‌ی حداقلی و بهینه از مجموعه ویژگی‌ها است. این زیرمجموعه، ویژگی‌های برجسته را شامل می‌شود در حالی که ویژگی‌های غیرمرتبط با برچسب کلاس و تکراری در آن قرار نمی‌گیرند. برای انجام این کار، بسیاری از روش‌های انتخاب ویژگی فعلی به کل ویژگی‌ها در ابتدا نیاز دارند و درصورتی‌که ویژگی جدیدی در آینده به مجموعه ویژگی‌ها اضافه شود، الگوریتم باید از ابتدا اجرا شود. به‌دست‌آوردن کل ویژگی‌ها و یا حتی منتظر ماندن برای آن غیرممکن در بسیاری از کاربردهای واقعی ممکن نیست؛ بنابراین برای این‌گونه مسائل که کل فضای ویژگی در ابتدا در اختیار ما قرار ندارد، روش‌های انتخاب ویژگی برخط ارائه شده‌اند. در این مقاله یک روش انتخاب ویژگی برخط با استفاده از مفهوم انتگرال فازی چوکت ارائه شده است. این روش در ابتدا جریان‌های ویژگی را بر اساس چندین معیار فیلتر ارزیابی می‌کند. سپس بر اساس عملگر چوکت (ادامه دارد)

عبارات و کلمات کلیدی: انتخاب ویژگی برخط، انتگرال فازی چوکت، جریان ویژگی، داده‌های با ابعاد بالا.

Email(s): hashemi.am@stu.yazd.ac.ir, pajoohan@yazd.ac.ir and dowlatshahi.mb@lu.ac.ir.

Mathematics Subject Classification: 03E72-68W01

نتایج آن‌ها ترکیب و برای حفظ یا نادیده‌گرفتن ویژگی تصمیم‌گیری می‌شود. در گام ارزیابی، عملکرد الگوریتم پیشنهادی با پنج روش انتخاب ویژگی برخط و بر اساس دو دسته‌بند مقایسه شده است. روش پیشنهادی بر اساس نتایج به‌دست‌آمده در هفت مجموعه‌داده دنیای واقعی نزدیک دو درصد بهبود نسبت به روش‌های مشابه بر اساس معیارهای دقت دسته‌بندی و امتیاز F داشته است. همچنین به دلیل محاسبات ساده در فرایند روش پیشنهادی، ارزیابی ویژگی‌ها در زمان کوتاهی انجام می‌گیرد.

۱ مقدمه

عصر جدید را می‌توان عصر داده‌های حجیم[1] نام‌گذاری کرد، زیرا امروزه داده‌ها با حجم و ابعاد بالا[2] با سرعت شگفت‌انگیزی در حال تولید هستند. این‌گونه داده‌های حجیم، چالش‌های بسیاری را برای الگوریتم‌های یادگیری ماشین [3] و داده‌کاوی[4] برای استخراج دانش ایجاد می‌کنند. مطالعات بسیاری با ارائه روش‌های متنوع، سعی در برطرف‌کردن این چالش‌ها دارند. این مسئله که با نام نفرین ابعاد[5] شناخته می‌شود، باعث کاهش دقت در الگوریتم‌های یادگیری، افزایش پیچیدگی محاسباتی و زمان یادگیری و همچنین بیش برازش[6] در الگوریتم‌ها می‌شود. در این موارد، روش‌های کاهش ابعاد[7] عملکرد مؤثری در کاهش چالش‌های یادشده دارند که انتخاب ویژگی[8] یکی از تکنیک‌های مؤثر در این حوزه است. اغلب در این نوع داده‌ها، ویژگی‌های غیرمرتبط[9] با برچسب کلاس و افزونه‌ی[10] بسیاری وجود دارد که حذف آنها و انتخاب زیرمجموعه‌ای کوچک و درعین‌حال مؤثر به برطرف‌کردن مشکلات یاد شده کمک می‌کند که این فرایند، انتخاب ویژگی نامیده می‌شود [۴]،[۷]،[۱۳] .

[1] Big data
[2] High dimensional
[3] Machine learning
[4] Data mining
[5] Curse of dimensionality
[6] Over fitting
[7] Dimensionality reduction
[8] Feature selection
[9] Irrelevant
[10] Redundant

روش‌های انتخاب ویژگی سنتی[11] برای دهه‌ها توسط پژوهشگران موردمطالعه قرار گرفته‌اند و مقالات بسیار زیادی در این حوزه تحقیقاتی تاکنون منتشر شده است [۵]،[۶]،[۸]،[۹] . روش‌های انتخاب ویژگی سنتی با فرض این که تمام ویژگی‌ها در ابتدا در دسترس هستند، ویژگی‌هایی را از فضای ویژگی و بر اساس رویکردهای مشخص انتخاب می‌کنند. درحالی‌که در کاربردهای دنیای واقعی، همواره تمام ویژگی‌ها در اختیار ما قرار ندارد. همچنین در برخی از کاربردها، فضای ویژگی ممکن است به‌صورت نامحدود باشد که در این حالت ویژگی‌ها به‌صورت پویا و در طی زمان به مجموعه‌داده اضافه می‌شوند. به‌عنوان مثال، در مسئله تشخیص موضوعات داغ[12] در شبکه‌های اجتماعی توییتر[13] و فیس‌بوک[14] که کلمات کلیدی به‌عنوان ویژگی‌ها در تشخیص این موضوعات مورداستفاده قرار می‌گیرند، فضای ویژگی به‌صورت نامحدود است. در طول زمان با درنظرگرفتن شرایط مختلف سیاسی، اقتصادی، علمی، پزشکی و غیره مباحث داغ دیگری ایجاد می‌شوند که باعث اضافه شده کلمات کلیدی جدید و در نتیجه ویژگی‌های جدیدی به مجموعه ویژگی‌ها می‌شود. به‌عنوان مثال بیماری کووید ۱۹[15] در سال‌های گذشته وجود نداشته است، اما با همه‌گیری این بیماری، کلمات کلیدی مربوط به آن به‌عنوان ویژگی‌های مؤثر به فضای ویژگی تشخیص موضوعات داغ اضافه شده‌اند. از طرفی با افزایش حجم و ابعاد داده‌ها، در برخی از مجموعه‌داده‌های بزرگ، نمی‌توان کل مجموعه‌داده را قبل از فرایند انتخاب ویژگی به طور کامل در حافظه بارگذاری کرد؛ بنابراین برای مجموعه‌داده‌های بزرگ، بهتر است که داده‌ها به‌صورت جریان[16] سطری (نمونه‌ها) و یا ستونی (ویژگی‌ها) پردازش شوند. روش‌های انتخاب ویژگی برخط[17] برای یافتن زیرمجموعه‌ی بهینه از ویژگی‌ها در مواردی که روش‌های انتخاب ویژگی سنتی قابل‌اجرا نیستند، ارائه شده‌اند. این روش‌ها، ویژگی‌ها را بر اساس جریان داده در پنجره‌های زمانی[18] مختلف ارزیابی می‌کنند و با ورود هر جریان داده جدید، زیرمجموعه‌ی

[11] Traditional
[12] Hot topic detection
[13] Twitter
[14] Facebook
[15] Covid-19
[16] Stream
[17] Online feature selection
[18] Sliding windows

ویژگی‌های انتخاب شده را به‌روزرسانی می‌کنند؛ بنابراین در هر لحظه، زیرمجموعه‌ی بهینه از ویژگی‌هایی که تاکنون در اختیار ما قرار گرفته‌اند، در دسترس است. جریان داده را می‌توان به دو دسته جریان نمونه و جریان ویژگی تقسیم کرد. در روش‌های انتخاب ویژگی برخط با جریان ویژگی فرض بر این است که تعداد نمونه‌های آموزشی ثابت است، درصورتی‌که ویژگی‌ها در هر پنجره زمانی افزایش می‌یابند که مسئله تشخیص موضوعات داغ مثالی از این کاربرد است. روش‌های انتخاب ویژگی برخط مبتنی بر جریان نمونه‌ها آن دسته از روش‌هایی هستند که فرض مسئله ثابت بودن تعداد ویژگی‌ها و ورود جریان نمونه‌های آموزشی در طی زمان است. به‌عنوان مثال در سامانه‌های تشخیص فیشینگ[۱۹] ، ویژگی‌ها ثابت هستند و در طول زمان وب‌سایت‌های عادی و فیشینگ به مجموعه‌داده اضافه می‌شوند [۱۰]،[۱۲]،[۲۰] .

روش‌های انتخاب ویژگی برخط مبتنی بر جریان ویژگی به دو دسته کلی تقسیم می‌شوند: روش‌های جریان واحد[۲۰] و روش‌های جریان گروهی[۲۱] . روش‌های جریان واحد در مسائلی کاربرد دارند که در هر زمان یک ویژگی و در روش‌های جریان گروهی در هر زمان چندین ویژگی به مجموعه ویژگی‌ها اضافه می‌شود. رویکرد اصلی در روش‌های انتخاب ویژگی برخط به این صورت است که تصمیم‌گیری بر اساس یک شاخص صورت می‌گیرد که یک ویژگی حفظ یا نادیده گرفته شود [۲]،[۱۱] .

روش‌هایی که تا به امروز در حوزه انتخاب ویژگی برخط مبتنی بر جریان ویژگی ارائه شده‌اند شامل ضعف‌هایی از جمله دقت پایین دسته‌بندی، پیچیدگی محاسباتی بالا، انتخاب تعداد زیاد ویژگی‌ها و زمان پردازش بالا هستند. ازاین‌رو هدف ما در این مقاله ارائه روشی مؤثر برای برقراری توازن بین دقت دسته‌بندی، زمان اجرا و تعداد ویژگی‌های انتخابی است که مهم‌ترین معیارهای سنجش الگوریتم‌ها در محیط برخط هستند. ما معتقد هستیم که استفاده از یک شورای تصمیم‌گیری می‌تواند عملکرد بهتری نسبت به تنها یک شاخص تصمیم‌گیری داشته باشد. انتخاب ویژگی شورایی یک رویکرد جدید در ارزیابی ویژگی‌ها است که بر اساس ترکیب نتایج چندین روش انتخاب ویژگی یا معیار ارزیابی ویژگی، زیرمجموعه‌ی نهایی ویژگی‌ها انتخاب می‌شود. منطق استفاده از این

[19] Phishing detection systems

[20] Single stream

[21] Group stream

رویکرد بر اساس این واقعیت است که هر روش و معیار انتخاب ویژگی ممکن است زیرمجموعه‌ای از ویژگی‌ها را انتخاب کند که در فضای ویژگی یک بهینه محلی باشد. ترکیب چندین روش و معیار مختلف می‌تواند از انتخاب بهینه محلی جلوگیری کند، چرا که تنوع معیارهای مختلف می‌تواند منجر به انتخاب ویژگی‌های بهتری شود. از طرفی تا به امروز استفاده از رویکرد شورایی در مسأله‌ی انتخاب ویژگی برخط مبتنی بر جریان ویژگی مورد مطالعه قرار نگرفته است.

روش‌های انتگرال فازی عملگرهای ترکیبی مؤثری هستند که می‌توانند نتایج به‌دست‌آمده از چندین معیار را با درجه تأثیر مختلف با هم ترکیب کنند. این عملگرها بر اساس مفهوم شاخص‌های فازی ایجاد شده‌اند که این شاخص‌ها، اطلاعات را از چندین منبع دریافت کرده و به یک مقدار واحد تبدیل می‌کنند. با توجه به این‌که در مسائل انتخاب ویژگی برخط مبتنی بر جریان ویژگی واحد در هر لحظه تنها در مورد یک ویژگی تصمیم‌گیری می‌شود، ما بر این باور هستیم که عملگرهای انتگرال فازی می‌توانند برای تصمیم‌گیری در این مسأله مؤثر باشند، زیرا این عملگرها در ترکیب مقادیر برداری و خطی، مؤثر و سریع هستند؛ بنابراین پیچیدگی محاسباتی زیادی ایجاد نمی‌شود و همچنین تصمیم‌گیری سریع که در تحلیل برخط یکی از موارد مهم است، فراهم می‌شود.

در این مقاله یک روش انتخاب ویژگی مبتنی بر جریان ویژگی واحد و با استفاده از مفهوم انتگرال فازی[۲۲] ارائه شده است. عملگر چوکت[۲۳] ، که یکی از عملگرهای انتگرال فازی است، به عنوان معیار اصلی برای انتخاب ویژگی در فضای برخط درنظر گرفته شده است. در این مقاله برخلاف روش‌های کنونی انتخاب ویژگی برخط مبتنی بر جریان ویژگی که فقط از یک معیار برای ارزیابی ویژگی‌ها استفاده می‌کنند، چند معیار برای انتخاب ویژگی استفاده شده است، زیرا تنوع روش‌های مختلف و بررسی یک ویژگی بر اساس چندین معیار متنوع می‌تواند منجر به نتیجه بهتری نسبت به یک معیار شود. در این الگوریتم فرایند ارزیابی ویژگی‌ها در دو سطح صورت می‌گیرد، با ورود هر ویژگی جدید ابتدا مرتبط‌بودن آن با برچسب کلاس بررسی و در صورتی که از یک حد آستانه بالاتر بود ویژگی حفظ می‌شود. در نهایت با انتخاب هر ویژگی، مجموعه فعلی ویژگی‌های انتخاب شده بر اساس میزان افزونگی و از طریق انتگرال فازی چوکت به‌روزرسانی می‌شود.

---

[22] Fuzzy integral

[23] Choquet

پنج الگوریتم انتخاب ویژگی برخط مبتنی بر جریان ویژگی واحد برای مقایسه با روش پیشنهادی درنظر گرفته شده است. مقایسه‌ها بر اساس دو دسته‌بند درخت تصمیم[۲۴] و $k-$ نزدیکترین همسایه[۲۵] و بر مبنای دو معیار دقت دسته‌بندی[۲۶] و امتیاز F [۲۷] صورت‌گرفته است. همچنین هفت مجموعه‌داده با ابعاد مختلف برای آزمایش‌های تجربی استفاده شده است.

در این مقاله، بخش ۲ به بررسی روش‌های ارائه شده در انتخاب ویژگی برخط مبتنی بر جریان ویژگی واحد اختصاص دارد. در بخش ۳، مفاهیم پایه شامل مفهوم انتگرال فازی و روش‌های ارزیابی ویژگی استفاده شده در روش پیشنهادی را تشریح می‌کنیم. بخش ۴ به تشریح روش پیشنهادی و بخش ۵ به ارائه نتایج آزمایش‌های تجربی می‌پردازد. در نهایت در بخش ۶ جمع‌بندی و نتیجه‌گیری مقاله ارائه شده است.

۲ کارهای مرتبط

با توجه به اینکه موضوع مورد نظر ما در این پژوهش از نوع انتخاب ویژگی برخط بر اساس جریان ویژگی واحد و گروهی است، در این بخش به معرفی روش‌های مطرح در این حوزه می‌پردازیم. در سال ۲۰۰۶، ژوجینگ [۲۸] و همکاران، روش $alpha-investing$ [۱۵] را به‌عنوان یکی از اولین روش‌های انتخاب ویژگی برخط ارائه کردند. این روش بر اساس یک مدل سراسری عمل نمی‌کند و عملکرد آن بر اساس یک رویکرد آماری است. این روش که برای ارزیابی جریان‌های ویژگی واحد ارائه شده است زمانی که یک ویژگی جدید به فضای ویژگی اضافه می‌شود، یک مقدار p برای آن محاسبه می‌کند که بیانگر احتمالی است که بر اساس آن در مورد انتخاب یا رد یک ویژگی تصمیم‌گیری می‌شود. هدف از این روش، کنترل حد آستانه در طول فرایند انتخاب ویژگی بوده است. این امر با انتخاب ویژگی‌های جدید در مدل امکان‌پذیر شده است که با انتخاب ویژگی‌های جدید حد آستانه نیز افزایش می‌یابد و امکان افزایش جزئی در گنجاندن ویژگی‌های نادرست

[24] Decision tree(DT)
[25] K nearest neighbor(KNN)
[26] Classification accuracy
[27] F-score
[28] Zhou Jing

در آینده فراهم می‌شود. به این معنا که افزایش تعداد ویژگی‌های انتخاب شده باعث افزایش مقدار حد آستانه می‌شود و در نتیجه تعداد ویژگی‌های بیشتری در آینده انتخاب می‌شود. همچنین اگر یک ویژگی حد آستانه را ارضا نکند، حد آستانه کاهش می‌یابد. یکی از مزایای استفاده از روش $alpha - investing$ توانایی آن در مدیریت مجموعه ویژگی‌هایی با اندازه‌های ناشناخته حتی تا بی‌نهایت است. از جمله معایب این روش می‌توان به موارد زیر اشاره کرد:

۱) در این روش افزونگی ویژگی‌های انتخاب شده در نظر گرفته نمی‌شود، زیرا تنها یک بار هر ویژگی را ارزیابی می‌کند.

۲) اگر تعداد جریان ویژگی‌ها و تعداد ویژگی‌های مدل فعلی زیاد باشد، از نظر محاسباتی کارآمد نیست، زیرا از مقدار p [۲۹] ویژگی‌ها استفاده می‌کند. از طرفی برای مجموعه‌داده‌های خلوت بسیار ناپایدار است و ممکن است اصلاً ویژگی را انتخاب نکند.

$SAOLA$ [۱۴] یک روش برخط مبتنی بر مقایسه‌های دو به دو است که هدف آن برطرف‌کردن دو چالش مهم در داده‌های حجیم است: ابعاد بسیار بالا و مقیاس‌پذیری آن‌ها. $SAOLA$ با استفاده از تکنیک‌های جدید مقایسه جفتی برخط، یک مدل ساده را در طول زمان به‌صورت برخط حفظ می‌کند. این روش بر روی مجموعه‌داده‌های با ابعاد بسیار بالا مقیاس‌پذیر است. الگوریتم $SAOLA$ مجموعه‌ای از مقایسه‌های جفتی را بین ویژگی‌های واحد به‌جای شرطی کردن مجموعه‌ای از ویژگی‌ها انجام می‌دهد که این امر باعث کاهش هزینه‌های محاسباتی می‌شود. همچنین، $SAOLA$ از یک استراتژی جستجوی حریصانه برای یافتن ویژگی‌های افزونه با بررسی زیرمجموعه‌های ویژگی مختلف در مجموعه ویژگی‌های فعلی استفاده می‌کند. درعین‌حال، نقطه ضعف $SAOLA$ این است که به‌دست‌آوردن یک مقدار بهینه برای حد آستانه دشوار است. $OSFSMI$ [۱۲] یک الگوریتم انتخاب ویژگی برخط است که جریان‌های ویژگی واحد را بر اساس مفهوم اطلاعات متقابل ارزیابی می‌کند. در این روش، هر زمانی که یک ویژگی جدید به فضای ویژگی وارد می‌شود، ابتدا میزان اطلاعات متقابل آن با برچسب کلاس محاسبه می‌شود و درصورتی‌که این مقدار بزرگ‌تر از صفر بود به طور موقت ذخیره می‌شود. ویژگی‌هایی که از این مرحله عبور می‌کنند به فاز بررسی افزونگی می‌رسند و در صورتی

[29] p-value

که ویژگی‌های افزونه‌ای با ورود ویژگی‌های جدیدتر کشف شوند از مجموعه ویژگی‌های انتخابی حذف می‌شوند. افزونگی ویژگی‌ها توسط معیار بهره تعاملی [۳۰] صورت می‌گیرد که مجموع همبستگی هر ویژگی با مجموعه ویژگی‌های فعلی و برچسب کلاس را محاسبه می‌کند. در واقع با انتخاب هر ویژگی و افزودن آن به مجموعه ویژگی‌ها، ویژگی‌های این مجموعه از نظر افزونگی بررسی می‌شوند. برای این منظور، مقدار بهره تعاملی برای همه‌ی ویژگی‌ها محاسبه شده و ویژگی‌ها به ترتیب نزولی مقدار بهره تعاملی مرتب می‌شوند. در صورتی که کمترین مقدار بهره تعاملی مربوط به جدیدترین ویژگی باشد، الگوریتم به مرحله بعد می‌رود و در انتظار ورود ویژگی جدید قرار می‌گیرد. در غیر این صورت، ویژگی با کمترین مقدار بهره تعاملی حذف می‌شود و سپس مقدار بهره تعاملی یک بار دیگر برای همه ویژگی‌ها محاسبه می‌شود و روال قبل تا زمانی که کمترین بهره‌ی تعاملی مربوط به ویژگی جدید باشد، تکرار می‌شود. $OFS-3AM$ [۱۸] یک الگوریتم انتخاب ویژگی برخط مبتنی بر جریان ویژگی واحد است که مرتبط‌ترین ویژگی‌ها را با کمترین افزونگی بر اساس یک مجموعه ناهموار همسایه سازگار [۳۱] انتخاب می‌کند. در این روش سه شاخص ارزیابی برای رویکردهای ناهموار وجود دارد:

۱) حداکثر وابستگی: در این شاخص هدف پیداکردن زیرمجموعه‌ای از ویژگی‌ها است که حداکثر وابستگی را داشته باشند و درعین‌حال حداقل تعداد ویژگی‌ها انتخاب شود. از دیدگاه نظری، حداکثر وابستگی بهترین معیار ارزیابی برای انتخاب ویژگی با مجموعه ناهموار همسایه است. بااین‌حال، تولید کلاس‌های هم‌ارزی حاصل با استفاده از حداکثر وابستگی در فضای با ابعاد بالا دشوار است. زیرا تعداد نمونه‌ها اغلب ناکافی است. در همین حال، سرعت محاسباتی پایین یکی دیگر از معایب وابستگی حداکثری است. علاوه بر این، وابستگی حداکثری برای انتخاب ویژگی برخط مبتنی بر جریان ویژگی واحد مناسب نیست زیرا ما فقط یک ویژگی را در هر پنجره زمانی دریافت می‌کنیم و از کل فضای ویژگی از قبل اطلاعی نداریم.

۲) حداکثر مرتبط بودن: هدف از این شاخص حداکثرسازی مرتبط بودن ویژگی‌های انتخابی با برچسب کلاس است. وابستگی بین ویژگی‌هایی که بر اساس حداکثر ارتباط انتخاب شده‌اند، می‌تواند افزونگی زیادی داشته باشد. به‌عنوان مثال، اگر دو ویژگی به

[30] Interaction gain

[31] Adapted Neighborhood Rough Set

شدت به یکدیگر وابسته باشند و هر دو در زیرمجموعهٔ ویژگی کاندید باشند، پس از حذف یکی از آن‌ها، دقت دسته‌بندی تغییر چندانی نخواهد کرد؛ بنابراین، «حداکثر مرتبط بودن» می‌تواند ویژگی‌هایی را با وابستگی بالا به برچسب‌های کلاس انتخاب کند، اما نمی‌تواند افزونگی را در زیرمجموعهٔ ویژگی انتخاب شده حذف کند.

۳) حداکثر اهمیت: در انتخاب ویژگی برخط مبتنی بر جریان ویژگی، نمی‌توان همه ترکیب‌های ویژگی کاندید را برای به حداکثر رساندن وابستگی مجموعه ویژگی انتخاب شده آزمایش کرد؛ بنابراین، از معیار «حداکثر مرتبط بودن» برای انتخاب ویژگی‌های مرتبط و حذف ویژگی‌های نامرتبط در ابتدا استفاده شده است. اگر یک ویژگی جدید بتواند وابستگی زیرمجموعه انتخاب شده را افزایش دهد، با معیار "حداکثر مرتبط بودن" انتخاب می‌شود. در غیر این صورت، اگر وابستگی افزودن ویژگی جدید برابر با زیرمجموعه انتخابی فعلی باشد، ابتدا ویژگی ورودی جدید و زیرمجموعه انتخابی فعلی ترکیب می‌شوند و سپس از معیار "حداکثر اهمیت" برای حذف ویژگی‌های غیرقابل‌توجه استفاده می‌شود.

الگوریتم $K-OSFD$ [۱۷] یک روش مبتنی بر مفهوم $k-$ نزدیک‌ترین همسایه است. در این روش اطلاعات از طریق نظریه مجموعه‌های ناهموار [۳۲] همسایه به دست می‌آیند و سپس ویژگی‌های باقدرت تمایزدهندگی بیشتر انتخاب می‌شوند. $OFS-Density$ [۱۶] از تئوری مجموعه‌های ناهموار برای حل مسأله انتخاب ویژگی برخط مبتنی بر جریان ویژگی واحد استفاده می‌کند. عملکرد این روش بر اساس یک رابطه همسایگی چگالی جدید است که به طور خودکار تعداد همسایگان را در طول محاسبه وابستگی توسط اطلاعات چگالی نمونه‌های اطراف تعیین می‌کند. با این رابطه همسایگی جدید، نیازی به تعیین هیچ پارامتری از قبل نیست. در همین حال، از یک محدودیت تساوی فازی برای تجزیه و تحلیل افزونگی استفاده شده است که زیرمجموعه ویژگی انتخاب شده را کوچک و متمایز می‌کند. هدف از این الگوریتم، انتخاب ویژگی‌های جدید با حداکثر همبستگی، حداکثر وابستگی و حداقل افزونگی است. $SFS-FI$ [۱۹] تعامل بین ویژگی‌ها را در فرایند انتخاب ویژگی برخط مبتنی بر جریان ویژگی واحد در نظر می‌گیرد. برای این منظور از معیاری به نام بهره تعاملی استفاده شده که قابلیت

[32] Rough set theory

اندازه‌گیری میزان تعامل بین ویژگی جدید و ویژگی‌هایی که قبلاً انتخاب شده‌اند را دارد. معیار بهره تعاملی در این روش بر اساس احتمال پسین تعریف شده است. در این روش برای هر ویژگی جدید مقدار بهره تعاملی آن با مجموعه ویژگی‌های انتخاب شده تا آن زمان، محاسبه شده و با یک حد آستانه از پیش تعریف شده مقایسه می‌شود. اگر مقدار بهره تعاملی ویژگی بیشتر از حد آستانه بود، ویژگی انتخاب و در صورتی که این مقدار از صفر کمتر بود، ویژگی نادیده گرفته می‌شود. در صورتی که مقدار بهره عاملی در بازه صفر تا حد آستانه قرار بگیرد، افزونگی آن مورد بررسی قرار می‌گیرد.

۳ مفاهیم پایه

۱.۳ انتگرال فازی چوکت

در مسائل ترکیبی که از چندین مقدار برای تصمیم‌گیری استفاده می‌کنیم، انتگرال‌های فازی مورداستفاده قرار می‌گیرند. انتگرال‌های فازی بر اساس شاخص‌های فازی[۳۳] عمل می‌کنند که به صورت زیر تعریف می‌شود:

تعریف ۱.۳ (شاخص فازی). فرض کنیم $N = \{۱, ۲, , n\}$. یک شاخص فازی را می‌توان به عنوان $\mu : ۲^N \to [۰, ۱]$ تعریف کرد که شرط‌های زیر را ارضا کند [۳] :

۱. یکنوا [۳۴] باشد: اگر $A \subset B$ ، آنگاه $\mu(A) \leq \mu(B)$

۲. نرمال شده باشد [۳۵] : $\mu(\mathcal{N}) = ۱$ و $\mu(\varnothing) = ۰$

تعریف ۲.۳ (شاخص تراکم فازی). اگر فرض کنیم $X = \{x_۱, x_۲, x_۳, ..., x_n\}$ یک مجموعه متناهی باشد و $\lambda \in (-۱, \infty)$ ، آنگاه شاخص تراکم فازی λ می‌تواند به صورت یک تابع $\mu_\lambda : ۲^x$ تعریف شود که شرایط زیر را ارضا کند [۳] :

۱. $\mu_\lambda(x) = ۱$

[33] Fuzzy measures
[34] Monotic
[35] Normalized

۲. اگر $B \in ۲^x$ ، آنگاه $\mu_\lambda(A \cup B) = \mu_\lambda(A) + \mu_\lambda(B) + \lambda\mu_\lambda(A)\mu_\lambda(B)$ که $A \cap B = \varnothing$

برای به‌دست‌آوردن μ_λ برای ترکیبات مختلف، باید در ابتدا λ محاسبه شود. شاخص تراکم فازی [36] $\mu(X) = \mu_\lambda(\{x_۱, x_۲, x_۳, ..., x_n\})$ به صورت زیر تعریف می‌شود [۱] :

(۱)

$$\mu_\lambda(\{x_۱, x_۲, x_۳, ..., x_n\}) = \sum_{i=۱}^{n} \mu_i + \sum_{i_۲=i_۱+۱}^{n} \mu_{i_۱} + \mu_{i_۲} + ... + \lambda^{n+۱}\mu_۱.\mu_۲...\mu_n$$
$$= \frac{۱}{\lambda}\left[\prod_{i=۱}^{n}(۱ + \lambda\mu_i) - ۱\right]$$

که $\lambda \in (-۱, \infty)$ است. همان‌طور که در قبل اشاره شد، $\mu(X) = ۱$ ، بنابراین می‌توانیم رابطه ۱ را برای محاسبه λ به صورت زیر ساده‌سازی کنیم.

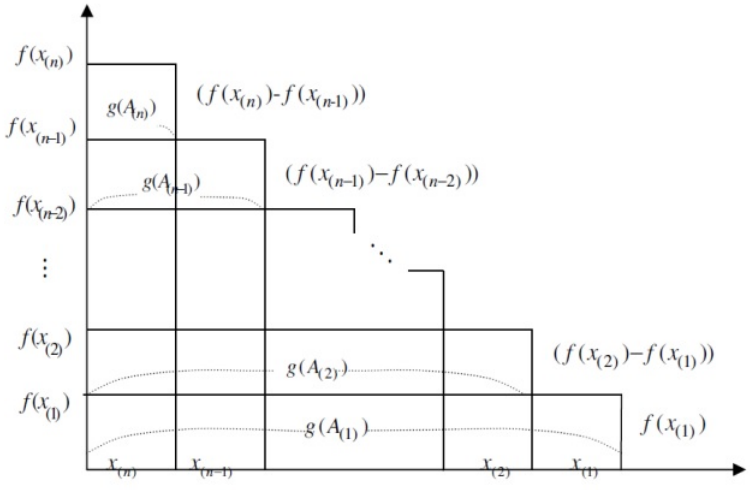
$$\lambda = \prod_{i=۱}^{n}(\lambda\mu_i + ۱) - ۱ \qquad (۲)$$

تعریف ۳.۳ (انتگرال فازی چوکت). فرض کنیم μ یک شاخص تراکم فازی در X باشد، آنگاه انتگرال فازی چوکت با شرط $f : X \to [۰, ۱]$ ،به صورت زیر تعریف می‌شود:

$$\int f d\mu = \sum_{i=۱}^{n}(f(x_i) - f(x_{i-۱})).\mu(A_i) \qquad (۳)$$

که $A_i \subset X$ برای $f(x_۱), f(x_۲), f(x_۳), \cdots, f(x_n), i = ۱, ۲, , n$ بازه‌ها [37] هستند که به صورتی تعریف می‌شوند که $f(x_۱) \leq f(x_۲) \leq f(x_۳) \leq \cdots \leq f(x_n)$ و $f(x_۰) = ۰$ است. مفهوم پایه انتگرال چوکت را می‌توان در تصویر ۱ مشاهده کرد و اگر بخواهیم رابطه ۳ را بر اساس شکل ۱ توجیه کنیم، بر اساس فرض زیر پیش می‌رویم. فرض کنیم $X = \{x_۱, x_۲, x_۳, , x_n\}$ بیانگر n شیء در بازه‌های $\{f(x_۱), f(x_۲), f(x_۳), , f(x_n)\}$ به طوری که

[36] Fuzzy density measure
[37] Ranges



شکل ۱: *مفهوم انتگرال فازی چوکت*

$f(x_۱) \leq f(x_۲) \leq f(x_۳) \leq \cdots \leq f(x_n)$ باشد. همچنین شاخص‌های این اشیاء $\{(A_۱), (A_۲), (A_۳), \cdots, (A_n)\}$ باشند. بنابراین مساحت مستطیل اول در شکل ۱ برابر $f(x_۱) \times \mu(A_۱)$ که $\mu(A_۱) = (\{x_۱, x_۲, x_۳, \cdots, x_n\})$ است. همچنین مساحت مستطیل دوم و سوم نیز به ترتیب برابر $(f(x_۲) - f(x_۱)) \times \mu(A_۲)$ و $(f(x_۳) - f(x_۲)) \times \mu(A_۳)$ هستند که $\mu(A_۱) = (\{x_۲, x_۳, \cdots, x_n\})$ و $\mu(A_۳) = (\{x_۳, \cdots, x_n\})$. بنابراین با توجه به مساحت‌های محاسبه شده می‌توانیم یک قاعده کلی برای محاسبه مساحت مستطیل n -ام ارائه کنیم که برابر است با $(f(x_n) - f(x_{n-۱})) \times \mu(A_n)$ که در آن $\mu(A_n) = (\{x_n\})$ است. حال مجموع مساحت مستطیل‌ها به صورت زیر محاسبه می‌شود.

(۴)

$$\int f d\mu = f(x_۱) \times \mu(A_۱) + (f(x_۲) - f(x_۱)) \times \mu(A_۲) + (f(x_۳) - f(x_۲)) \times \mu(A_۳) + \cdots + (f(x_n) - f(x_{n-۱})) \times \mu(A_n)$$

حال اگر رابطه بالا را ساده کنیم و به یک حالت عمومی تبدیل کنیم، رابطه ۳ که انتگرال چوکت است را بدست می‌آوریم.

برای درک بهتر مسئله از یک مثال عددی استفاده می‌کنیم. مجموعه $X = \{x_1, x_2, x_3\}$ را درنظر بگیرید که بازه‌ها به صورت $f(x_1) = 0.4$ ، $f(x_2) = 0.6$ و $f(x_3) = 0.8$ هستند. همچنین مقادیر شاخص‌های تراکم فازی در آن به صورت $\mu_\lambda(\{x_1\}) = 0.4$ ، $\mu_\lambda(\{x_2\}) = 0.3$ و $\mu_\lambda(\{x_3\}) = 0.2$ وجود دارند. حال برای بدست آوردن μ_λ برای ترکیبات مختلف عناصر X باید ابتدا λ را با استفاده از رابطه ۲ محاسبه کنیم:

$$\lambda = (0.4\lambda - 1)(0.3\lambda - 1)(0.2\lambda - 1) - 1$$

اگر این معادله را حل کنیم به مقادیر $\lambda = [0, -11.87, 0.3719]$ می‌رسیم. از قبل می‌دانیم که $\lambda \in (-1, \infty)$ ،پس مقادیر قابل قبول برای λ برابر عدد صفر و ۳۷۱۹.۰ هستند. در این مثال، رابطه را برای $\lambda = 0.3719$ محاسبه می‌کنیم. مقادیر μ_λ برای محاسبه تمام حالت‌ها در X بر اساس تعریف ۲.۳ بدست می‌آیند:

$$\mu_\lambda(\{x_1, x_2\}) = \mu_\lambda(\{x_1\}) + \mu_\lambda(\{x_2\}) + \lambda\mu_\lambda(\{x_1\})\mu_\lambda(\{x_2\}) = 0.7446$$

$$\mu_\lambda(\{x_2, x_3\}) = \mu_\lambda(\{x_2\}) + \mu_\lambda(\{x_3\}) + \lambda\mu_\lambda(\{x_2\})\mu_\lambda(\{x_3\}) = 0.5223$$

$$\mu_\lambda(\{x_1, x_3\}) = \mu_\lambda(\{x_1\}) + \mu_\lambda(\{x_3\}) + \lambda\mu_\lambda(\{x_1\})\mu_\lambda(\{x_3\}) = 0.7323$$

$$\mu_\lambda(\{x_1, x_2, x_3\}) = 1$$

در نهایت خواهیم داشت:

$$\int f d\mu = \sum_{i=۱}^{n} (f(x_i) - f(x_{i-۱})).\mu(A_i) = \int f d\mu = f(x_۱) \times \mu(A_۱)$$
$$+ (f(x_۲) - f(x_۱)) \times \mu(A_۲) + (f(x_۳) - f(x_۲)) \times \mu(A_۳)$$
$$= ۰/۴ \times ۱ + (۰/۶ - ۰/۹) \times ۰/۵۲۲۳ + (۰/۸ - ۰/۶) \times ۲ = ۰/۵۴۴۵$$

۲.۳ عدم قطعیت متقارن

عدم قطعیت متقارن[۳۸] یک روش برای اندازه‌گیری میزان مرتبط بودن یک ویژگی نسبت به برچسب کلاس است. این روش بر اساس مفاهیم بهره و آنتروپی اطلاعات[۳۹] تعریف می‌شود. اگر ما X را یک متغیر گسسته درنظر بگیریم، آنگاه آنتروپی X به صورت زیر محاسبه می‌شود:

$$H(X) = -\sum_{x \in X} P(x) \log_۲(P(x)) \qquad (۵)$$

که $x \in X$ و $p(X)$ برابر احتمال x برای همه مقادیر محتمل در X است. از طرفی دیگر، آنتروپی شرطی بین دو متغیر تصادفی گسسته X و Y با استفاده از رابطه زیر محاسبه می‌شود:

$$H(X|Y) = -\sum_{x \in X} P(y) \sum_{x \in X} P(x|y) \log_۲(P(x|y)) \qquad (۶)$$

با توجه به روابط ۵ و ۶ بهره اطلاعاتی[۴۰] به‌صورت زیر تعریف می‌شود:

$$IG(X|Y) = H(X) - H(X|Y) \qquad (۷)$$

[38] Symmetrical uncertainty(SU)

[39] Information entropy

[40] Information gain(IG)

بهره اطلاعاتی در واقع میزان همبستگی دو متغیر تصادفی را می‌سنجد. اگر این مقدار برای دو متغیر برابر صفر باشد، به این معناست که از هم مستقل هستند و میزان ارتباط آنها با افزایش مقدار بهره اطلاعاتی افزایش می‌یابد. ضعف عمده روش بهره اطلاعاتی این است که به سمت متغیرهای با مقادیر متنوع متمایل می‌شود. بنابراین برای حل این مشکل این مقادیر در بازه [۰, ۱] نرمال می‌شوند تا این ضعف پوشش داده شود. این مقدار نرمال اصطلاحاً عدم قطعیت متقارن نامیده می‌شود که از طریق رابطه زیر محاسبه می‌شود:

$$SU(X,Y) = ۲[\frac{IG(X|Y)}{H(X)+H(Y)}] \quad (۸)$$

مقادیر صفر و یک به ترتیب به معنای مستقل بودن و وابستگی کامل دو متغیر است.

۴ روش پیشنهادی

در این بخش به معرفی و تشریح روش پیشنهادی می‌پردازیم. این روش برای مسئله انتخاب ویژگی برخط واحد ارائه شده است که در آن از یک رویکرد ترکیبی برای تصمیم‌گیری در مورد جریان‌های ویژگی استفاده شده است. در پژوهش‌های مختلف به‌دفعات نشان داده شده است که تنوع روش‌های مختلف در رویکردهای ترکیبی از به دام افتادن در بهینه محلی جلوگیری می‌کند و همچنین باعث بهبود عملکرد الگوریتم‌ها می‌شود. در مسائل انتخاب ویژگی، این بهبود به دلیل استفاده از چندین شاخص مختلف در ارزیابی هر ویژگی است که هر شاخص، ویژگی‌ها را از یک جنبه مشخص ارزیابی می‌کند؛ بنابراین ویژگی‌های متمایز و بهینه آن دسته از ویژگی‌ها هستند که در برآیند کلی شاخص‌ها امتیاز بالاتری را کسب کنند. در روش پیشنهادی هر ویژگی به محض ورود از نظر مرتبط‌بودن با برچسب کلاس مورد ارزیابی قرار می‌گیرد و سپس مجموعه ویژگی‌های انتخاب شده در هر مرحله بر اساس ترکیب سه معیار افزونگی با استفاده مفهوم انتگرال فازی چوکت به‌روزرسانی می‌شود. مراحل گام‌به‌گام روش پیشنهادی در الگوریتم ارائه شده در شکل ۲ نمایش داده شده است و در ادامه نیز این مراحل را شرح خواهیم داد.

شکل ۲: *شبه کد الگوریتم پیشنهادی*

در مسئله انتخاب ویژگی برخط مبتنی بر جریان ویژگی واحد در هر لحظه یک ویژگی وارد فضای ویژگی می‌شود و باید نسبت به انتخاب یا نادیده‌گرفتن آن به‌صورت برخط تصمیم‌گیری شود. روش پیشنهادی در این مقاله مبتنی بر یک رویکرد ترکیبی است که از ترکیب چندین روش ارزیابی ویژگی‌ها برای تصمیم‌گیری استفاده می‌کند. با توجه به این که فرایند ترکیب شاخص‌های ارزیابی توسط عملگر انتگرال فازی چوکت انجام می‌گیرد، در ابتدای الگوریتم باید مقادیر تراکم فازی برای هر یک از شاخص‌ها تنظیم شود. این روند در گام دوم الگوریتم و پس از درنظرگرفتن یک بردار تهی برای ذخیره ویژگی‌های انتخاب شده قرار گرفته است. به طور کلی الگوریتم پیشنهادی در این مقاله در دو سطح مرتبط‌بودن و افزونگی ویژگی‌ها صورت می‌گیرد. افزونگی مجموعه ویژگی‌های انتخاب شده بر اساس مفهوم فاصله اقلیدسی انجام می‌شود که سه مقدار ۰/۴، ۰/۴ و ۰/۲ به‌عنوان مقادیر تراکم فازی به ترتیب برای مقدار حداکثر، حداقل و مجموع فاصله اقلیدسی بین هر ویژگی با سایر ویژگی‌ها تعیین شده‌اند. همان‌طور که در بخش ۱.۳ اشاره شد، برای محاسبه مقدار انتگرال فازی چوکت باید مقادیر μ_λ برای تمام حالت‌های ممکن روش‌های دخیل در فرایند ترکیب محاسبه شوند. بنابراین در گام سوم الگوریتم، ابتدا مقدار λ بر اساس مقادیر تراکم فازی و رابطه ۲ محاسبه می‌شود، که معادله زیر بدست می‌آید:

$$\lambda = (۰/۴\lambda - ۱)(۰/۴\lambda - ۱)(۰/۲\lambda - ۱) - ۱$$

با حل این معادله به ریشه‌های ۱۰- و صفر می‌رسیم که با توجه به‌شرط $\lambda \in (-۱, \infty)$ ،مقدار $\lambda = ۰$ قابل قبول است. سپس مقدار μ_λ را برای همه ترکیبات سه تابع محاسبه می‌شوند:

$$\mu_\lambda(\{max, min\}) = ۰/۴ + ۰/۴ = ۰/۸$$

$$\mu_\lambda(\{min, sum\}) = ۰/۴ + ۰/۲ = ۰/۶$$

$$\mu_\lambda(\{max, sum\}) = ۰/۴ + ۰/۲ = ۰/۶$$

$$\mu_\lambda(\{max, min, sum\}) = ۱$$

در گام چهارم الگوریتم، با استفاده از یک ساختار تکراری ویژگی‌های جدید ارزیابی می‌شوند. با ورود هر ویژگی جدید، مقادیر عدم قطعیت متقارن بر اساس رابطه ۸ بین آن ویژگی و برچسب کلاس محاسبه می‌شود. این فرایند در گام‌های ۵ تا ۶ الگوریتم پیشنهادی در شکل ۲ نشان داده شده است. اگر تاکنون هیچ ویژگی انتخاب نشده باشد، محاسبه مقدار افزونگی امکان‌پذیر نیست؛ بنابراین ویژگی فقط بر اساس مقدار همگرایی محاسبه شده ارزیابی می‌شود. اگر مقدار شاخص از حد آستانه α بزرگ‌تر بود، ویژگی انتخاب و در غیراینصورت نادیده گرفته می‌شود و الگوریتم منتظر ورود جریان ویژگی بعدی می‌شود که گام‌های ۸ تا ۹ الگوریتم بیانگر این مراحل هستند. با انتخاب اولین ویژگی برای سایر جریان‌های ویژگی، مقادیر حداکثر، حداقل و مجموع فاصله اقلیدسی بین هر ویژگی با سایر ویژگی‌ها برای سنجش افزونگی محاسبه می‌شود. بنابراین با ورود ویژگی‌های جدید، به‌روزرسانی مجموعه ویژگی‌هایی که تاکنون انتخاب شده‌اند بر اساس ترکیب این سه شاخص و توسط انتگرال فازی چوکت انجام می‌شود.

همچنین با توجه به درنظرگرفتن شرط $f(x_۱) \le f(x_۲) \le f(x_۳) \le \cdots \le f(x_n)$ برای بازه‌ها در انتگرال چوکت، برای هر ویژگی جدید، پس از محاسبه مقادیر شاخص ها، این مقادیر به صورت صعودی مرتب می‌شوند و تحت عنوان $f(x_۱), f(x_۲), f(x_۳)$ درنظر گرفته می‌شوند. بنابراین در ارزیابی هر ویژگی، هر بار ممکن است ترتیب متفاوتی از روش‌های ارزیابی را داشته باشیم. شاخص‌های تراکم فازی متناظر نیز بر همین اساس در هر محاسبه انتگرال فازی چوکت متناظر با تغییرات در مقادیر شاخص‌ها به شاخص متناظر خود نسبت داده می‌شوند.

برای این منظور، مقدار انتگرال فازی چوکت بر اساس مقادیر تراکم فازی و مقادیر به‌دست‌آمده توسط سه شاخص برای همه ویژگی‌هایی که تاکنون انتخاب شده‌اند با استفاده از رابطه ۳ به‌دست می‌آید. در مرحله بعد ویژگی با کمترین مقدار انتگرال فازی چوکت مشخص می‌شود. اگر این ویژگی جدیدترین ویژگی وارد شده به مجموعه باشد، هیچ ویژگی حذف نخواهد شد. درغیراینصورت ویژگی با کمترین مقدار انتگرال فازی چوکت حذف می‌شود و این فرایند برای ویژگی‌های باقی‌مانده ادامه می‌یابد تا زمانی که ویژگی با کمترین مقدار جدیدترین ویژگی شود. در نهایت در صورتی که جریان ویژگی جدیدی به فضای

ویژگی وارد نشود، مجموعه ویژگی‌های انتخابی برای فرایند یادگیری استفاده می‌شود. این مراحل در گام های ۱۱ تا ۱۷ الگوریتم پیشنهادی در شکل ۲ قرار دارند. قابل ذکر است که مقدار α برابر ۳/۰ در این روش به صورت آزمون و خطا تنظیم شده است.

۵ نتایج آزمایش‌های تجربی

در این بخش، عملکرد روش پیشنهادی در مقایسه با چندین روش انتخاب ویژگی برخط مبتنی بر جریان ویژگی واحد مورد ارزیابی قرار گرفته است. برای این منظور الگوریتم‌های $OSFSMI$ [۱۲] ، $OFS-Density$ [۱۶] ، $OFS-3AM$ [۱۸] ، $Alpha-investigating$ [۱۵] و $OFSS-FI$ [۱۹] برای مقایسه با روش پیشنهادی درنظر گرفته شده‌اند.

۱.۵ مجموعه‌داده‌ها

هفت مجموعه‌داده دنیای واقعی برای ارزیابی عملکرد الگوریتم‌های انتخاب ویژگی برخط با جریان ویژگی واحد و مقایسه آنها با الگوریتم پیشنهادی مورداستفاده قرار گرفته‌اند. مشخصه‌های این مجموعه‌داده‌ها در جدول ۱ ارائه شده است.

جدول ۱: *مشخصات مجموعه‌داده‌ها*

مجموعه داده	تعداد نمونه‌ها	تعداد ویژگی‌ها	تعداد کلاس‌ها	نوع مجموعه‌داده
Sorlie	۸۵	۴۵۷	۵	ریزآرایه
ColonTumor	۶۲	۲۰۰۰	۲	ریزآرایه
Prostate-GE	۱۰۲	۵۹۶۷	۲	زیست‌شناسی
۱clean	۴۷۶	۱۶۸	۲	تصویر
SRBCT	۸۲	۲۳۰۹	۴	ریزآرایه
Leukemia	۶۲	۲۰۰۱	۲	تصویر
Subramanian	۵۰	۱۰۱۰۱	۲	تصویر

۲.۵ شاخص‌های ارزیابی عملکرد

شاخص‌های دقت و امتیاز F برای مقایسه عملکرد الگوریتم‌ها در این مقاله انتخاب شده‌اند که بر اساس مفاهیم زیر تعریف می‌شوند:

۱. مثبت کاذب[۴۱] FP : درصدی از نمونه‌های با کلاس منفی که به صورت نادرست، مثبت دسته‌بندی شده‌اند.

۲. مثبت حقیقی[۴۲] TP : درصدی از نمونه‌های با کلاس مثبت که به‌درستی دسته‌بندی شده‌اند.

۳. منفی حقیقی[۴۳] TN : درصدی از نمونه‌های با کلاس منفی که به درستی دسته‌بندی شده‌اند.

۴. منفی کاذب[۴۴] FN : ددرصدی از نمونه‌های با کلاس مثبت که به صورت نادرست، منفی دسته‌بندی شده‌اند.

با استفاده از این مفاهیم، شاخص‌های دقت و امتیاز F به صورت زیر تعریف می‌شوند:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{۹}$$

$$F - Score = \frac{TP}{TP + \frac{۱}{۲}(FP + FN)} \tag{۱۰}$$

۳.۵ نتایج

در این بخش نتایج به‌دست‌آمده از آزمایش‌های تجربی ارائه شده‌اند. در این مقایسه‌ها، برای تمام روش‌های انتخاب ویژگی برخط که با روش پیشنهادی مقایسه شده‌اند، مقادیر پارامترها بر اساس مقادیر پیشنهادی در مقاله‌های مذکور تنظیم شده‌اند. مقایسات بر

[41] False Positive(FP)
[42] True Positive(TP)
[43] True Negative(TN)
[44] False Negative(FN)

اساس عملکرد الگوریتم‌ها در عمل دسته‌بندی با استفاده از دو دسته‌بند k -نزدیکترین همسایه و درخت تصمیم صورت‌گرفته است. تعداد همسایه‌های نزدیک در این آزمایش‌ها برابر عدد ۳ قرار گرفته است.

در گام یادگیری ۶۰ درصد از نمونه‌های آموزشی هر مجموعه‌داده به‌صورت تصادفی به مجموعه آموزشی و ۴۰ درصد باقی‌مانده به‌عنوان مجموعه آزمون در نظر گرفته شده‌اند. همچنین نتایج ارائه شده در این مقاله، میانگین ۳۰ بار اجرای مجزا برای هر الگوریتم در هر مجموعه‌داده است. نتایج دسته‌بندی به ترتیب برای دسته‌بند k -نزدیکترین همسایه و درخت تصمیم بر اساس معیار دقت در جداول ۲ و ۳ و برای معیار امتیاز F در جداول ۴ و ۵ گزارش شده‌اند. همچنین میانگین تعداد ویژگی‌های انتخابی در هر الگوریتم در جدول ۶ و میانگین زمان اجرای الگوریتم‌ها در جدول ۷ ارائه شده‌اند. در نهایت نتایج دسته‌بندی بر اساس یک رویکرد آماری در جدول ۸ با هم مقایسه شده‌اند. این مقایسه بر اساس روش آماری فریدمن [۴۵] صورت گرفته است. در این جدول مقادیر P [۴۶] حاصل از مقایسه آماری روش پیشنهادی و الگوریتم‌های مشابه گزارش شده‌اند. همچنین حد آستانه مورد نظر در این روش ۰۵.۰ درنظر گرفته شده است. به این صورت که اگر مقدار P حاصل از مقایسه آماری کمتر از حد آستانه باشد، دو روش از نظر آماری فاصله معناداری دارند. درغیراینصورت دو روش تقریباً در وضعیت یکسانی قرار دارند. علامت $+$ در جدول، بیانگر برتری، علامت $-$ نشان‌دهنده شکست و علامت $=$ بیانگر تساوی روش پیشنهادی در مقابل روش مقایسه‌شده دارد. همچنین تعداد برد/تساوی/باخت روش پیشنهادی در مقابل هر روش در سطر آخر جدول قرار گرفته است. گفتنی است که روش فریدمن بین روش پیشنهادی و سایر روش‌ها بر اساس نتایج آنها در دو دسته‌بند و طبق دو معیار ارزیابی دسته‌بندی صورت گرفته است.

روش پیشنهادی در این مقاله، یک الگوریتم جدید برای انتخاب ویژگی برخط مبتنی بر جریان ویژگی واحد است که از عملگر انتگرال فازی چوکت برای ترکیب تکنیک‌های ارزیابی ویژگی متعدد استفاده می‌کند. در این روش، از ترکیبی از معیارهای مبتنی بر افزونگی و مبتنی بر مرتبط بودن با برچسب کلاس برای درنظرگرفتن تنوع و همگرایی در الگوریتم پیشنهادی استفاده شده است. این فرایند باعث انتخاب زیرمجموعه ویژگی‌های

[45] Friedman

[46] P-value

جدول ۲: *نتایج بر اساس معیار دقت در دسته‌بند k -نزدیکترین همسایه*

مجموعه داده	OSFSMI	M۳OFS-A	OFS-Density	Alpha-Investing	OFSS-FI	روش پیشنهادی
ColonTumor	۷۱۲۵/۰	۶۷۹۲/۰	۶۸۹۶/۰	۵۹۹۲/۰	۷۲۰۸/۰	**۷۲۱۴/۰**
Leukemia	۸۹۸۲/۰	۸۲۳۲/۰	۹۱۴۳/۰	۸۶۷۹/۰	۹۱۹۶/۰	**۹۳۰۴/۰**
Sorlie	۵۷۲۱/۰	۷۰۰۰/۰	۶۶۱۸/۰	۲۶۶۲/۰	**۷۲۰۶/۰**	۶۳۹۷/۰
Prostate-GE	۸۳۰۰/۰	۷۴۷۵/۰	۸۹۸۸/۰	۸۶۶۳/۰	۸۵۷۵/۰	**۹۰۲۵/۰**
SRBCT	۷۱۳۶/۰	۸۰۹۱/۰	۸۳۱۸/۰	۵۶۳۶/۰	۹۰۱۵/۰	**۹۰۳۰/۰**
Subramanian	۶۱۵۰/۰	۶۳۰۰/۰	۶۱۷۵/۰	۴۸۰۰/۰	**۶۵۲۵/۰**	۶۲۷۵/۰
\Clean	۶۹۵۵/۰	۹۲۰۸/۰	۹۲۰۸/۰	۸۰۳۷/۰	۸۵۳۲/۰	**۹۲۵۰/۰**

جدول ۳: *نتایج بر اساس معیار دقت در دسته‌بند درخت تصمیم*

مجموعه داده	OSFSMI	M۳OFS-A	OFS-Density	Alpha-Investing	OFSS-FI	روش پیشنهادی
ColonTumor	۷۰۸۳/۰	۶۷۵۰/۰	**۷۳۷۵/۰**	۶۰۶۳/۰	۷۱۰۴/۰	۵۶۸۸/۰
Leukemia	۸۸۷۵/۰	۸۲۵۰/۰	۹۰۸۹/۰	۸۹۶۴/۰	۸۷۵۰/۰	**۹۰۵۴/۰**
Sorlie	۵۸۹۷/۰	۵۲۷۹/۰	۵۸۶۸/۰	۲۹۷۱/۰	۵۷۹۴/۰	**۶۰۰۰/۰**
Prostate-GE	۸۰۷۵/۰	۷۷۸۸/۰	**۸۶۵۰/۰**	۸۲۸۸/۰	۸۳۵۰/۰	۸۲۳۸/۰
SRBCT	۷۶۰۶/۰	۸۱۸۲/۰	۸۰۰۰/۰	۶۸۹۴/۰	۷۹۲۴/۰	**۷۹۲۴/۰**
Subramanian	۵۷۰۰/۰	۶۰۷۵/۰	۵۹۲۵/۰	۵۵۵۰/۰	۶۳۷۵/۰	**۶۴۲۱/۰**
\Clean	۶۹۰۳/۰	۹۱۶۸/۰	۹۱۶۸/۰	۸۲۱۸/۰	۶۹۰۳/۰	**۹۱۹۵/۰**

با افزونگی کمتر و همبسته با برچسب کلاس می‌شود. در روش‌های انتخاب ویژگی برخط مبتنی بر جریان ویژگی که تاکنون ارائه شده‌اند، رویکرد تصمیم‌گیری برای انتخاب ویژگی‌ها با استفاده از یک معیار مشخص است. این ویژگی در صورتی انتخاب می‌شود که مقدار به‌دست‌آمده بالاتر از یک آستانه از پیش تعریف شده باشد، در غیر این صورت نادیده گرفته می‌شود. در این مقاله روش جدیدی ارائه شده است که معیار تصمیم‌گیری در مورد پذیرش ویژگی‌ها، بر خلاف روش‌های فعلی، بر اساس چندین معیار است.

ازآنجایی‌که تنوع روش‌های مختلف می‌تواند منجر به نتیجه بهتری نسبت به یک معیار شود، انتگرال فازی چوکت برای ترکیب معیارهای تصمیم‌گیری استفاده شده است. انتگرال فازی چوکت برای این کار استفاده می‌شود زیرا می‌تواند چندین مقدار با تأثیر متفاوت بر نتایج نهایی را بر اساس یک دیدگاه فازی ترکیب کند. جداول ۲، ۳، ۴ و ۵ نشان می‌دهند که روش پیشنهادی نسبت به روش‌های رقیب، بر اساس دو معیار دسته‌بندی و دو دسته‌بند متفاوت برتری دارد. جدول ۶ میانگین تعداد ویژگی‌های انتخاب شده برای هر الگوریتم را نشان می‌دهد. الگوریتم پیشنهادی مجموعه کوچکی از ویژگی‌ها را انتخاب می‌کند و درعین‌حال بهترین عملکرد را در فرایند دسته‌بندی دارد. همچنین روش پیشنهادی میانگین زمان اجرای پایینی را ثبت کرده است درحالی‌که برخلاف سایر روش‌ها از سه معیار متفاوت در تصمیم‌گیری استفاده می‌کند. روش پیشنهادی بر اساس

جدول ۴: *نتایج بر اساس معیار امتیاز F در دسته‌بند k -نزدیکترین همسایه*

مجموعه داده	OSFSMI	M۳OFS-A	OFS-Density	Alpha-Investing	OFSS-FI	روش پیشنهادی
ColonTumor	۷۷۱۲/۰	۷۴۶۷/۰	۷۳۸۲/۰	۵۶۹۳/۰	۷۸۱۱/۰	۷۹۰۶/۰
Leukemia	۹۲۰۳/۰	۸۶۹۰/۰	۹۳۵۲/۰	۸۹۸۷/۰	۹۴۰۵/۰	۹۴۴۰/۰
Sorlie	۵۵۱۰/۰	۶۷۵۹/۰	۶۴۷۷/۰	۳۱۸۲/۰	۷۳۱۱/۰	۶۲۴۸/۰
Prostate-GE	۸۳۲۷/۰	۷۳۷۰/۰	۸۹۳۵/۰	۸۶۷۷/۰	۸۴۷۷/۰	۹۰۰۷/۰
SRBCT	۶۸۹۸/۰	۷۹۷۴/۰	۸۲۵۰/۰	۵۷۵۱/۰	۹۰۰۹/۰	۸۹۸۱/۰
Subramanian	۷۲۸۱/۰	۷۱۵۲/۰	۷۱۶۹/۰	۶۰۴۷/۰	۷۴۸۷/۰	۷۴۵۰/۰
۱Clean	۷۲۸۱/۰	۹۲۹۳/۰	۹۲۹۳/۰	۸۲۴۹/۰	۸۶۱۳/۰	۹۳۰۲/۰

جدول ۵: *نتایج بر اساس معیار امتیاز F در دسته‌بند درخت تصمیم*

مجموعه داده	OSFSMI	M۳OFS-A	OFS-Density	Alpha-Investing	OFSS-FI	روش پیشنهادی
ColonTumor	۷۸۵۹/۰	۷۰۱۷/۰	۷۳۵۷/۰	۷۱۹۵/۰	۷۵۳۰/۰	۷۷۷۳/۰
Leukemia	۹۱۴۹/۰	۸۶۸۸/۰	**۹۲۹۶/۰**	۹۲۳۵/۰	۹۰۶۶/۰	۹۲۵۹/۰
Sorlie	۵۷۲۳/۰	۵۱۰۴/۰	۵۸۱۱/۰	۳۳۵۳/۰	۵۷۰۱/۰	**۵۷۶۰/۰**
Prostate-GE	۷۹۸۸/۰	۷۵۵۴/۰	**۸۶۱۷/۰**	۸۲۲۴/۰	۸۲۵۰/۰	۸۲۵۲/۰
SRBCT	۷۷۹۹/۰	۸۰۳۹/۰	۸۱۰۴/۰	۷۰۹۷/۰	۷۹۸۰/۰	۷۸۰۰/۰
Subramanian	۶۶۰۱/۰	۶۸۸۳/۰	۶۸۵۸/۰	۶۶۸۵/۰	۷۱۴۱/۰	**۷۲۶۸/۰**
۱Clean	۷۳۳۲/۰	۹۲۶۲/۰	۹۲۶۲/۰	۸۴۱۷/۰	۸۱۰۸/۰	**۹۲۸۴/۰**

نتایج به‌دست‌آمده تقریباً دو درصد بهبود نسبت به روش‌های مشابه داشته است.

۶ نتیجه‌گیری

این مقاله یک روش انتخاب ویژگی برخط مبتنی بر جریان ویژگی را با استفاده از انتگرال فازی چوکت پیشنهاد می‌کند که در آن سه معیار ارزیابی ویژگی برای تصمیم‌گیری در خصوص انتخاب ویژگی‌ها در نظر گرفته می‌شود. این مقادیر به عملگر انتگرال فازی چوکت تحویل داده می‌شود تا فرایند انتخاب ویژگی بر اساس ترکیب این مقادیر انجام پذیرد. نتایج اجرای الگوریتم بر روی مجموعه‌داده‌های مختلف نشان‌دهنده کارایی و بهینه بودن روش پیشنهادی است، زیرا همگرایی و تنوع الگوریتم با انواع روش‌های انتخاب ویژگی متعادل می‌شود که در نتایج به‌دست‌آمده عملکرد الگوریتم پیشنهادی بر روش‌های رقیب برتری دارد. همچنین به دلیل محاسبات ساده در فرایند الگوریتم، روش در زمان اجرای بسیار کوتاهی اجرا می‌شود و از طرفی مجموعه ویژگی‌های انتخاب شده در هر مجموعه‌داده دارای اعضای کمی است. به‌عنوان کارهای آینده قصد داریم از تکنیک‌های ترکیبی در انتخاب ویژگی برای سایر کاربردهای یادگیری ماشین استفاده کنیم و رویکرد شورایی را در مسائل مختلف بهینه‌سازی در یادگیری ماشین مورد آزمایش قرار دهیم.

جدول ۶: *میانگین تعداد ویژگی‌های انتخاب شده*

مجموعه داده	OSFSMI	M۳OFS-A	OFS-Density	Alpha-Investing	OFSS-FI	روش پیشنهادی
ColonTumor	۴/۹	۲۵/۲۲	۹۵/۴	۶/۱	۱۹۲۱	۱۵/۲
Leukemia	۰۵/۷	۵۵/۹	۱۰/۱۰	۷۵/۴	۱۲۲۲	۳/۱۷
Sorlie	۳/۹	۲۵/۳۸	۸/۴	۲/۱	۴۵۶	۶/۹
Prostate-GE	۲۵/۹	۹/۳۸	۵۵/۶	۸	۹۵/۱۸۶	۶۵/۴
SRBCT	۲۵/۱۷	۴۵/۱۲	۲/۱۱	۵/۴	۹/۱۳	۲۲۷۳
Subramanian	۱۵/۱۱	۳/۱۵	۳۵/۴	۵۵/۱	۲۴۳۶	۴/۱۰
۱Clean	۵/۴	۱	۱	۷۵/۱۳	۱۶۷	۱
میانگین	۹	۵/۱۹	۲/۵	۴/۶	۵/۱۲۳۷	۹/۸

جدول ۷: *میانگین زمان اجرا*

مجموعه داده	OSFSMI	M۳OFS-A	OFS-Density	Alpha-Investing	OFSS-FI	روش پیشنهادی
ColonTumor	۷۱/۸	۶۴/۰	۲۶/۰	۰۳/۰	۶۶/۱۱	۳/۰
Leukemia	۷۵/۳۴	۲۹/۴	۱۰/۱	۲۰/۰	۷۹/۴۵	۱۴/۰
Sorlie	۷۵/۱	۳۴/۰	۱۱/۰	۰۰۵/۰	۸۵۳۵	۰۱/۰
Prostate-GE	۱۷/۲۰	۳۸/۴	۵۸/۱	۱۹/۰	۵۱/۲	۱۰/۰
SRBCT	۳۷/۱۲	۸۱/۱	۴۶/۰	۰۷/۰	۳۹/۱۷	۰۵.۰
Subramanian	۳۹/۶۲	۹۱/۳	۰۷/۱	۳۱/۰	۸۳/۱۰۵	۱۸/۰
۱Clean	۶۵/۱۳	۷۴/۰	۷۱/۰	۰۲/۰	۱۳/۱۳	۰۳/۰
میانگین	۹۷/۲۱	۳۰/۲	۷۵/۰	۱۱/۰	۱۷/۲۸	۰۷/۰

همچنین، قصد داریم کار خود را در مسائل انتخاب ویژگی برخط مبتنی بر جریان ویژگی بهبود بخشیم و روش‌های قدرتمندتری هم در جریان‌های واحد و گروهی پیشنهاد کنیم.

مراجع

[1] Larbani, M., Chi, H., Gwo, T. (2011) A Novel Method for Fuzzy Measure Identification. *International Journal of Fuzzy Systems*, **13**, 24–34.

[2] Bai, S., Lin, Y., Lv, Y., Chen, J. and Wang, C. (2021) Kernelized fuzzy rough sets based online streaming feature selection for large-scale hierarchical classification. *Applied Intelligence*, **51**, 1602–1615.

[3] CBeliakov, G. and Divakov, D. (2020) On representation of fuzzy measures for learning Choquet and Sugeno integrals. *Knowledge-Based Systems*, **189**, 105134.

[4] Dhal, P. and Azad, C. (2021) A comprehensive survey on feature selection in the various fields of machine learning. *Applied Intelligence*.

جدول ۸: نتایج مقایسه آماری فریدمن

روش پیشنهادی در مقابل	OSFSMI	MOFS-A	OFS-Density	Alpha-Investing	OFSS-FI
معیار دقت در دسته‌بند k- همسایه نزدیکتر	۰۳۰۲/۰(+)	۰۷۱/۰(=)	۰۳۰۲/۰(+)	۰۰۳۹/۰(+)	۰۷۱/۰(=)
معیار دقت در دسته‌بند درخت تصمیم	۰۳۰۲/۰(+)	۰۳۰۲/۰(+)	۱۴۸۶/۰(=)	۰۳۰۲/۰(+)	۰۴۴۶/۰(+)
معیار امتیاز F در دسته‌بند k- همسایه نزدیکتر	۰۰۳۹/۰(+)	۰۳۰۲/۰(+)	۰۳۰۲/۰(+)	۰۰۳۹/۰(+)	۱۴۸۶/۰(=)
معیار امتیاز F در دسته‌بند درخت تصمیم	۰۳۰۲/۰(+)	۰۳۰۲/۰(+)	۲۷۸۶/۰(=)	۰۰۳۹/۰(+)	۰۳۰۲/۰(+)
برد/مساوی/باخت	۴/۰/۰	۳/۱/۰	۲/۲/۰	۴/۰/۰	۲/۲/۰

[5] Hashemi, A., Dowlatshahi, M.B. and Nezamabadi-Pour, H. (2021) An efficient Pareto-based feature selection algorithm for multi-label classification. *Information Sciences*, **581**, 428–447.

[6] Hashemi, A., Dowlatshahi, M.B. and Nezamabadi-Pour, H. (2021) VMFS: A VIKOR-based multi-target feature selection. *Expert Systems with Applications*, 115224.

[7] Hashemi, A., Dowlatshahi, M.B. and Nezamabadi-pour, H. (2021) Ensemble of feature selection algorithms: a multi-criteria decision-making approach. *International Journal of Machine Learning and Cybernetics*, 1–21.

[8] Hashemi, A., Dowlatshahi, M.B. and Nezamabadi-pour, H. (2020) MFS-MCDM: Multi-label feature selection using multi-criteria decision making. *Knowledge-Based Systems*,**206**, 106365.

[9] Hashemi, A., Dowlatshahi, M.B. and Nezamabadi-pour, H. (2020) MGFS: A multi-label graph-based feature selection algorithm via PageRank centrality. *Expert Systems with Applications*, **142**, 113024.

[10] Hu, X., Zhou, P., Li, P., Wang, J. and Wu, X. (2018) A survey on online feature selection with streaming features. *Frontiers of Computer Science*, **12**, 479–493.

[11] Jialei Wang, Peilin Zhao, Hoi, S.C.H. and Rong Jin. (2014) Online Feature Selection and Its Applications. *IEEE Transactions on Knowledge and Data Engineering*, **26**, 698–710.

[12] Rahmaninia, M. and Moradi, P. (2018) OSFSMI: Online stream feature selection method based on mutual information. *Applied Soft Computing*, **68**, 733–746.

[13] Tiwari, S.R. and Rana, K.K. (2021) Feature Selection in Big Data: Trends and Challenges. *Data Science and Intelligent Applications*, **52**, 83–98.

[14] Yu, K., Wu, X., Ding, W. and Pei, J. (2014) Towards Scalable and Accurate Online Feature Selection for Big Data. *IEEE International Conference on Data Mining*, 660-669.

[15] Zhou, J., P. Foster, D., A. Stine, R. and H. Ungar, L. (2006) Streamwise feature selection. *Journal of Machine Learning Research*, **3**, 1532–4435.

[16] Zhou, P., Hu, X., Li, P. and Wu, X. (2019) OFS-Density: A novel online streaming feature selection method. *Pattern Recognition*, **86**, 48–61.

[17] Zhou, P., Hu, X., Li, P. and Wu, X. (2017) Online feature selection for high-dimensional class-imbalanced data. *Knowledge-Based Systems*, **136**, 187–199.

[18] Zhou, P., Hu, X., Li, P. and Wu, X. (2019) Online streaming feature selection using adapted Neighborhood Rough Set. *Information Sciences*, **481**, 258–279.

[19] Zhou, P., Li, P., Zhao, S. and Wu, X. (2021) Feature Interaction for Streaming Feature Selection. *IEEE Transactions on Neural Networks and Learning Systems*, **32**, 4691–4702.

[20] Zhou, P., Li, P., Zhao, S. and Zhang, Y. (2021) Online early terminated streaming feature selection based on Rough Set theory. *Applied Soft Computing*, **113**, 107993.