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

مجله‌ی علمی پژوهش در علوم رایانه (سال چهارم)
شماره ۱۵ / پاییز ۱۳۹۸ / ص ۱۵ - ۱

انتخاب ویژگی‌های تأثیرگذار در عملیات دسته‌بندی با استفاده از الگوریتم جستجوی فاخته (CSA)

محمدرضا فرهمند[۱]، الهام سادات حجازی[۲]

[۱] استادیار، هیئت علمی، دانشگاه آزاد اسلامی، واحد ابرکوه.

[۲] کارشناس ارشد، گروه کامپیوتر، دانشگاه آزاد اسلامی، واحد یزد.

نام نویسنده مسئول:
الهام سادات حجازی

چکیده

مساله انتخاب زیرمجموعه ویژگی‌ها، به مفهوم شناسایی و انتخاب یک زیرمجموعه مفید از ویژگی‌ها، از میان مجموعه داده اولیه می‌باشد و همچنین مبحث مهمی در تحلیل میزان همبستگی در زمینه‌های دسته‌بندی می‌باشد که در کاهش ابعاد مجموعه ویژگی‌ها به کار می‌آید. این کار با حذف ویژگی‌هایی که ایجاد نویز می‌کنند و یا اینکه با دیگر ویژگی‌ها همبستگی کمی دارند، انجام می‌شود. در بسیاری از مجموعه داده‌ها برخی از ویژگی‌ها در تصمیم‌گیری نقشی ندارند و به‌نوعی می‌توان آن‌ها را اضافی تلقی نمود. پس انتخاب یک زیرمجموعه‌ی مناسب از ورودی‌ها می‌تواند هم در دقت طبقه‌بندی و هم در سرعت آن تاثیر داشته باشد در این مطالعه، با استفاده از الگوریتم جستجوی فاخته و دو معیار درجه تفکیک پذیری واطلاعات متقابل برای محاسبه همبستگی در میان مشخصه ها، رویکرد جدیدی جهت انتخاب ویژگی های تاثیر گذار وبهینه ، ارائه شده است. ویژگی های به دست آمده با سه دسته بندی کننده ماشین بردار پشتیبان، درخت تصمیم و **K** نزدیک‌ترین همسایه مورد ارزیابی قرارگرفته اند. سپس با مقایسه روش پیشنهاد شده با نتایج به دست آمده از کل مجموعه ویژگی، درجه **F** و روش های انتخاب ویژگی مبتنی بر همبستگی، نشان داده ایم که روش پیشنهاد شده به صورت کلی بسیار کارآمد می باشد. در انتها ازآنجایی که ماشین بردار پشتیبان جهت دسته بندی، نتایج بهتری را ارائه داده است، با دیگر روش های انتخاب ویژگی به علاوه بردار پشتیبان نیز مقایسه انجام گرفته است که راهکار پیشنهادی نسبت به دیگر روش ها نتایج بهتری را حاصل نموده است. از آنجایی که روش پیشنهادی یک راهکار انتخاب ویژگی را معرفی می‌کند می‌توان از آن در سایر زمینه‌های تحقیقاتی نظیر تشخیص پزشکی، در سیستم های تعیین خسارت و یا سیستم های تولیدی استفاده نمود.

واژگان کلیدی: انتخاب ویژگی، الگوریتم جستجوی فاخته، درجه تفکیک پذیری، اطلاعات متقابل با همبستگی دسته‌بندی.

مجلـه‌ی علمی پژوهش در علوم رایانه ، شماره ۱۵، پاییز ۱۳۹۸، ص ۱۵- ۱

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

مقدمه

داده کاوی، به دلیل حجم انبوه داده های روزانه و ضرورت تبدیل این داده ها به اطلاعات مفید، سریع ترین زیرحوزه رو به رشد فناوری اطلاعات محسوب می شود [۱]. داده کاوی شامل پیش پردازش چندگانه (ادغام، فیلتر، تبدیل، کاهش، و غیره)، ارائه دانش و همچنین ارزیابی الگو می گردد [۲]. یکی از مراحل پیش پردازش اصلی، انتخاب ویژگی نام دارد که هدف آن پاک کردن ویژگی های غیر مرتبط و غیرفعال مجموعه داده های خاص است [۳]. وجود مقادیر کلان داده های موجود در مسائل واقعی در جهان وحضور داده های نامرتبط و حشو روند تحلیل آنها را چالش برانگیز و اغلب نادرست می سازد[۴]. انتخاب ویژگی (FS)یک گام پیش پردازش است که ویژگی های نامرتبط و زائد را حذف می کند و مجموعه نهایی از مفیدترین ویژگی ها را پیدا می کند که به نوبه خود به عملکرد بهتر در تکنیک های داده کاوی (به عنوان مثال طبقه بندی) منجر می شود[۵]. ویژگی های نامرتبط و زائد، جستجوی بیشتری در بر می گیرند؛ چرا که آنها الگوهایی با قابلیت تشخیص کم و قوانین لازم برای پیش بینی یا طبقه بندی نسبتا آشکار و نیز ریسک بیش برازش بالا می سازند. انتخاب زیر مجموعه های ویژگی نیاز به تعیین ویژگی مناسب برای به حداکثر رساندن دقت پیش بینی یا طبقه بندی دارد. در واقع، طبقه بندی سریع و دقیق، با استفاده از حداقل تعداد ویژگی ها انتخاب می شود. این ظاهرا از طریق انتخاب ویژگی به دست می آید[۶].

به طور گسترده طبقه بندی داده ها در وظایف مختلف و زمینه های چند رسانه ای مانند طبقه بندی تصویر [۷]، [۸]، طبقه بندی متن [۹]، طبقه بندی عواطف [۱۰]، تقسیم بندی [۱۱]، تشخیص شی [۱۲]، [۱۳]، [۱۴]، [۱۵]، تشخیص رویداد [۱۶]، تشخیص برجستگی[۱۷]، [۱۸]، ردیابی بصری [۱۹] و غیره کاربرد دارد. با این حال، با رشد انفجاری داده ها، و اینکه چنین داده هایی همیشه ساختار داده پیچیده دارند (به عنوان مثال چند کلاسی، چندین دیدگاه و بُعد بالا)، طبقه بندی داده ها هنوز هم یک کار بسیار پیچیده است[۲۰].

در این تحقیق روشی خودکار و سریع جهت انتخاب تأثیرگذارترین ویژگی‌های ارائه می‌گردد که در آن با استفاده از الگوهای آزمایشی و استفاده از الگوریتم‌های مختلف، توانایی هر ویژگی به‌تنهایی و در کنار سایر ویژگی‌ها ارزیابی و نهایتاً ویژگی‌هایی که بهترین میزان تابع ارزیابی عملکرد را از خود نشان دهند انتخاب می‌گردند.

انتخاب ویژگی

روش انتخاب ویژگی برای حذف اصطلاحات بی اهمیت و کاهش ابعاد کلان مجموعه ویژگی ها و بهینه سازی کارایی و اثربخشی طبقه بندی استفاده می شود[۲۱]. مسأله انتخاب ویژگی، یکی از مسائلی است که در مبحث یادگیری ماشین و همچنین شناسایی آماری الگو مطرح است. یافتن مجموعه ای بهینه از ویژگی ها چالش برانگیز و از نظر محاسباتی هزینه براست. به تازگی، الگوریتم‌های فراابتکاری به عنوان ابزارهایی موثر و قابل اطمینان برای حل مسائل بهینه سازی (به عنوان مثال، یادگیری ماشین، مساله داده کاوی، طراحی مهندسی و انتخاب ویژگی)معرفی شدند [۳]،[۲۲]. این مسئله در بسیاری از کاربردها (مانند طبقه‌بندی) اهمیت به سزایی دارد، زیرا در این کاربردها تعداد زیادی ویژگی وجود دارد، که بسیاری از آن‌ها یا بلااستفاده هستند و یا اینکه بار اطلاعاتی چندانی ندارند. حذف نکردن این ویژگی‌ها مشکلی ازلحاظ اطلاعاتی ایجاد نمی‌کند ولی بار محاسباتی را برای کاربرد موردنظر بالا می‌برد و علاوه بر این باعث می‌شود که اطلاعات غیرمفید زیادی را به همراه داده‌های مفید ذخیره کنیم. برای مسئله انتخاب ویژگی، راه‌حل‌ها و الگوریتم‌های فراوانی ارائه‌شده است که بعضی از آن‌ها قدمت سی یا چهل‌ساله دارند. مشکل بعضی از الگوریتم‌ها درزمانی که ارائه‌شده بودند، بار محاسباتی زیاد آن‌ها بود، اگرچه امروزه با ظهور کامپیوترهای سریع و منابع ذخیره‌سازی بزرگ این مشکل، به چشم نمی‌آید ولی از طرف دیگر، مجموعه‌های داده‌ای بسیار بزرگ برای مسائل جدید باعث شده است که همچنان پیدا کردن یک الگوریتم سریع برای این کار مهم باشد. انتخاب ویژگی (حذف متغیر) در درک اطلاعات، کاهش نیاز به محاسبه، کاهش اثر مشقت بعدچندی و بهبود عملکرد پیش بینی کننده همواره موثر بوده است [۲۳].

انتخاب زیر مجموعه های ویژگی به تعیین ویژگی مناسب برای افزایش دقت پیش بینی یا طبقه بندی نیاز دارد. هدف اصلی تحقیقات موجود، تعیین زیر مجموعه ویژگی بهینه است. انتخاب ویژگی ها معمولا بر اساس پارامترهای زمان محاسباتی و کیفیت راه حل های زیر مجموعه ویژگی های تولید شده انجام می شود. در واقع، طبقه بندی سریع و دقیق، با استفاده از حداقل تعداد ویژگی ها انتخاب می شود. بدیهی است که این امراز طریق انتخاب ویژگی ها به دست می آید [۲۴].

روش انتخاب ویژگی (FS)، شرایط را با بالاترین نمره بر اساس معیار پیش تعیین شده اهمیت شرایط حفظ می کند. به طور گسترده مشاهده شد که انتخاب ویژگی می تواند ابزاری قدرتمند برای ساده سازی و یا تسریع محاسبات باشد، و درصورت استفاده مناسب و صحیح، اتلاف و زیان کمتری در کیفیت طبقه بندی به همراه دارد. هدف اصلی روش انتخاب ویژگی، بهبود کارایی طبقه بندی و کارایی محاسباتی است. به رغم رویکردهای متعدد در ادبیات، انتخاب ویژگی هنوزهم در زمره موضوعات پژوهشی در حال پیشرفت قرار دارد. محققان همچنان

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

به دنبال تکنیک های جدید برای انتخاب ویژگی های متمایز هستند تا دقت طبقه بندی را بهبود ببخشند و زمان پردازش نیز کاهش یابد. تعداد زیادی از تکنیک های مبتنی بر فیلتر و تکنیک های بسته بند برای انتخاب ویژگی های متمایز در طبقه بندی متن وجود دارد. روش های انتخاب ویژگی خودکار(FS) عبارتند از حذف عبارات غیرآموزنده و کم اهمیت بر اساس آمارمجموع[۲۱]،[۲۵].

روش‌های انتخاب ویژگی

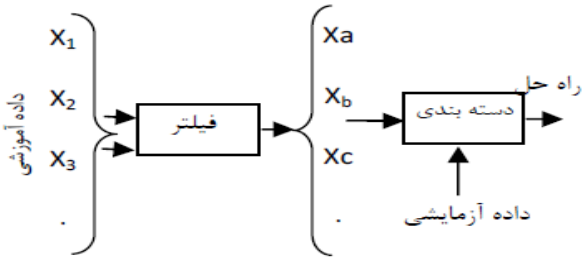
FS را می توان نوعی مساله بهینه سازی در نظر گرفت. با توجه به دشواری این مساله و داشتن تعداد زیادی از راه حل های محلی، الگوریتم های بهینه سازی تصادفی، روشهایی امیدوار کننده برای حل این مشکل هستند. هدف FS یافتن زیر مجموعه ای از ویژگی های M از مجموعه کامل با ویژگی های N است، این کار باعث بهبود عملکرد الگوریتم یادگیری (سرعت یادگیری و یا دقت طبقه بندی) می شود. FS می تواند به عنوان مساله بهینه سازی در نظر گرفته شود؛ چون زیر مجموعه (نزدیک) مطلوب را جستجو می کند[۵].

الگوریتم های انتخاب ویژگی به دو دسته اصلی یعنی الگوریتم های مبتنی بر بسته بند و الگوریتم های مبتنی بر فیلتر تقسیم می شوند[۲۶]. الگوریتم های مبتنی بر بسته بند به استفاده از الگوریتم های یادگیری ماشین برای ارزیابی آنها وابستگی دارد. از سوی دیگر، الگوریتم های مبتنی بر فیلتر از روش های آماری برای انتخاب زیر مجموعه ویژگی استفاده می کنند. با این حال، الگوریتم های مبتنی بر بسته بند نتایج بهتری کسب می کنند؛ و از لحاظ محاسباتی گران هستند. بنابراین، برای کاهش زمان محاسبات الگوریتم جستجوی هوشمند مورد نیاز است [۲۷] [۲۸].

روش های انتخاب ویژگی سنتی را می توان به روش فیلتر، بسته بند، و هیبرید تقسیم بندی کرد. روش هیبرید، همانطور که از نامش پیداست، ویژگی های هر دو روش فیلتر و بسته بند را ترکیب می کند. هدی و همکاران (۲۰۱۱) امتیاز رتبه بندی ویژگی فیلتر را در گام بسته بند گنجاندند تا روند جستجو را تسریع بخشند. ژو و همکاران (۲۰۰۷) روش رتبه بندی فیلتر را در الگوریتم تکاملی ممتیک ادغام کردند، این روش جستجو و شناسایی زیر مجموعه ویژگیهای اصلی را تسریع می کند [۲۹]. روشهای حذف متغیر به طور کلی به روش های فیلتر(Filter) و بسته‌بندها (Wrapper) طبقه بندی شده اند. روشهای فیلتر به عنوان پیش پردازش برای رتبه بندی ویژگی هایی عمل می کنند که در آن ویژگی های عالی رتبه انتخاب و به پیش بینی کننده اعمال می شوند. در روش بسته بندها، معیار انتخاب ویژگی عملکرد پیش بینی کننده است یعنی پیش بینی کننده بر روی الگوریتم جستجو پیچیده می شود تا زیر مجموعه ای را پیدا کند که بالاترین عملکرد پیش بینی کننده را نشان می دهد. روش های تعبیه شده شامل انتخاب متغیر به عنوان بخشی از روند آموزشی بدون تقسیم داده ها به مجموعه های آموزش و آزمون است [۲۳].

روش فیلتر

در این روش، از هیچ تابع دسته‌بندی استفاده نمی‌شود. به‌عبارت‌دیگر از هیچ فیدبکی از الگوریتم یادگیری اعمال‌شده استفاده نخواهد شد. این‌یک روش از پیش انتخاب‌شده‌ای است که مستقل از الگوریتم یادگیری ماشین اعمال‌شده می‌باشد. زیرمجموعه‌های ویژگی به‌وسیله مفاهیم دیگری ارزیابی می‌شوند. در الگوریتم Focus نوعی از روش‌های فیلتر (یک جستجوی جامعی برای آزمایش همه زیرمجموعه‌های ویژگی انجام می‌شود) و سپس این روش، زیرمجموعه‌ای با مینیمم تعداد ویژگی‌هایی را مشخص می‌کند که می‌توانند نمونه‌های مجموعه آموزشی را با دقت قابل قبولی دسته‌بندی کنند روش. Relief [۳۰] یک روش جستجوی تصادفی بر اساس مدل روش فیلتر می‌باشد که در این روش، به هر ویژگی بر اساس ، مرتبط بودن با مفهوم هدف، یک وزن نسبت داده می‌شود و نمونه‌ها به‌طور تصادفی برای پیدا کردن ویژگی‌های مرتبط انتخاب می‌شوند. روش فیلتر به‌صورت زیر کار می‌کند: در مرحله اول، برای هر ویژگی یک رتبه اعتبار محاسبه می‌شود. سپس این رتبه‌ها مرتب‌شده و ویژگی‌هایی با کمترین رتبه حذف می‌شوند. (از یک آستانه برای رتبه‌های ویژگی استفاده می‌شود). سپس نتایج زیرمجموعه‌ای از ویژگی‌ها به‌عنوان ورودی به یک سیستم دسته‌بند داده می‌شود. شکل (۱) طرز کار روش فیلتر را نشان می‌دهد.

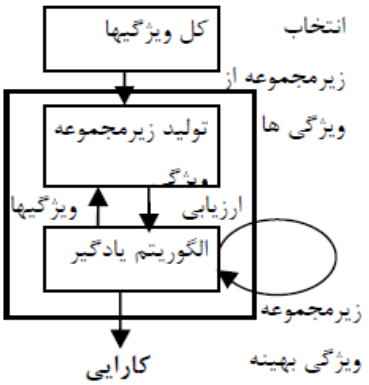


شکل ۱: طرز کار روش فیلتر

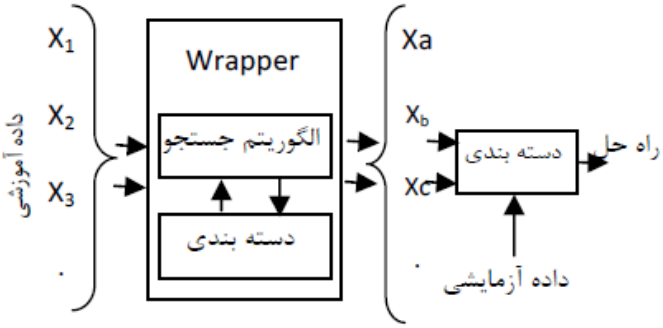
ISSN: 2745-3673
http://www.Rcsj.ir

روش Wrapper

این روش به جعبه سیاه معروف است. در این روش از یک تابع دسته‌بندی برای ارزیابی شایستگی زیرمجموعه‌های ویژگی استفاده می‌شود. این روش از فیدبکی از الگوریتم یادگیری اعمال‌شده استفاده می‌کند. از یک الگوریتم ژنتیک هم برای جستجوی ویژگی‌های معتبر استفاده‌شده است. دلیل استفاده از الگوریتم ژنتیک این است که این الگوریتم می‌تواند یک جستجوی تصادفی را انجام دهد و مستعد گیر افتادن در مینیمم محلی نمی‌باشد. به‌هرحال عملگر بازترکیبی استفاده‌شده در این الگوریتم، مفهوم راه‌حل‌های ترکیبی را دارد، درحالی‌که انتخاب‌های موفقیت‌آمیز ویژگی قبلی را حفظ می‌کند. به‌عبارت‌دیگر این روش، یک روش بازخورد می‌باشد که از الگوریتم یادگیری ماشین در فرآیند انتخاب ویژگی استفاده می‌کند. ارزیابی به‌وسیله اجرای الگوریتم استقرایی در طول فازهای یادگیری و آزمون در هر انتخاب ویژگی، انجام می‌شود. در این مقاله از شبکه عصبی برای این کار استفاده‌شده است. شکل (۲) انتخاب ویژگی را با استفاده از فرآیند Wrapper و شکل (۳) طرز کار این فرآیند را نشان می‌دهد.



شکل ۲: انتخاب ویژگی با استفاده از فرآیند Wrapper



شکل۳: طرز کار روش Wrapper

الگوریتم جستجوی فاخته

روش جستجوی فاخته (CS) یک روش بهینه‌سازی فرا اکتشافی است که رویکردی تکاملی در جستجوی راه‌حل بهینه دارد و در سال ۲۰۰۹ توسط Yang و Deb پیشنهادشده است. این روش از رفتار جالب‌توجه گونه‌هایی از پرنده‌ی فاخته در پرورش تخم الهام گرفته است و آن را با پرواز لووی که نوعی گشت تصادفی است ترکیب می‌کند. برخی از گونه‌های فاخته به‌جای ساختن لانه، تخم‌های خود را در لانه‌ی پرنده‌ای از گونه‌های دیگر می‌گذارند و آن‌ها را با تقلید از شکل تخم‌ها و جوجه‌های پرنده‌ی میزبان وادار به مشارکت در بقای نسل خود می‌کنند.

جستجوی فاخته یک الگوریتم فرا ابتکاری که تقلیدی از استراتژی تولید مثل هوشمندانه کوکو است. این پرنده به دقت لانه میزبان را از دیگر پرندگان انتخاب می‌کند و تخم خود را در میان تخم‌های موجود می‌گذارد. پرنده میزبان اشتباهی از تخم کوکو مراقبت می‌کند. با این حال برخی ممکن است تخم را تشخیص دهند و آن را از بین ببرند یا به خارج از لانه بفرستند. از رفتار هوشمندانه این پرنده برای توسعه یک الگوریتم بهینه‌سازی جدید تقلید شده است. این الگوریتم شامل مجموعه‌ای از لانه‌های میزبان با یک تخم است (راه‌حل) که هر

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

کدام یک نسل را ایجاد می‌کند. بهترین راه‌حل به دنبال تولید راه‌حلی جدید بر اساس یکی از راه‌حل‌های موجود و تغییر مشخصه‌های خاصی است.

همانند سایر الگوریتم‌های تکاملی، این روش هم با یک جمعیت اولیه کار خود را شروع می‌کند (جمعیتی متشکل از فاخته‌ها). این جمعیت از فاخته‌ها تعدادی تخم دارند که آن‌ها را در لانه تعدادی پرنده‌ی میزبان خواهند گذاشت. تعدادی از این تخم‌ها که شباهت بیشتری به تخم‌های پرنده میزبان دارند شانس بیشتری برای رشد و تبدیل شدن به فاخته بالغ خواهند داشت. سایر تخم‌ها توسط پرنده میزبان شناسایی شده و از بین می‌روند. میزان تخم‌های رشد کرده، مناسب بودن لانه‌های آن منطقه را نشان می‌دهند. هرچه تخم‌های بیشتری در یک ناحیه قادر به زیست باشند و نجات یابند به همان اندازه سود بیشتری به آن منطقه اختصاص می‌یابد. بنابراین موقعیتی که در آن بیشترین تعداد تخم‌ها نجات یابند پارامتری خواهد بود که الگوریتم فاخته قصد بهینه‌سازی آن را دارد.

روش پیشنهادی

کاهش بُعد می‌تواند با دو رویکرد به دست آید: استخراج ویژگی و انتخاب ویژگی. استخراج ویژگی بیانگر تبدیل خطی یا غیرخطی از فضای اصلی ویژگی به یک فضای جدید با ابعاد کمتر می‌باشد. از سوی دیگر، انتخاب ویژگی، زیرمجموعه‌ای از ویژگی را به‌واسطه انتخاب ویژگی‌های مهم از نمونه‌های اصلی بدون هیچ‌گونه تغییری، تولید می‌کند. به صورت کلی، یک زیرمجموعه ویژگی خوب می‌بایست دارای خصوصیات زیر باشد. اولاً، ویژگی‌های انتخاب شده به خوبی می‌تواند ارائه‌دهنده الگوهای ورودی باشد به طوری که ویژگی‌های نامربوط انتخاب نشده، تنها فضای جستجو را بزرگ‌تر کنند. دوما، ویژگی‌های انتخاب شده دربرگیرنده همه اطلاعات استفاده شده برای تمایز گذاری الگوها با چندین برچسب دسته داده می‌باشد. چنین زیرمجموعه ویژگی‌ای می‌تواند به عملکرد رضایت‌بخش و مورد انتظار دسته‌بندی کننده فارغ از اینکه از چه الگوریتم آموزشی استفاده شده است، دست یابد. سوما، حذف ویژگی‌های زائد می‌تواند اندازه مجموعه داده‌ها را کاهش دهد. در صورتی که ویژگی‌های نامربوط و زاید به صورتی نامناسب انتخاب شوند ، فرآیند یادگیری را تحت‌الشعاع قرار داده و باعث می‌شود تا فرآیند آموزش ناکارآمد گردد.

پیدا کردن زیرمجموعه ویژگی بهینه F_s از مجموعه بزرگ ویژگی F که منجر به رشد نمایی[1] (تعریفی) فضای جستجو (2^d، که d در اینجا دلالت بر اندازه F دارد) می‌شود که از نظر محاسباتی جذاب نمی‌باشد. بسیاری از روش‌های جستجوی غیرمستدل در پی پیدا کردن راه‌حل‌های نزدیک به بهینه بوده‌اند. از آنجایی که الگوریتم فاخته یکی از ابزارهای بهینه‌سازی می‌باشد که برای پیدا کردن راه‌حل‌ها در فضاهای جستجوی پیچیده و غیرخطی به گستردگی مورد استفاده قرار گرفته، به صورت طبیعی برای حل مسائل انتخاب ویژگی نیز بکار گرفته شده است.

الگوریتم بهینه‌سازی فاخته (COA) یک الگوریتم جدید جمعیت محور است که از روش زندگی گونه‌ای پرنده به نام فاخته الهام گرفته است. در این تحقیق ما رویکردی جدید را بر اساس COA برای انتخاب زیرمجموعه ویژگی معرفی می‌کنیم. برای بررسی کارایی الگوریتم، آزمایشات بر روی چند مجموعه داده انجام شده است و نتایج نشان می‌دهد که روش پیشنهادی می‌تواند راه‌حلی بهینه را برای مساله انتخاب زیرمجموعه ویژگی فراهم آورد.

چارچوب روش پیشنهادی

ما در این بخش یک سیستم هوشمند جدید را برای انتخاب ویژگی ارائه می‌دهیم. این سیستم بر روی انتخاب مبتنی بر الگوریتم جستجوی فاخته استوار است که برای انتخاب یک زیرمجموعه مناسب ویژگی ها، از یک معیار استقلال و تفکیک‌پذیری و یک معیار همبستگی مبتنی بر اطلاعات متقابل برای کاهش بعد داده‌ها و کسب دقت بالای دسته بندی کننده، تمرکز نموده است. از آنجایی که نسبت ماتریس پراکندگی بین دسته‌ای به پراکندگی دورن دسته‌ای زیرمجموعه‌ها، می‌تواند مشارکت آن‌ها در دسته‌بندی را ارائه دهد، زیرمجموعه ویژگی بهینه بر حسب نمره مستقل بودن دسته انتخاب می‌گردد. بعلاوه، متغیرهای زائد نیز در فرآیند انتخاب ویژگی مدنظر قرار گرفته‌اند. ماتریس پراکندگی میان دسته ای بر رابطه بین متغیرها بر حسب برچسب‌های دسته‌شان، دلالت می‌کند و هرچه که این معیار بزرگ‌تر باشد و در سوی دیگر افزایش همبستگی ویژگی ها با کلاس داده و کاهش افزونگی بین ویژگی های حاصل ، نشان دهنده اینکه رو متغیر کمتر تکراری می باشند.شکل (۴) زیر مراحل روش پیشنهادی را به خوبی نشان می‌دهد.

[1] Exponential Growth

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

شکل ۴: فلوچارت کلی روش پیشنهادی

رویکرد پیشنهاد شده با داده‌های استاندارد تائید شده و راه‌حل معرفی شده برای آموزش سه دسته‌بندی کننده مورد استفاده قرار گرفته‌اند، این دسته‌بندی کننده‌ها عبارتند از: ماشین بردار پشتیبانی (SVM)، درخت تصمیم (Dt)و K نزدیک‌ترین همسایه (KNN). نتایج ما همچنین به ترتیب با نتایج کسب شده توسط کل مجموعه ویژگی، نمره F[۲] و انتخاب ویژگی مبتنی بر همبستگی[۳](CFS)، مورد مقایسه قرار گرفته‌اند.

رویکرد پیشنهادی بر اساس الگوریتم جستجوی فاخته

در این بخش، الگوریتم پیشنهادی برای انتخاب ویژگی ارائه شده است. گام‌های الگوریتم پیشنهادی با جزئیات در ادامه ارائه شده است.

ایجاد سکونت‌گاه اولیه فاخته

در الگوریتم ژنتیک و بهینه‌سازی ازدحام ذرات، هر راه‌حل از مسئله را به ترتیب کرموزوم و مکان ذره نامیدیم. اما در COA آن را سکونت‌گاه[۴] می‌نامیم. در یک مساله N بعدی، یک سکونت‌گاه آرایه‌ای N × 1 است که نشان‌دهنده محل فعلی زندگی فاخته است. این آرایه به‌صورت زیر است.

$$habitat = (x_1, x_2, x_3, \dots, x_n)$$

در رویکرد پیشنهادی برای انتخاب ویژگی، هر سکونت‌گاه، یک رشته‌ای از اعداد دودویی است. هنگامی که مقدار متغیر ۱ است، پس این ویژگی انتخاب می‌شود و هرگاه ۰ است ویژگی متناظرش انتخاب نمی‌شود. شکل (۵) نمایش ویژگی را به عنوان یک سکونت‌گاه در رویکرد پیشنهادی را نشان می‌دهد.

F_1	F_2	F_3	...	F_{n-1}	F_{n-2}

زیستگاه

0	1	0	...	1	0

زیرمجموعه ویژگی ها: $\{F_2,\dots,F_{n-1}\}$

شکل۵: مثالی از بیان ویژگی‌ها در روش پیشنهادی

[2] F-score
[3] Correlation-based feature selection
[4] habitat

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

سود و یا شایستگی یک سکونت‌گاه تحت یک تابع شایستگی که در ادامه بیان می‌شود، تعریف می‌شود. بیشتر طبقه‌بندی کننده‌ها می‌توانند برای محاسبه بهره استفاده شوند. برای مثال K نزدیک‌ترین همسایه (KNN)، درخت تصمیم (DT) و ماشین‌های پشتیبان بردار (SVM) سه طبقه بند محبوب هستند

الگوریتم با N_{pop} سکونت‌گاه اولیه تصادفی در اندازه جمعیت آغاز می‌شود. سکونت‌گاه واقعی فاخته جایی است که آن‌ها با بیشترین فاصله از سکونت‌گاهشان تخم‌گذاری می‌کنند. این ماکزیمم محدوده را شعاع تخم‌گذاری[۵] (ELR) گویند که به‌صورت زیر تعریف می‌شود. شعاع پوششی تخم به‌صورت زیر می‌باشد:

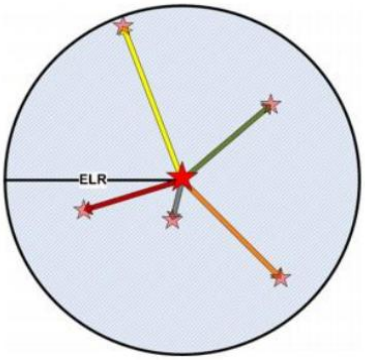
رابطه (۱)

$$ELR=\alpha \times \frac{\text{تعداد تخم های فاخته}}{\text{تعداد نهایی تخم ها}} \times (Var_{hi} - Var_{low})$$

در اینجا α یک عدد صحیح بوده که برای کنترل حداکثر اندازه ELR در نظر گرفته می‌شود. Var_{low} و Var_{hi} به ترتیب کران پایین و بالای تعداد تخم‌های هر فاخته می‌باشد. α عددی صحیح است، var_{hi} و var_{low} کران بالا و پایین به ترتیب هستند. مطابق با معادله بالا، ELR متناسب با تمام تخم‌ها، تعداد تخم‌های فعلی و محدودیت‌های متغیر است.

تخم‌گذاری فاخته

هر فاخته تخم‌گذاری را بطور تصادفی در دیگر لانه‌های پرندگان میزبان در محدوده ELR شان آغاز می‌کنند. شکل۶ تصویری واضح از این مفهوم را ارائه می‌دهد. ستاره قرمز در وسط شکل، زیستگاه اولیه فاخته‌ای با ۵ تخم است. ستاره‌های دیگر محل هر یک از تخم‌های فاخته را نشان می‌دهد. بعد از تخم‌گذاری، تخم‌ها با کمترین مقدار بهره، شناسایی و از بین برده می‌شوند. دیگر تخم‌ها در لانه‌های میزبان رشد می‌کند و توسط پرندگان میزبان تغذیه می‌شوند. قابل توجه است که فقط یک تخم در هر لانه شانس رشد دارد زیرا فاخته بیشترین مقدار غذا آورده شده توسط پرنده میزبان را می‌خورد.



شکل۶: نمایی از تخم‌گذاری تصادفی در الگوریتم فاخته

مهاجرت فاخته‌ها در روش پیشنهادی

وقتی فاخته‌ها رشد کردند و بالغ شدند در جمعیت خود زندگی می‌کنند. اما در زمان تخم‌گذاری، آن‌ها به جمعیت جدید و بهتری با مشابهت بیشتر تخم‌ها با پرنده میزبان مهاجرت می‌کنند. بعد از اینکه فاخته‌ها در نواحی مختلفی شکل گرفتند، جمعیت با بهترین میزان تابع هدف به عنوان نقطه هدف برای مهاجرت دیگر فاخته‌ها انتخاب می‌شود. بعد از استقرار، تمیز دادن اینکه فاخته متعلق به کدام گروه است سخت می‌باشد. برای حل این مساله، خوشه‌بندی انجام می‌گیرد. بعد از گروه‌بندی فاخته، ماکزیمم میانگین بهره هر گروه گروه هدف را تعیین می‌کند. همان‌طور که قبلاً ذکر شد، فاخته‌ها سکونت‌گاه خود را برای تخم‌گذاری با انتقال همه فاخته‌ها به نقطه هدف بهبود می‌بخشند. نسخه اصلی COA بر روی مسائل پیوسته عمل می‌کند. چون انتخاب ویژگی مساله ای گسسته است، در این تحقیق یک روش مهاجرت جدید که برای مسائل گسسته مناسب است ارائه شده است. این عملگر به‌صورت فرمان های زیر است.

- برای هر زیستگاه تکرار کن
- فاصله بلوک شهری را بین هر زیستگاه و نقطه هدف را محاسبه کن

---

[۵] Egg Laying Radius (ELR)

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

- یک رشته باینری (S) اولیه با طول N به تعداد ویژگی‌ها با مقداردهی اولیه به صفر بساز
- تعدادی از سلول‌های باینری را به ۱ مقداردهی کن
- سلول‌های ۱ موجود در نقطه هدف را به همان سلول‌های زیستگاه‌ها در S کپی کن

از بین بردن فاخته در بدترین سکونت‌گاه‌ها

به دلیل تعادل جمعیت در پرندگان، پارامتری جدید تعریف شده است که ماکزیمم تعداد فاخته‌های زنده در جمعیت را محدود می‌کند. برای مدل‌سازی این محدودیت، N_{max} تعداد فاخته‌های زنده‌ای است که بهره بهتری دارند و دیگر فاخته‌ها می‌میرند.

تابع هدف روش پیشنهادی

در این قسمت برای انتخاب یک زیرمجموعه مناسب ویژگی ها ، از یک معیار استقلال و تفکیک‌پذیری و یک معیار همبستگی مبتنی بر اطلاعات متقابل برای کاهش بعد داده‌ها و کسب دقت بالای دسته بندی کننده، تمرکز نمودهشده است.معیار تفکیک پذیری با ماتریس پراکندگی میان دسته ای بر رابطه بین متغیرها بر حسب برچسب‌های دسته‌شان، دلالت می‌کند و و هرچه که این معیار بزرگ‌تر باشد و در سوی دیگر افزایش همبستگی ویژگی ها با کلاس داده و کاهش افزونگی بین ویژگی های حاصل ، نشان دهنده این است که متغیرها کمتر تکراری می باشند. در بسیاری از رویکرد های انتخاب ویژگی قبلی تنها از یک تابع هدف بر اساس بررسی توزیع داده ها پشتیبانی می گردد که این امر منجر به حذف بسیاری از راه حل های کارآمدی می گرددکه ممکن است بر اساس اهداف کارایی دیگری دارای قدرت باشند. در واقع ارزیابی این زیر مجموعه ها با یک روش انجام می شود که منجر به پس زدن زیر مجموعه هایی با خواص خوب با توجه به معیار دیگری می شود. در این زمینه، راه حل با افزونگی بسیار کم یا همبستگی بسیار بالا که دو هدف کارایی بسیار مهم می باشند، می تواند توسط فرآیند انتخاب رد شود. برای افزایش تنوع زیر مجموعه های انتخاب شده ، در این پایان نامه فضای جستجوی بسیار بزرگ مسئله با یک روش چند منظوره مورد بررسی قرار می گیرد. برای این کار دو هدف کارایی بر اساس درجه همبستگی مبتنی بر پراکندگی داده ها و سپس همبستگی و افزونگی مبتنی بر اطلاعات متقابل تعریف می گردد. در اینجا به بررسی دو معیار کارایی مختلف و سپس تابع هدف روش پیشنهادی برای ارزیابی الگوریتم بهینه‌سازی جستجوی فاخته می‌پردازیم.

معیار اول: درجه تفکیک‌پذیری

از آنجایی که نسبت ماتریس پراکندگی بین دسته‌ای به پراکندگی دورن دسته‌ای زیرمجموعه‌ها، می‌تواند مشارکت آن‌ها در دسته‌بندی را ارائه دهد، زیرمجموعه ویژگی بهینه بر حسب میزان استقلال ویژگی‌ها در فرآیند دسته‌بندی انتخاب می‌گردد. هدف درجه تفکیک‌پذیری[6]، انتخاب ویژگی‌های بهینه برای دسته‌بندی می‌باشد. تصور کنید که $(x, y)\in(R_d \times Y)$ یک نمونه است که در اینجا R_d یک فضای ویژگی d بعدی و $Y=\{1, 2,\ldots, c\}$، مجموعه برچسب دسته می‌باشد. نماد n_i بیانگر تعداد نمونه‌های متعلق به طبقه i ام بوده و N تعداد کلی نمونه‌ها را ارائه می‌دهد. فکر کنید که x_{ij} بر نمونه j ام در دسته i ام دلالت داشته، u میانگین نمونه همه دسته‌ها و u_i میانگین نمونه دسته i ام می‌باشد.

$$u_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$$

رابطه (۲)

$$u = \frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{n_i} X_{ij}$$

رابطه (۳)

پس ماتریس پراکندگی بین دسته‌ای[7] S_b و ماتریس پراکندگی میان دسته‌ای[8] S_w، به‌صورت زیر تعریف می‌شوند:

$$S_b = c\sum_{i=1} n_i (u_i - u)$$

رابطه (۴)

[6] The Separability Score
[7] Between-Class Scatter Matrix
[8] Within-Class Scatter Matrix

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

$$S_w = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (u_i - x_{ij})$$

رابطه (۵)

در اینجا، S_b فواصل میان بردار میانگین هر یک از دسته‌ها و میانگین کلی را اندازه‌گیری کرده درحالی‌که S_w میانگین پراکندگی دسته‌ها در حول بردارهای میانگین‌شان را اندازه‌گیری می‌کند. برای یک زیرمجموعه ویژگی موردنظر، تفکیک‌پذیری دسته مبتنی بر ماتریس پراکندگی[9]، ارزیابی‌کننده نسبت ردیابی[10] یا عامل تعیین‌کننده ماتریس پراکندگی بین دسته‌ای با ماتریس پراکندگی درون دسته‌ای می‌باشد. تفکیک‌پذیری دسته Obj_1 به‌صورت زیر ارائه‌شده است:

$$Obj_1 = \frac{S_b}{S_w}$$

رابطه (۶)

یک زیرمجموعه با Obj_1 بزرگ، به عنوان یک زیرمجموعه خوب در نظر گرفته شده و به معنی پراکندگی درون دسته‌ای کوچک و پراکندگی بین دسته‌ای بزرگ می‌باشد. از این رو، یک Obj_1 بزرگ تضمین می‌کند که دسته‌ها توسط میانگین‌های پراکندگی‌شان به خوبی پراکنده شده‌اند. این معیاری ساده، قدرتمند و یکپارچه برای دسته‌بندی می‌باشد. ایده تفکیک‌پذیری دسته به منظور انتخاب زیرمجموعه بهینه برای دسته‌بندی مورد استفاده قرار گرفته، زیرا مشارکت‌های این زیرمجموعه‌ها برای دسته‌بندی را منعکس می‌کند.

معیار دوم: همبستگی مبتنی بر اطلاعات متقابل

برای افزایش تنوع زیرمجموعه‌های انتخاب شده، در این تحقیق فضای جستجوی بسیار بزرگ مسئله با یک روش چند منظوره موردبررسی قرار می‌گیرد. ارزیابی کیفیت اهداف کارایی بیان شده براساس اطلاعات متقابل[11] صورت می‌پذیرد تا به‌طور جداگانه میزان همبستگی داده‌ها و افزونگی زیرمجموعه ویژگی‌های انتخابی را اندازه‌گیری کند.این دو معیار هر دو کیفیت اختصاصی ویژگی‌های انتخاب شده و کیفیت زیرمجموعه را اندازه‌گیری می‌کند.

در این بخش ما فرموله سازی معیار دیگری را برای انتخاب ویژگی به عنوان تابع هدف الگوریتم بهینه‌سازی جستجوی فاخته و نحوه محاسبه آن‌ها را ارائه می‌دهیم. اطلاعات متقابل یک شاخص خوب برای مطالعه وابستگی بین یک ویژگی و طبقه‌بندی و افزونگی بین ویژگی‌های تصادفی است. X و Y را به عنوان دو متغیر تصادفی با قوانین احتمال گسسته در نظر بگیرید. اطلاعات متقابل دو متغیر X و Y با I(X,Y) نشان داده شده و از طریق P(X) و P(Y)و P(X,Y) به صورت زیر تعریف می‌شود:

$$I(X;Y) = \sum_{y \in \Omega Y} \sum_{x \in \in \Omega} P(x,y).\log \frac{P(x,y)}{(P(x).P(y)}$$

رابطه(۷)

که در آن Ω_X و Ω_Y به ترتیب فضاهای نمونه Xو Yبه ترتیب هستند. هنگامی که دو متغیر Xو Y به یکدیگر وابسته هستند. I(X,Y) زیاد است و در نقطه مقابل هنگامی که Xو Y به‌طور کامل از هم مستقل هستند، I(X,Y) برابر با صفر است. بعد از محاسبه اطلاعات متقابل، ما دو هدف کارایی مختلف را برای انتخاب بهترین زیرمجموعه از بین راه‌حل‌های تولیدی توسط الگوریتم بهینه‌سازی جستجوی فاخته را دنبال می‌کنیم. اولین هدف کیفیت مجموعه ویژگی‌های انتخابی در راستای پیش‌بینی هدف داده‌ها که تحت عنوان کلاس بیان می‌گردد، را ارزیابی می‌کند و دومین هدف به بررسی افزونگی و تکراری بودن مجموعه ویژگی‌ها می‌پردازد. این دو هدف در ادامه بیان شده‌اند.

همبستگی ویژگی‌ها با کلاس

برای هر زیرمجموعه از ویژگی‌ها، ما مفهوم ارتباط را که از طریق وابستگی بیان می‌گردد را تعریف می‌کنیم که در واقع از طریق محاسبه میانگین اطلاعات متقابل بین هر یک از ویژگی‌های موجود در مجموعه ویژگی‌های انتخابی با مجموعه هدف داده‌ها یعنی کلاس داده‌ها که از طریق متغیر C نشان داده می‌شود، به دست می‌آید. نحوه محاسبه این رابطه در زیر نشان داده شده است:

[9] Scatter-Matrix-based class separability
[10] The Ratio of The Trace
[11] Nutual Information (MI)

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

$$D_S = \frac{1}{|S|} \sum_{Xi \in S} I(X_i; c)$$

رابطه (۸)

که در آن I(xi, c) بیان‌کننده مقدار اطلاعات متقابل بین ویژگی iام در زیرمجموعه ویژگی انتخابی کاندید و کلاس داده‌ها می‌باشد و بیان‌کننده این واقعیت است که این ویژگی با چه کیفیتی کلاس داده‌ها را بیان می‌کند.

کاهش افزونگی بین ویژگی‌های انتخابی

در یک راه‌حل انتخابی ممکن است، دو یا بیش از دو ویژگی در هدف اول یعنی ارتباط و همبستگی با کلاس بسیار خوب باشند، اما این ویژگی‌ها به نحوی دارای افزونگی بوده و بتوان با داشتن یکی از آن‌ها به مقادیر دیگر ویژگی‌ها دست یافت. در این شرایط می‌بایست ویژگی‌های انتخابی در راستای کاهش افزونگی و حذف ویژگی‌های تکراری بررسی گردند. در اینجا از اطلاعات متقابل برای ارزیابی افزونگی بین ویژگی‌ها استفاده می‌گردد. در اینجا دو متغیر i و j از یک مجموعه ویژگی‌های کاندید توسط الگوریتم بهینه‌سازی ازدحام ذرات به‌صورت $(Xi; X_j)_{i,j = ۱,\ldots, m_i \neq j}$ نشان داده شده و از طریق رابطه زیر محاسبه می‌گردد:

$$R_S = \frac{1}{|S|^2} \sum_{Xi, X_j \in S} I(X_i; X_j)$$

رابطه(۹)

هدف ما افزایش معیار اول یعنی همبستگی ویژگی‌ها با کلاس و کاهش هدف دوم یعنی افزونگی مابین ویژگی‌هاست. در اینجا از یک تابع هدف به‌صورت رابطه زیر استفاده می‌گردد:

$$Obj_2 = \frac{D_S}{R_S}$$

رابطه (۱۰)

در این تابع هدف همبستگی داده‌ها در صورت کسر و هدف دوم یعنی افزونگی در مخرج کسر قرار می‌گیرد. با بزرگ شدن صورت و کوچک شدن مخرج حاصل کسر بزرگ شده و کیفیت بالاتری را نشان می‌دهد. بنابراین یک تعادل و توازن ما بین این دو هدف کارایی برقرار می‌گردد.

تابع هدف نهایی الگوریتم بهینه‌سازی جستجوی فاخته

در بخش قبل دو هدف کارایی مختلف برای ارزیابی راه‌حل‌های تولید شده توسط هر فاخته در الگوریتم بهینه‌سازی جستجوی فاخته معرفی گردید. در این بخش یک تابع هدف کلی بر اساس این دو تابع هدف کارایی تعریف می‌گردد که از حاصل جمع دو تابع کارایی بیان شده حاصل می‌گردد. این تابع در رابطه زیر نشان داده شده است:

$$Fitness = Obj_1 + Obj_2$$

رابطه (۱۱)

از آنجایی که در هر یک از این فاکتورها، بالاتربودن بیانگر کیفیت بالاتر روش پیشنهادی می‌باشد، بنابراین بالابودن تابع هدف پیشنهادی برای هر راه حل تولید شده توسط الگوریتم بهینه‌سازی جستجوی فاخته بیانگر کارآمدی بالاتر آن می‌باشد.

ارزیابی روش پیشنهادی

روش پیشنهادی بر اساس پارامترهای متعددی مورد ارزیابی قرار گرفته است که در این فصل جزییات شبیه‌سازی و پیاده‌سازی روش مورد نظر و همچنین نتایج حاصل شده مورد بررسی و ارزیابی قرار می‌گیرند.

ویژگی‌های محیط پیاده‌سازی

برای پیاده‌سازی روش پیشنهادی از نرم‌افزار MATLAB R2010a استفاده گردیده است. این شبیه‌سازی بر روی سیستمی کامپیوتری با پردازنده corei7 با میزان حافظ اصلی ۴ گیگابایت و بر روی سیستم عامل ویندوز ۷ اجرا شده است.

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

داده‌های مورد استفاده

داده‌های مربوط به داده‌های حجیم که در این تحقیق مورداستفاده قرارگرفته، مربوط به حمله های شبکه ها که در ۵ دسته قرار دارند و این مجموعه داده ها مربوط به مخزن داده UCI بوده و از آدرس اینترنتی[۱۲] قابل دسترسی می‌باشد. جدول ۱ این داده ها را با مشخصات هریک که بیانگر تعداد رکورد و ویژگی است، توصیف می کند.

جدول ۱: مشخصات مجموعه داده های مورد استفاده

مجموعه داده	تعداد رکوردها	تعداد ویژگی ها	تعداد طبقات داده
KddCup99	۴۹۴۰۲	۴۱	۵

در اینجا داده‌ها به صورت تصادفی بین دو مجموعه، یعنی ۷۰ درصد نمونه‌ها آموزشی و ۳۰ درصد نمونه‌های آزمایشی، تقسیم شده‌اند. ما الگوریتم فاخته را بر روی مجموعه آموزش، بکار گرفته‌ایم. در انتهای هر دور اجرا، بهترین زیرمجموعه‌ای که کشف گردیده را برای آموزش سه طبقه‌بندی کننده مختلف و ارزیابی عملکردشان مورد استفاده قرار می‌گیرند. پارامترهای بهینه الگوریتم فاخته، به صورت غیر مستدل توسط بکارگیری یک مجموعه از آزمایشات مقدماتی، که در جدول (۱) فهرست شده‌اند انتخاب گردیده و برای همه آزمایشات گزارش شده بکار گرفته شده‌اند. مدل‌های طبقه‌بندی کننده بر روی نمونه‌های آموزش ایجاد گردیده‌اند. عملکردهای الگوریتم‌های آموزشی بر روی نمونه-های آموزشی مورد ارزیابی قرار گرفته و بر اساس تکرارهای متفاوت تقسیم داده‌ها، میانگین‌گیری شده‌اند.

ارزیابی کیفیت روش انتخاب ویژگی پیشنهادی

از چندین الگوریتم طبقه‌بندی کننده برای ارزیابی زیرمجموعه ویژگی به دست آمده، مورد استفاده قرار گرفته‌است. عملکرد الگوریتم پیشنهاد شده با طبقه‌بندی کننده های مختلف بر روی دو حالت کل ویژگی‌ها، ویژگی‌های انتخابی توسط الگوریتم جستجوی فاخته بر اساس پراکندگی و همبستگی داده ها در جدول (۱) نشان داده شده است. بهترین دقت‌های طبقه‌بندی با افزایش تعداد دفعات تکرار، نشان داده شده‌اند. به منظور تایید بیشتر کارآمدی و عمومیت روش‌مان، عملکردهای به دست آمده با عملکردهای به دست آمده از مجموعه کل ویژگی‌های مقایسه شده‌اند. دقت‌ها برای زمانی می‌باشند که همه ویژگی‌های و زیرمجموعه ویژگی بهینه فراهم شده‌اند. دراین حالت از ۳ طبقه بندی کننده ماشین بردار پشتیبان، درخت تصمیم و K- نزدیک ترین همسایه استفاده می گردد.

مقایسه با حالت عدم انتخاب ویژگی تاثیرگذار

سه طبقه‌بندی کننده‌ای که در بالا مورد بحث و بررسی قرار گرفت برای ارزیابی زیرمجموعه ویژگی به دست آمده، مورد استفاده قرار گرفته‌اند. عملکرد الگوریتم پیشنهاد شده با سه طبقه‌بندی کننده در جدول (۲) نشان داده شده است.

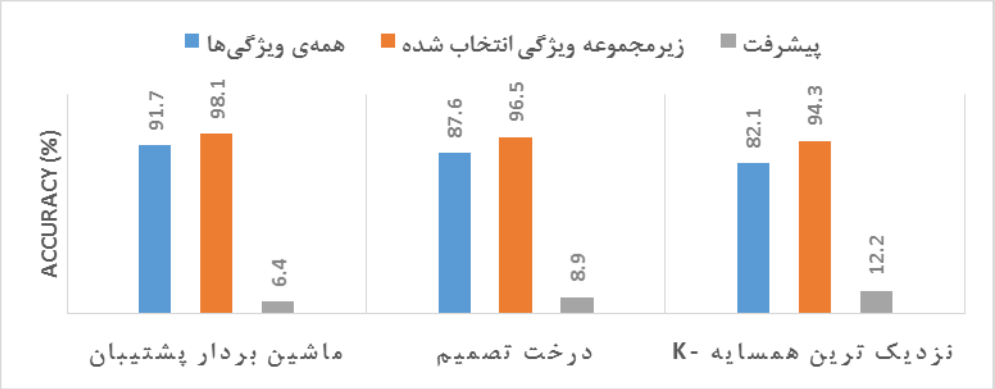
جدول ۲: دقت طبقه‌بندی (درصد) به‌واسطه استفاده از کل ویژگی‌ها و ویژگی‌های بهینه انتخاب شده

داده ها/ طبقه بندی کننده	ماشین بردار پشتیبان	درخت تصمیم	k- نزدیک ترین همسایه
همه‌ی ویژگی‌ها	۹۱,۷	۸۷,۶	۸۲,۱
زیرمجموعه ویژگی انتخاب شده	۹۸,۱	۹۶,۵	۹۴,۳
پیشرفت	۶,۴	۸,۹	۱۲,۲

ذکر این مطلب مناسب است که عملکرد به دست آمده با استفاده از زیرمجموعه ویژگی انتخاب شده در مقایسه با مجموعه کل ویژگی‌ها دارای برتری می‌باشد. به صورت ویژه، ماشین بردار پشتیبان، بالاترین دقت طبقه‌بندی را در مقایسه با درخت تصمیم و K نزدیک‌ترین همسایه دارد. از این رو، ماشین بردار پشتیبان به عنوان طبقه‌بندی کننده انتخاب شده است. ما به این نتیجه رسیدیم که K نزدیک‌ترین همسایه نسبتاً ضعیف تر از درخت تصمیم بر روی داده های مورد استفاده می‌باشد. با مقایسه نتایج کسب شده بر اساس مجموعه کل ویژگی‌ها، عملکرد طبقه‌بندی روش پیشنهادی پیشرفت کرده است. نمودار (۱) این مقایسه را به خوبی نشان می‌دهد.

[12] https://archive.ics.uci.edu/ml/datasets/

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



نمودار(۱): نمودار مقایسه دقت دسته‌بندی (درصد) به‌واسطه استفاده از کل ویژگی‌ها و ویژگی‌های بهینه انتخاب شده

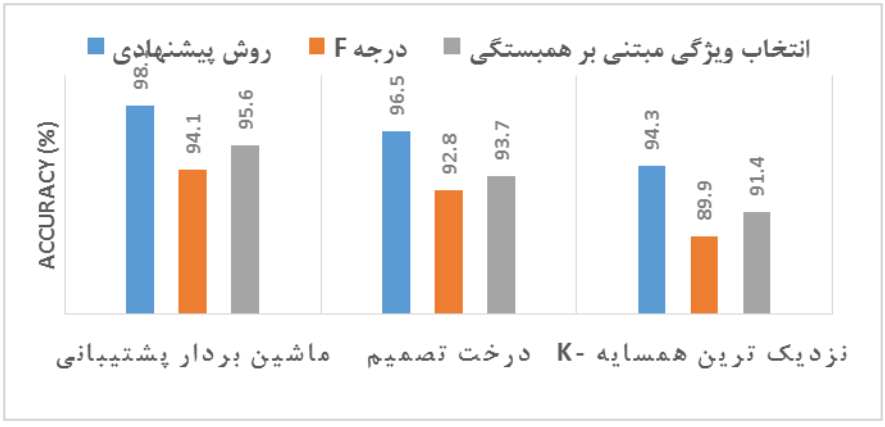
مقایسه با سایر روش های انتخاب ویژگی

به منظور تائید کارآمدی روش پیشنهاد شده انتخاب ویژگی مبتنی بر الگوریتم فاخته، نتایج با نتیجه‌های به دست آمده از دو روش انتخاب ویژگی دیگر که قبلاً نیز آن‌ها را در فصل دوم ذکر کرده‌ایم، مورد مقایسه قرار گرفته‌اند. نتایج طبقه‌بندی در جدول (۳) ارائه شده‌اند.

جدول ۳: مقایسه درصد دقت کسب شده توسط روش پیشنهادی و سایر روش‌ها

روش پیشنهادی	ماشین بردار پشتیبانی	درخت تصمیم	k- نزدیک ترین همسایه
درجه F	۹۸,۱	۹۶,۵	۹۴,۳
انتخاب ویژگی مبتنی بر همبستگی	۹۴,۱	۹۲,۸	۸۹,۹
	۹۵,۶	۹۳,۷	۹۱,۴

همان‌طور که ملاحظه می‌کنید روش ما نتایج بهتری را در مقایسه با دیگر روش‌ها ارائه می‌دهد. روش ما با طبقه‌بندی کننده ماشین بردار پشتیبان، بیشترین پیشرفت عملکرد طبقه‌بندی را در میان این سه روش انتخاب ویژگی دارد. به صورت کلی، روش پیشنهاد شده انتخاب ویژگی بر مبنای الگوریتم فاخته در امتداد ماشین بردار پشتیبان بسیار کارآمد می‌باشد. نمودار (۲) این مقایسه را نشان می‌دهد.



نمودار۲: نمودار مقایسه درصد دقت کسب شده توسط روش پیشنهادی و سایر روش‌های انتخاب ویژگی

همان‌طور که ملاحظه می‌کنید روش ما نتایج بهتری را در مقایسه با دیگر روش‌ها ارائه می‌دهد. روش مبتنی بر همبستگی دارای عملکرد نسبتاً بدتری در مقایسه با روش ما و عملکرد نسبتاً بهتری در مقایسه با روش انتخاب ویژگی مبتنی بر F-Score می‌باشد. روش ما با طبقه‌بندی کننده ترکیبی پیشنهادی، بیشترین پیشرفت عملکرد طبقه‌بندی را در میان این سه روش انتخاب ویژگی دارد. به صورت کلی، روش پیشنهادی برای انتخاب ویژگی بر مبنای الگوریتم جستجوی فاخته در امتداد طبقه‌بندی کننده های مختلف بسیار کارآمد می‌باشد.

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

مقایسه با روش های انتخاب ویژگی ارائه شده در تحقیقات قبلی

روش پیشنهادی را با سایر روش‌های که قبلاً در این زمینه کارکرده‌اند نیز مورد مقایسه قرار می گیرد. این مقایسه بر این دلالت دارد که روش ما دارای بالاترین دقت تشخیص می‌باشد. مقایسه‌ی موجود در جدول (۴) نشان می‌دهد که دقت روش ما مقداری بالاتر از دقت سایر روش‌ها می‌باشد، همچنین لازم به ذکر است که روش ما دارای پیچیدگی نسبتاً کمتری می‌باشد. بنابراین روش ما به صورت کلی بسیار کارآمد می‌باشد.

جدول ۴: مقایسه نتایج

روش کار	دقت طبقه بندی (Accuracy)
تبرید شبیه سازی شده + ماشین بردار پشتیبان [۳۱]	۹۴,۹
بهینه سازی ازدحام ذرات+ ماشین بردار پشتیبان [۳۲]	۹۶,۱
الگوریتم ژنتیک + ماشین بردار پشتیبان [۳۳]	۹۶,۷
روش پیشنهادی	۹۸,۱

نتیجه گیری

انتخاب ویژگی به‌منظور انتخاب زیرمجموعه مناسب از بین ویژگی‌های استخراج‌شده ،انجام می‌شود. انتخاب ویژگی یک مسئله‌ی جستجو در فضای بزرگی از راه‌حل‌ها شامل ترکیبات متفاوتی از ویژگی‌ها می‌باشد و منجر به بهبود محاسبه کار آیی دسته‌بندی، ایجاد دسته‌بندهای سریع و کم‌هزینه می‌گردد. در این تحقیق، یک رویکرد برای کاهش ابعاد ویژگی‌های مجموعه‌های داده‌ای با استفاده از الگوریتم جستجوی فاخته ارائه‌شده است. ازآنجاکه تعداد ویژگی‌ها زیاد است لازم است برای کار آیی بالاتر طبقه‌بندی کننده ویژگی‌های مؤثر انتخاب و بقیه حذف شوند. از الگوریتم فاخته به این منظور استفاده‌شده است. از نکات مهم، انتخاب تابع هدف، ماکزیمم نمودن تابع هدف و نحوه کدگذاری داده‌ها است. نتایج نشان داد که روش انتخاب ویژگی بر اساس الگوریتم فاخته بر مبنای تابع سازگاری نسبت به سایر روش انتخاب ویژگی‌ها برتری قابل‌توجهی دارد. همچنین دسته‌بندی استفاده شده بر مبنای داده‌های استخراجی توسط الگوریتم فاخته، نسبت به سایر روش‌های معرفی شده برای طبقه‌بندی مناسب دارای دقت تشخیص بالاتری می‌باشد که بیانگر کیفیت روش معرفی شده از جنبه‌های مختلف می‌باشد.

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

منابع و مراجع

[1] J. Han, J. Pei, and M. Kamber, Data mining: concepts and techniques. Elsevier, 2011.
[2] H. Liu and H. Motoda, Feature selection for knowledge discovery and data mining, vol. 454. Springer Science & Business Media, 2012.
[3] Al-Tashi, Q., Abdulkadir, S. J., Rais, H. M., Mirjalili, S., & Alhussian, H. (2019). Binary Optimization Using Hybrid Grey Wolf Optimization for Feature Selection. IEEE Access, 1–1.doi:10.1109/access.2019.2906757
[4] Li, Y. , Li, T. , & Liu, H. (2017). Recent advances in feature selection and its applications. Knowledge and Information Systems, 53 (3), 551–577 .
[5] M. Mafarja et al. (2019), Binary grasshopper optimisation algorithm approaches for feature selection problems, Expert Systems With Applications 117 (2019) 267–286, https://doi.org/10.1016/j.eswa.2018.09.015
[6] Ahmed KHARRAT and Mahmoud NEJI, (2019), Feature selection based on hybrid optimization for magnetic resonance imaging brain tumor classification and segmentation, Applied Medical Informatics Original Research Vol. 41, No. 1 /2019, pp:9-23.
[7] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: remote sensing image scene classification via learning discriminative cnns," TGRS, vol. 56, no. 5, pp. 2811–2821, 2018.
[8] G. Cheng, P. Zhou, and J. Han, "Duplex metric learning for image set classification," TIP, vol. 27, no. 1, pp. 281–292, 2018.
[9] B. Tang, S. Kay, and H. He, "Toward optimal feature selection in naive bayes for text categorization," TKDE, vol. 28, no. 9, pp. 2508–2521, 2016.
[10] F.Wu, Z. Yuan, and Y. Huang, "Collaboratively training sentiment classifiers for multiple domains," TKDE, vol. 29, no. 7, pp. 1370–1383 ,2017.
[11] J. Han, R. Quan, D. Zhang, and F. Nie, "Robust object cosegmentation using background prior," TIP, vol. 27, no. 4, pp.1639–1651, 2018
[12] D. Zhang, D. Meng, L. Zhao, and J. Han, "Bridging saliency detection to weakly supervised object detection based on selfpaced curriculum learning," in IJCAI, 2016, pp. 3538–3544.
[13] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu, "Advanced deep-learning techniques for salient and category-specific object detection: a survey," SPM, vol. 35, no. 1, pp. 84–100, 2018.
[14] D. Zhang, J. Han, L. Zhao, and D. Meng, "Leveraging priorknowledge for weakly supervised object detection under a collaborative self-paced curriculum learning framework," IJCV, p. DOI:https://doi.org/10.1007/s1126, 2018.
[15] G. Cheng, J. Han, P. Zhou, and D. Xu, "Learning rotation-invariant and fisher discriminative convolutional neural networks for object detection," TIP, vol. 28, no. 1, pp. 265–278, 2019.
[16] X. Chang, Y.-L. Yu, Y. Yang, and E. P. Xing, "Semantic pooling for complex event analysis in untrimmed videos," TPAMI, vol. 39,no. 8, pp. 1617–1632, 2017.
[17] D. Zhang, D. Meng, and J. Han, "Co-saliency detection via a self-paced multiple-instance learning framework," TPAMI, vol. 39, no. 5, pp. 865–878, 2016.
[18] J. Han, G. Cheng, Z. Li, and D. Zhang, "A unified metric learningbased framework for co-saliency detection," TCSVT, vol. 28, no. 10, pp. 2473–2483, 2018.
[19] S. Zhang, X. Yu, Y. Sui, S. Zhao, and L. Zhang, "Object tracking with multi-view support vector machines," TMM, vol. 17, no. 3, pp. 265–278, 2015.
[20] Xu, J., Han, J., Nie, F., & Li, X. (2019). Multi-view Scaling Support Vector Machines for Classification and Feature Selection. IEEE Transactions on Knowledge and Data Engineering, 1–1. doi:10.1109/tkde.2019.2904256

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

[21] B. S. Harish, M. B. Revanasiddappa, (2017), A Comprehensive Survey on various Feature Selection Methods to Categorize Text Documents, International Journal of Computer Applications (0975 – 8887) Volume 164 - No.8, April 2017.
[22] Al-Tashi et al., (2019), Binary Optimization Using Hybrid Grey Wolf Optimization for Feature Selection, IEEE, DOI 10.1109/ACCESS.2019.2906757, IEEE Access
[23] G. Chandrashekar, F. Sahin, (2014), A survey on feature selection methods Computers and Electrical Engineering 40 (2014) 16–28.
[24] Kharrat, A., & Neji, M. (2018). A Hybrid Feature Selection for MRI Brain Tumor Classification. Innovations in Bio-Inspired Computing and Applications, 329–338. doi:10.1007/978-3-319-76354-5_30
[25] Forman George. An extensive empirical study of feature selection metrics for text classification. Journal of machine learning research, 3(Mar):1289–1305, 2003.
[26] Kohavi R, John G (1997) Wrappers for feature subset selection.A rtif Intell 97(1):273–324.
[27] Guyon I, Elisseeff A (2003) An introduction to variable and attribute selection. Machine Learning Research 3:1157–1182.
[28] Sayed, G. I., Hassanien, A. E., & Azar, A. T. (2017). Feature selection via a novel chaotic crow search algorithm. Neural Computing and Applications. doi:10.1007/s00521-017-2988-6
[29] Melo, A., & Paulheim, H. (2017). Local and global feature selection for multilabel classification with binary relevance. Artificial Intelligence Review.doi:10.1007/s10462-017-9556-4
[30] I. Mejía-Guevara and Ákuri-Morales, "Evolutionary feature and parameter selection in support vector regression," MICAI 2007: Advances inArtificial Intelligence, pp. 399-408, 2007.
[31] S. Lin, Z .Lee, S. Chen, and T. Tseng, "Parameter determination of support vector machine and feature selection using simulated annealing approach," Applied soft computing, vol. 8, pp. 1505-1512, 2008.
[32] S. Lin, K. Ying, S. Chen, and Z. Lee, "Particle swarm optimization for parameter determination and feature selection of support vector machines," Expert systems with applications, vol. 35, pp. 1817-1824, 2014.
[33] C. Huang and C. Wang, "A GA-based feature selection and parameters optimizationfor support vectormachines," Expert systems with applications, vol. 31, pp. 231-240, 2012.