کاهش ابعاد روش پنهان‌شکنی CDF با استفاده از یک روش انتخاب ویژگی مبتنی بر تئوری گراف

سعید آزادی‌فر، سیدحسین خواسته و محمدهادی ادریسی

چکیده: پنهان‌شکنی دانش کشف حضور داده پنهان در یک رسانه پوششی است. هدف پنهان‌شکنی جلوگیری از رسیدن روش‌های پنهان‌نگاری به اهداف خود می‌باشد. یکی از معروف‌ترین روش‌های پنهان‌شکنی روش CDF است که در این پژوهش استفاده شده است.

یکی از چالش‌های عمده در مسئله پنهان‌شکنی تصاویر تعداد زیاد ویژگی‌های استخراج‌شده برای این کار است. مجموعه‌های داده‌ای با ابعاد بالا از دو جهت باعث کاهش عملکرد پنهان‌شکنی می‌شود. از یک طرف با افزایش ابعاد داده‌ها، حجم محاسبات افزایش پیدا می‌کند و از طرف دیگر مدلی که بر اساس داده‌های با ابعاد بالا ساخته می‌شود دارای قابلیت تعمیم پایینی است و احتمال بیش‌برازش افزایش می‌یابد. در نتیجه، کاهش ابعاد مسئله می‌تواند هم پیچیدگی محاسباتی را کاهش داده و هم باعث بهبود عملکرد پنهان‌شکنی شود. در این مقاله تلاش شده با ترکیب مفهوم زیرگراف کامل بیشینه وزن‌دار و معیار مرکزیت یال و در نظر گرفتن مناسب‌بودن هر ویژگی، ویژگی‌های تأثیرگذار و دارای حداقل افزونگی به‌عنوان ویژگی‌های نهایی انتخاب شوند. نتایج شبیه‌سازی بر روی مجموعه داده‌های SPAM و CC-PEV نشان داد روش پیشنهادی دارای عملکرد مناسبی است و به دقت حدود ۹۶٪ در تشخیص جاسازی داده در تصاویر دست پیدا کرده و همچنین این روش در مقایسه با روش‌های شناخته‌شده قبلی دارای دقت بالاتری است.

***کلیدواژه:* پنهان‌شکنی، پنهان‌نگاری، انتخاب ویژگی، کاهش بعد.**

۱- مقدمه

پنهان‌نگاری در بسترهای دیجیتال موضوعی است که در یک دهه اخیر توجه بسیاری از محققین را به خود جلب کرده است. روش‌های پنهان‌نگاری روش‌هایی هستند که هدف آنها پنهان‌کردن یک پیام سرّی در یک رسانه پوششی (مانند صوت، تصویر و فیلم) است به نحوی که وجود پیام سری برای کسی غیر از گیرنده پیام قابل کشف نباشد. در مقابل پنهان‌نگاری و برای کشف حضور داده پنهان در یک رسانه پوششی، دانشی به نام پنهان‌شکنی به وجود آمده است.

تحلیل‌ها و حملات پنهان‌شکنی بر خلاف رمزشکنی، به دنبال کشف مفهوم و محتوای پیام‌های سری است. یک حمله پنهان‌شکنی موفق، حمله‌ای است که با احتمال بهتر نسبت به یک حدس تصادفی بتواند محیط گنجانه حاوی اطلاعات جاسازی‌شده را از محیط پوشانه اصلی و طبیعی متمایز سازد. از سوی دیگر، اکثر حملات پنهان‌شکنی بر خلاف حملاتی که علیه سیستم‌های نهان‌نگاری[1] استفاده می‌شوند، حتی به دنبال تغییر پیام جاسازی‌شده در پوشانه هم نیستند و فقط در پی کشف وجود پیام می‌باشند.

روش‌های پنهان‌شکنی را بر مبنای دو پیش‌فرض اساسی متفاوت که دشمن از الگوریتم مورد استفاده برای پنهان‌نگاری، مطلع است یا خیر [۱]، به دو گروه پنهان‌شکنی خاص یا هدفمند[2] و پنهان‌شکنی عام یا کور[3] تقسیم می‌کنند. در روش‌های هدفمند، الگوریتم استفاده‌شده در پنهان‌نگاری برای دشمن مفروض یا معلوم است و حملات پنهان‌شکنی به طور خاص برای آن الگوریتم طراحی و اجرا می‌شوند. در حالی که در روش‌های کور، دشمن بدون اطلاع از نوع الگوریتمی که احتمالاً برای پنهان‌نگاری استفاده شده است، به طور عام سعی در تشخیص پدیده جاسازی اطلاعات در محیط رسانه را دارد. وجه مشترک هر دو گروه روش‌های پنهان‌شکنی این است که حملات بر اساس استخراج یک یا چند ویژگی[4] کلیدی حاصل‌شده از برخی پردازش‌ها و تبدیلات اولیه مناسب بر روی محیط رسانه مشکوک به جاسازی انجام می‌شوند [۲]. در روش‌های هدفمند با توجه به معلوم‌بودن الگوریتم جاسازی، طراحی تبدیلات لازم برای استخراج بردارهای ویژگی به مراتب آسان‌تر از روش‌های کور است زیرا در روش‌های کور ویژگی‌ها باید آن قدر جامع و فراگیر باشند که تقریباً تمامی مشخصات ممکن سیگنال‌های پوشانه را در خود نشان دهند.

یکی از چالش‌های عمده در مسئله پنهان‌شکنی تصاویر تعداد زیاد ویژگی‌های استخراج‌شده برای این کار است. مجموعه‌های داده‌ای با ابعاد بالا از دو جهت باعث کاهش عملکرد پنهان‌شکنی می‌شود. از یک طرف با افزایش ابعاد داده‌ها، حجم محاسبات افزایش می‌یابد و از طرف دیگر مدلی که بر اساس داده‌های با ابعاد بالا ساخته می‌شود دارای قابلیت تعمیم پایینی است و احتمال بیش‌برازش افزایش پیدا می‌کند. در نتیجه، کاهش ابعاد مسئله می‌تواند هم پیچیدگی محاسباتی را کاهش داده و هم باعث بهبود عملکرد پنهان‌شکنی شود. بنابراین برای ارائه یک روش با عملکرد بالا در پنهان‌شکنی نیاز به استفاده از الگوریتم‌های انتخاب ویژگی وجود دارد.

هدف اصلی این تحقیق ارائه یک روش مبتنی بر خوشه‌بندی ویژگی‌ها برای انتخاب ویژگی‌های مناسب در پنهان‌شکنی تصویر است. در روش پیشنهادی ویژگی‌های مجموعه داده‌ای به تعدادی خوشه تقسیم‌بندی شده و سپس از هر خوشه ویژگی‌های مناسب انتخاب می‌شود. با استفاده از انتخاب ویژگی و کاهش ابعاد داده‌ای تلاش می‌شود که هم دقت

این مقاله در تاریخ ۱۹ دی ماه ۱۳۹۶ دریافت و در تاریخ ۱۵ آبان ماه ۱۳۹۷ بازنگری شد.

سعید آزادی‌فر، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی خواجه نصیرالدین طوسی، تهران، ایران، (email: saeid.azadifar@email.kntu.ac.ir).

سیدحسین خواسته (نویسنده مسئول)، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی خواجه نصیرالدین طوسی، تهران، ایران، (email: khasteh@kntu.ac.ir).

محمدهادی ادریسی، دانشکده مهندسی کامپیوتر، دانشگاه اصفهان، اصفهان، ایران، (email: h.edrisi@ui.ac.ir).

1. Watermarking
2. Targeted (Specific) Steganalysis
3. Blind (Universal) Steganalysis
4. Feature

پنهان‌شکنی مناسب باشد و هم پیچیدگی محاسباتی و زمان اجرا کاهش پیدا کند.

در این تحقیق جدیدترین مقالات و تحقیقات انجام‌شده در زمینه پنهان‌شکنی تصاویر مطالعه شده و مزایا و معایب هر کدام از این روش‌ها معرفی می‌شود. همچنین چالش‌های موجود در زمینه پنهان‌شکنی تصاویر بررسی شده و برای حل این مشکلات و چالش‌ها، با استفاده از انتخاب ویژگی مبتنی بر خوشه‌بندی گراف یک روش جدید برای بهبود عملکرد پنهان‌شکنی تصاویر ارائه می‌شود.

در روش پیشنهادی ویژگی‌های اولیه در پنهان‌شکنی تصویر به صورت یک گراف وزن‌دار مدل می‌شوند. در این گراف گره‌ها، وزن‌های گراف و یال‌ها، شباهت بین ویژگی‌ها را نشان می‌دهند. پس از بازنمایی گرافی مسئله و یافتن زیرگراف‌های کامل بیشینه وزن‌دار در یک فرایند تکراری ویژگی‌های مناسب با استفاده از معیار مرکزیت یال و امتیاز فیشر انتخاب می‌شوند. با توجه به استفاده از معیار امتیاز فیشر، این کار سبب می‌شود که شباهت بین ویژگی‌های انتخابی به حداقل رسیده و همچنین ارتباط بین ویژگی‌ها با کلاس هدف نیز بیشترین مقدار خود را دارا باشد.

در پنهان‌شکنی پیچیدگی محاسباتی و زمانی بسیار بااهمیت است زیرا معمولاً وجود داده در یک تصویر باید به صورت برخط و در مدت زمان بسیار کمی تشخیص داده شود [۳]. بر اساس تحقیقات انجام‌شده در این حوزه، زمانی که از مجموعه داده‌ای ویژگی‌های نامناسب و دارای افزونگی حذف می‌شوند عملکرد پنهان‌شکنی از چند جهت افزایش پیدا می‌کند. از یک طرف به جهت کاهش تعداد ویژگی‌ها، پیچیدگی محاسباتی و زمان اجرا کاهش پیدا می‌کند و از طرف دیگر با کاهش تعداد ویژگی‌ها و پارامترهای الگوریتم طبقه‌بندی احتمال بیش‌برازش کاهش پیدا می‌کند. با توجه به این دلایل می‌توان انتظار داشت که روش پیشنهادی هم از نظر دقت و هم از نظر پیچیدگی محاسباتی نسبت به روش‌های پیشین دارای برتری خواهد بود.

همچنین با استفاده از مجموعه‌های داده‌ای استاندارد پنهان‌شکنی تصاویر مانند SPAM و CC-PEV [۴] عملکرد روش پیشنهادی مورد ارزیابی قرار می‌گیرد.

پس از پیاده‌سازی روش پیشنهادی و همچنین تهیه مجموعه‌های داده‌ای، عملکرد روش پیشنهادی با جدیدترین و شناخته‌شده‌ترین روش‌های انتخاب ویژگی بر روی مجموعه داده‌های ذکرشده مقایسه می‌شود. نتایج مقایسه روش‌های مختلف در قالب شکل‌ها و جدول‌های مختلف ارائه شده و به صورت کامل تحلیل نتایج صورت خواهد گرفت. همچنین از دیدگاه آماری و استفاده از آزمون‌های مختلف آماری مانند فریدمن روش‌های مختلف با یکدیگر مقایسه خواهند شد. در واقع با تحلیل آماری آزمون فریدمن تلاش می‌شود که عملکرد روش پیشنهادی از دیدگاه آماری با سایر روش‌های پیش‌تر ارائه‌شده مقایسه گردد.

در ادامه مقاله در بخش دوم روش‌های پیشتر ارائه‌شده برای پنهان‌شکنی تصویر مورد مطالعه قرار می‌گیرد. در بخش سوم با استفاده از الگوریتم‌های گراف یک روش انتخاب ویژگی کارا برای پنهان‌شکنی تصویر ارائه می‌شود. پس از انتخاب ویژگی‌های مناسب، دقت پنهان‌شکنی تصویر مورد ارزیابی قرار می‌گیرد. در بخش چهارم برای بررسی کارایی روش پیشنهادی آزمایش‌های گوناگونی طراحی شده و عملکرد روش پیشنهادی مورد بررسی قرار می‌گیرد. همچنین عملکرد روش پیشنهادی از جنبه‌های مختلف با روش‌های پیشتر ارائه‌شده مورد مقایسه قرار می‌گیرد. در بخش پنجم به یک جمع‌بندی کلی از روش ارائه‌شده پرداخته و مهم‌ترین نوآوری‌های موجود در این تحقیق بیان می‌شود.

۲- کارهای پیشین

همان طور که گفته شد روش‌های پنهان‌شکنی را بر مبنای دو پیش‌فرض اساسی متفاوت که دشمن از الگوریتم مورد استفاده برای پنهان‌نگاری، مطلع است یا خیر [۱]، به دو گروه پنهان‌شکنی خاص یا هدفمند و پنهان‌شکنی عام یا کور تقسیم می‌کنند. ساده‌ترین روش پنهان‌نگاری، جایگزینی بیت‌های کم‌ارزش فضای نمونه (LSB-F) است که ظرفیت مناسبی را ارائه داده و توسط گوش و چشم نیز قابل کشف نیست [۱]. حملات زیادی ارائه شده‌اند که می‌توانند وجود داده پنهان‌شده به روش جایگزینی بیت کم‌ارزش را آشکار کنند. در میان این روش‌ها، روش‌هایی هستند که به صورت ترکیبی از سایر روش‌ها استفاده می‌کنند. این روش‌ها دقت بیشتری دارند و میزان جاسازی را بهتر تخمین می‌زنند. یکی از این روش‌ها تست مربع کای (χ^2) است. این حمله قادر به کشف روش‌هایی است که برای جاسازی از روش LSB-F در حوزه مکان و هر حوزه دیگر استفاده کرده باشند. Westfeld و Pfitzman در [۵] از یک تست χ^2 استفاده کرده‌اند تا تشخیص دهند که آیا توزیع فرکانسی مشاهده‌شده در یک تصویر با توزیع مورد انتظار مطابق است یا خیر؟ که انحراف ناشی از جاسازی داده مخفی‌شده را نشان می‌دهد. اگرچه تصویر پوشانه شناخته‌شده نیست اما مجموعه ضرایب DCT همسایه ثابت می‌ماند که این نکته باعث می‌شود تا محاسبه توزیع مورد انتظار در تصویر گنجانه قابل محاسبه باشد.

Fridrich و همکارانش در [۶]، روش حمله مؤثری موسوم به RS را در یک حمله پنهان‌شکنی کمّی برای آشکارسازی LSB-F حتی با نرخ جاسازی کم در حوزه مکان، مطرح کرده‌اند.

یکی از دقیق‌ترین و مؤثرترین حملات کمی پنهان‌شکنی در مقابل جایگزینی بیت‌های کم‌ارزش در حوزه مکان هر نوع سیگنال دیجیتال، روش ارائه‌شده در [۷] توسط Dumitrescu و همکارانش، مبتنی بر تحلیل رفتار زوج نمونه‌های مجاور در قبل و بعد از جاسازی است که به عنوان روش SPA شناخته می‌شود. این روش نیز مانند روش RS از راهبرد استفاده از خواص معلوم در سیگنال پوشانه برای طراحی یک حمله هدفمند بهره جسته و در صورت انجام جاسازی LSB-F در سیگنال مورد بررسی می‌تواند با دقت نسبتاً بالایی، طول پیام جاسازی‌شده را تخمین بزند.

در روش LSB-M به جای آن که بیت‌های کم‌ارزش نمونه‌ها با بیت‌های داده جایگزین شوند با آن تطابق داده می‌شوند. در ساده‌ترین روش از این خانواده، اگر بیتی که قرار است در مقدار یک نمونه جاسازی شود با کم‌ارزش‌ترین بیت آن نمونه تطابق داشته باشد، مقدار نمونه تغییری نمی‌کند، در غیر این صورت به صورت کاملاً تصادفی، مقدار ۱ یا ۱- به نمونه افزوده می‌شود [۸]. برخلاف وجود شباهت زیاد دو روش LSB-F و LSB-M، موفقیت حملات ارائه‌شده علیه این دو روش قابل مقایسه نیست. اگرچه تا به حال حملات موفق بسیار برای روش LSB-F پیشنهاد شده است اما تعداد حملات موفق برای روش LSB-M بسیار کم است. با توجه به این که در این روش LSB-M تمام مقادیر می‌توانند کاهش یا افزایش یابند و این تقارن مانع ایجاد زوج مقدارها در هیستوگرام می‌شود، بنابراین در برابر روش‌های حمله به LSB-F مقاوم است. از جمله حملات ارائه‌شده برای LSB-M می‌توان به روش‌های [۹] و [۱۰] اشاره کرد اما تمام این روش‌ها در همه شرایط کارایی یکسانی ندارند و در اکثر موارد کارایی آنها بستگی به نوع تصویر پوشانه مورد استفاده دارد. روش ارائه‌شده توسط Westfeld، یکی از حملات موفق به روش LSB-M است. فرض موجود در این روش این است که سیگنال پوشانه شامل تعداد

نسبتاً کمی از رنگ‌های مختلف است. روش LSB-M باعث می‌شود مقدارهای نزدیک به هم زیادی در پیکسل‌های تصویر ایجاد شود.

یکی از معروف‌ترین این حملات که اولین بار در [۱۱] توسط Harmsen ارائه شد، حمله‌ای مبتنی بر جابه‌جایی مرکز ثقل تابع مشخص هیستوگرام (HCF-COM) است که می‌تواند معیار مناسبی برای ارزیابی بسیاری از روش‌های پنهان‌نگاری تطبیقی از جمله LSB-M باشد.

انتخاب ویژگی، فرایندی است که در آن زیرمجموعه‌ای از ویژگی‌های موجود بر اساس معیارهای اهمیت ویژگی انتخاب می‌شود. انتخاب ویژگی از دهه ۱۹۷۰ تا کنون به‌عنوان یک موضوع تحقیقاتی مهم و فعال در شناسایی الگو، یادگیری ماشین و داده‌کاوی شناخته می‌شود. هدف نهایی فرایند انتخاب ویژگی، حذف ویژگی‌های نامرتبط[۱] و دارای افزونگی[۲] است. در طی فرایند انتخاب، یک زیرمجموعه از ویژگی‌های اولیه انتخاب شده و مناسب‌بودن آن با استفاده از یک معیار ارزیابی اندازه گرفته می‌شود. از یک دیدگاه کلی، روش‌های انتخاب ویژگی به دو دسته باناظر[۳] و بدون ناظر[۴] تقسیم‌بندی می‌شوند [۱۲] تا [۱۴]. در روش‌های باناظر یک مجموعه از الگوهای آموزشی وجود دارند که هر الگو به‌وسیله برداری از مقادیر ویژگی به همراه برچسب دسته توصیف می‌شود، در حالی که در روش‌های بدون ناظر با مجموعه داده‌ای بدون برچسب دسته مواجه هستیم. راهکارهای انتخاب ویژگی را می‌توان به چهار دسته فیلتر[۵]، پوششی[۶]، ترکیبی[۷] و تعبیه‌شده[۸] تقسیم کرد. با توجه به این که روش پیشنهادی یک روش فیلتر هست در ادامه به بررسی مهم‌ترین روش‌های فیلتر ارائه‌شده بر مبنای تئوری گراف پرداخته می‌شود.

اخیراً روش‌های مبتنی بر تئوری گراف مانند روش تعبیه طیفی[۹] [۱۵]، خوشه‌بندی طیفی[۱۰] [۱۶] و یادگیری نیمه‌نظارتی[۱۱] [۱۷] به دلیل داشتن قابلیت تعریف صحیح از رابطه شباهت (نزدیکی) در میان نمونه‌های یک مجموعه داده، نقش مهمی در حوزه یادگیری ماشین ایفا می‌کنند. در مسئله انتخاب ویژگی، با نمایش فضای ویژگی‌ها در مدل گراف، روش‌های مبتنی بر گراف یک چارچوب کلی و انعطاف‌پذیر ارائه می‌دهند که ساختار چندگانه و روابط بین بردارهای ویژگی را به صورت گرافیکی نمایش می‌دهد. از معروف‌ترین روش‌های انتخاب ویژگی، روش‌های امتیاز فیشر [۱۸] و امتیاز لاپلاسین [۱۹] هستند که از تئوری گراف به منظور انتخاب ویژگی استفاده می‌کنند.

در [۲۰] یک معیار انتخاب ویژگی نیمه‌نظارتی طیفی ارائه شده که نام آن s-Laplacian score است. بر اساس این معیار، در این مقاله یک روش جدید انتخاب ویژگی مبتنی بر گراف به نام GSFS ارائه شده است که در آن از تئوری گراف طیفی و تئوری اطلاعات متقابل شرطی[۱۲] به منظور انتخاب ویژگی‌های مرتبط و همچنین حذف ویژگی‌های افزونه استفاده می‌شود.

همچنین در [۱۲] از تئوری اطلاعات مبتنی بر ابرگراف[۱۳] به منظور انتخاب ویژگی استفاده شده است. در این روش ابتدا فضای ویژگی‌ها با استفاده از یک ابرگراف مدل‌سازی می‌شود که در آن هر گره از ابرگراف مربوط به یک ویژگی است و هر یال از ابرگراف شامل یک مقدار وزنی است که متناسب با اطلاعات تعاملی چندبعدی میان ویژگی‌هایی که از طریق یالی در ابرگراف به یکدیگر متصل شده‌اند می‌باشد. در مرحله بعد این روش با اعمال یک الگوریتم خوشه‌بندی خاص بر روی گراف، زیرمجموعه‌ای از ویژگی‌های اطلاعاتی انتخاب می‌شود.

در [۲۱] یک الگوریتم سریع انتخاب ویژگی مبتنی بر خوشه‌بندی به منظور انتخاب ویژگی از مجموعه داده‌هایی با ابعاد بالا پیشنهاد شده که شامل ۴ مرحله زیر است: در مرحله اول ویژگی‌های نامرتبط ار مجموعه داده حذف می‌شوند. در مرحله دوم بر روی ویژگی‌های باقیمانده درخت پوشای کمینه (MST)[۱۴] ساخته می‌گردد. در مرحله سوم عمل تقسیم‌بندی درخت پوشای کمینه انجام می‌شود. در مرحله چهارم ویژگی‌های مهم و اطلاعاتی انتخاب می‌گردند.

به منظور حل مشکل مقاله قبلی در [۲۲] با ترکیب مفهوم یافتن زیرگراف با بیشترین چگالی با مفهوم خوشه‌بندی گراف، نویسندگان مقاله یک روش انتخاب ویژگی بدون ناظر پیشنهاد داده‌اند. این روش شامل دو مرحله است: در گام اول زیرگراف با بیشترین چگالی تعیین شده که در نتیجه ویژگی‌های انتخابی در این مرحله حداکثر عدم افزونگی را در میان یکدیگر دارند. در گام دوم با استفاده از الگوریتم خوشه‌بندی ویژگی‌ها که بر روی ویژگی‌های انتخابی مرحله قبل اعمال می‌شود مجموعه‌ای با تعداد ویژگی‌های انتخابی کمتر تولید می‌شود.

در [۲۳] دو روش انتخاب ویژگی بدون ناظر مبتنی بر فیلتر و از نوع چندمتغیره پیشنهاد شده که با تحلیل ارتباط و افزونگی ویژگی‌ها، بهترین آنها را انتخاب می‌کند. در هر دو روش، ابتدا فضای جستجو به صورت یک گراف نمایش داده شده و در مرحله بعد ویژگی‌ها با استفاده از روش بهینه‌سازی کلونی مورچگان رتبه‌بندی می‌شوند. علاوه بر این در این روش یک معیار اطلاعاتی و اکتشافی جدید به منظور بهبود دقت نتایج انتخاب ویژگی پیشنهاد شده که شباهت بین زیرمجموعه‌های ویژگی را در نظر می‌گیرد.

در [۲۴] یک روش جدید انتخاب ویژگی مبتنی بر گراف به نام GCNC ارائه شده که شامل سه مرحله است: در گام اول مجموعه تمام ویژگی‌ها در قالب یک گراف وزنی نمایش داده می‌شود. در گام دوم با استفاده از یک الگوریتم تشخیص انجمن[۱۵] مجموعه گره‌های گراف (ویژگی‌ها) به تعدادی خوشه تقسیم‌بندی می‌شوند. در گام سوم با استفاده از یک روال جستجوی تکراری و بر اساس معیار گره مرکزی[۱۶] زیرمجموعه نهایی از ویژگی‌ها انتخاب می‌شود.

در [۲۵] یک روش انتخاب ویژگی جدید بر مبنای اطلاعات متقابل ارائه شده که بر اساس نرمال‌سازی حداکثر میزان ارتباط و حداقل میزان افزونگی عمل می‌کند. همچنین در [۲۶] یک روش انتخاب ویژگی با استفاده از الگوریتم بهینه‌سازی جنگل ارائه شده است.

در [۱۳] یک روش انتخاب ویژگی بدون ناظر بر مبنای خوشه‌بندی زیرفضا معرفی شده که یادگیری برچسب‌های خوشه‌های حاوی نمونه‌های

1. Irrelevant
2. Redundant
3. Supervised
4. Unsupervised
5. Filter Approach
6. Wrapper Approach
7. Hybrid Approach
8. Embedded Approach
9. Spectral Embedding
10. Spectral Clustering
11. Semi-Supervised Learning
12. Conditional Mutual Information
13. Hypergraph
14. Minimum Spanning Tree
15. Community Detection
16. Node Centrality

جدول ۱: مقایسه روش‌های انتخاب ویژگی.

سال	توضیحات	بدون ناظر	باناظر	روش
۲۰۱۱	یک روش انتخاب ویژگی تک‌متغیره مبتنی بر راهکار فیلتر که ویژگی‌های با بهترین توانایی تمیز را می‌یابد. تعدادی ویژگی با رتبه بالا را انتخاب می‌کند که دارای بیشترین حفظ قدرت یا نگهدارندگی باشند [۱۸].		√	FS
۲۰۰۵	یک روش انتخاب ویژگی چندمتغیره که به دنبال انتخاب ویژگی‌هایی با بیشترین ارتباط با دسته هدف با استفاده از یک معیار ارتباط می‌باشد. همچنین از یک معیار افزونگی به منظور کاهش افزونگی میان ویژگی‌ها استفاده می‌کند [۱۴].		√	mRMR
۲۰۱۳	یک روش انتخاب ویژگی سریع مبتنی بر خوشه‌بندی است. ویژگی‌ها بر اساس روش‌های خوشه‌بندی تئوری گراف به تعدادی خوشه تقسیم می‌شوند. از هر خوشه ویژگی‌های نماینده‌گر که ارتباط قوی‌تری با دسته هدف دارند انتخاب می‌شوند [۲۷].		√	Fast
۲۰۱۵	یک روش انتخاب ویژگی مبتنی بر راهکار فیلتر که از ترکیب مفهوم خوشه‌بندی گراف و مرکزیت گره استفاده کرده است. این روش توانایی خوبی در حذف ویژگی‌های نامرتبط دارد [۲۴].	√		GCNC
۲۰۱۲	یک راهکار فیلتر چندمتغیره است. ابتدا ویژگی‌ها بر اساس یک معیار مشخص، جهت ارزیابی مناسب‌بودن، رتبه‌دهی و مرتب می‌شوند و مناسب‌ترین آنها به‌عنوان اولین ویژگی انتخاب می‌گردد. سپس در هر تکرار از الگوریتم، یک ویژگی انتخاب می‌شود اگر شباهت آن به آخرین ویژگی انتخاب‌شده از یک حد آستانه کمتر باشد [۲۸].	√		RRFS
۲۰۱۴	یک روش انتخاب ویژگی مبتنی بر یافتن زیرگراف چگال است. از واریانس برای فاز خوشه‌بندی به منظور انتخاب نمونه ویژگی‌های اولیه استفاده شده است. در حالی که از اطلاعات متقابل نرمالایزشده به منظور اطلاق هر ویژگی انتخاب‌نشده به نزدیک‌ترین خوشه نماینده آن بهره برده است [۲۲].	√		DSFFC

آموزشی با استفاده از بازنمایی بر اساس خوشه‌بندی زیرفضا انجام می‌شود و ویژگی‌هایی که توانایی خوبی در حفظ برچسب خوشه‌ها دارند انتخاب می‌شوند. مقایسه روشهای مختلف انتخاب ویژگی در جدول ۱ قابل مشاهده است.

۳- روش پیشنهادی

در این مقاله تلاش می‌شود که با استفاده از یک روش انتخاب ویژگی مبتنی بر خوشه‌بندی گراف، ویژگی‌های دارای افزونگی و یا نامرتبط حذف شده و در نتیجه دقت الگوریتم پنهان‌شکنی مناسب باشد. برای این کار ویژگی‌های اولیه در مجموعه داده‌ای مورد استفاده در پنهان‌شکنی به تعدادی خوشه تقسیم‌بندی شده و سپس با استفاده از معیار انتخاب ویژگی پراکندگی داده، از هر خوشه ویژگی‌های مناسب انتخاب می‌شوند.

روش پیشنهادی از سه مرحله تشکیل شده است. در مرحله اول، فضای مسئله به‌صورت گرافی بازنمایی می‌شود. در مرحله دوم، در یک فرایند تکراری زیرگراف کامل بیشینه وزن‌دار یافته شده [۲۹] و سپس در مرحله سوم از روش پیشنهادی با استفاده از ترکیب مرکزیت یال [۳۰] و امتیاز فیشر در حالت باناظر (و امتیاز لاپلاسین در حالت بدون ناظر) از هر زیرگراف کامل ویژگی‌هایی که دارای حداقل افزونگی و دارای حداکثر ارتباط با دسته هدف هستند شناسایی و به‌عنوان ویژگی‌های نهایی انتخاب می‌شوند. در واقع در روش پیشنهادی سعی شده که یک معیار جدید برای حذف افزونگی میان ویژگی‌ها معرفی شود. همچنین در این روش، تعداد ویژگی‌های نهایی به‌صورت خودکار تعیین می‌شوند.

در مرحله اول مجموعه ویژگی‌های یک مجموعه داده به صورت ساختار گرافی مدل می‌شوند که در این ساختار گره‌های گراف به عنوان ویژگی‌های مجموعه داده و یال‌های آن ارتباط یا شباهت میان ویژگی‌های یک مجموعه داده را نشان می‌دهند. برای نمایش شباهت میان ویژگی‌ها، یال‌های گراف به صورت وزن‌دار نمایش داده می‌شوند که به منظور محاسبه وزن یال‌ها، از معیار شباهت پیرسون که در [۲۳] و [۲۴] مورد استفاده قرار گرفته است به صورت زیر استفاده می‌شود

$$w_{i,j}=\left|\frac{\sum_{p}(x_i-\overline{x_i})(x_j-\overline{x_j})}{\sqrt{\sum_{p}(x_i-\overline{x_i})^2}\sqrt{\sum_{p}(x_j-\overline{x_j})^2}}\right| \quad (۱)$$

در معادله فوق پارامترهای x_i و x_j بردار ویژگی‌های F_i و F_j و متغیرهای $\overline{x_i}$ و $\overline{x_j}$ مقدار میانگین بردارهای x_i و x_j بر روی p نمونه آموزشی را نشان می‌دهند. مقدار حاصل از معادله فوق در بازه صفر و یک است که مقدار صفر نشان‌دهنده عدم شباهت کامل و مقدار یک، شباهت کامل میان ویژگی‌ها را نشان دهد. به منظور نرمال‌سازی این مقدار از معادله زیر استفاده شده است

$$\hat{w}_{i,j}=\frac{1}{1+\exp(-\frac{w_{i,j}-\overline{w}}{\sigma})} \quad (۲)$$

در معادله فوق $\overline{w}$ و σ به ترتیب میانگین و واریانس از توزیع مقادیر وزنی است.

در مرحله اول از روش پیشنهادی، مسئله انتخاب ویژگی در یک گراف بدون جهت و وزن‌دار $G=\langle F,E\rangle$ بازنمایی می‌شود. در این گراف، $F=\{F_1,F_2,\ldots,F_n\}$ مجموعه ویژگی‌های اولیه هستند که هر ویژگی توسط یک گره نمایش داده می‌شود و $E=\{(F_i,F_j):F_i,F_j\in F\}$ مجموعه یال‌های گراف هستند که شباهت میان ویژگی‌ها را نشان می‌دهند. در مرحله دوم به منظور محاسبه شباهت بین ویژگی‌ها از ضریب همبستگی پیرسون استفاده شده است. همچنین به منظور کاهش تعداد یال‌های گراف و کارایی بهتر الگوریتم شباهت‌هایی که مقدار آنها کمتر از یک آستانه θ باشند حذف می‌شوند. هدف اصلی این روش پیشنهادی این است که یک زیرمجموعه از ویژگی‌ها اولیه به‌گونه‌ای انتخاب شود که الف) دارای کمترین شباهت با یکدیگر و ب) دارای بیشترین ارتباط با دسته هدف باشند. به منظور دستیابی به هدف اول از دو مفهوم زیرگراف کامل وزن‌دار و مرکزیت یال استفاده شده است. زیرگراف کامل وزن‌دار یعنی مجموعه‌ای از رئوس گراف که دوبه‌دو به هم متصل هستند و دارای بیشترین وزن در میان بقیه زیرگراف‌ها می‌باشد. در مرحله سوم در یک رویه تکرارشونده به این صورت عمل می‌شود که در هر تکرار یک زیرگراف کامل بیشینه وزن‌دار انتخاب شده و سپس از بین ویژگی‌های موجود در این زیرگراف کامل بیشینه ویژگی‌های مناسب با استفاده از دو مفهوم مرکزیت یال و امتیاز فیشر در حالت باناظر و امتیاز لاپلاسین در حالت بدون ناظر انتخاب می‌شود. سپس ویژگی‌های موجود در این زیرگراف کامل بیشینه از مجموعه ویژگی‌های اولیه حذف شده و فرایند

یک روش انتخاب ویژگی مبتنی بر گراف کامل بیشینه وزن‌دار و مرکزیت یال

ورودی
D_T : مجموعه داده آموزشی
G : گراف وزن‌دار ویژگی‌های اولیه
θ : آستانه برای شباهت
δ : آستانه برای EC

خروجی $\{f'_1,\ldots,f'_m\}$ = مجموعه ویژگی‌های انتخاب‌شده

شروع الگوریتم

۱ محاسبه مقدار ارتباط میان همه ویژگی‌های موجود با اندیس i $(i=1,2,\ldots,n)$
۲ نرمال‌سازی ارتباط میان ویژگی‌ها
۳ اعمال حد آستانه بر روی گراف اولیه G و به دست آوردن گراف G'
۴ **حلقه ۱**
۵ یافتن گراف کامل بیشینه وزن‌دار و نام‌گذاری آن با Max_Clique
حلقه ۲
۶ برای هر ویژگی موجود در Max_Clique
۷ مقدار امتیاز فیشر (یا لاپلاسین) محاسبه شود
۸ این ویژگی از Max_Clique حذف شود
انتهای حلقه ۲
۹ ویژگی با بیشترین مقدار امتیاز فیشر (یا لاپلاسین) به مجموعه ویژگی‌های انتخاب‌شده اضافه شود
حلقه ۳
۱۰ تا زمانی که مقدار شدت ارتباط (EC) میان ویژگی‌های با بیشترین مقدار امتیاز فیشر (یا لاپلاسین) و ویژگی‌های موجود در مجموعه ویژگی‌های انتخاب‌شده از مقدار آستانه دلتا کمتر است
ویژگی با بیشترین مقدار امتیاز فیشر (یا لاپلاسین) به مجموعه ویژگی‌های انتخاب‌شده اضافه شود
۱۱ این ویژگی از Max_Clique حذف شود
انتهای حلقه ۳
۱۲ Max_Clique از گراف G' حذف شود
۱۳ تا زمانی که حداقل یک گراف کامل بیشینه وزن‌دار پیدا می‌شود این حلقه تکرار شود
انتهای حلقه ۱
۱۴ همه ویژگی‌های باقیمانده به مجموعه ویژگی‌های انتخاب‌شده اضافه شوند.
انتهای الگوریتم

شکل ۱: شبه‌کد روش پیشنهادی.

ذکرشده برای ویژگی‌های باقیمانده تکرار می‌شود. به عبارت دیگر با استفاده از مفهوم مرکزیت یال و امتیاز فیشر، تلاش می‌شود از هر زیرگراف کامل بیشینه ویژگی‌هایی انتخاب شوند که بهترین نماینده‌ها برای ویژگی‌های موجود در آن زیرگراف کامل بیشینه باشند. یافتن زیرگراف کامل بیشینه وزن‌دار مشابه خوشه‌بندی گراف است و باعث می‌شود که ویژگی‌های اولیه بر اساس شباهتشان با یکدیگر به تعدادی زیرگراف کامل بیشینه (خوشه) مختلف، تقسیم شوند. بنابراین ویژگی‌های موجود در هر زیرگراف کامل بیشینه (خوشه) دارای شباهت بیشتری با یکدیگر و ویژگی‌های موجود در زیرگراف کامل بیشینه (خوشه) مختلف دارای شباهت کمتری با هم هستند. این کار سبب می‌شود که افزونگی مابین ویژگی‌های انتخاب‌شده نهایی حداقل شده و ویژگی‌ها انتخاب‌شده به خوبی نماینده کل ویژگی‌های اولیه باشند. برای کاهش افزونگی بین ویژگی‌های انتخاب‌شده نهایی علاوه بر مفهوم زیرگراف کامل بیشینه از مفهوم مرکزیت یال استفاده شده است.

در تمام روش‌های پیش‌تر ارائه‌شده برای انتخاب ویژگی که از تئوری گراف بهره گرفته‌اند تنها از شباهت مستقیم بین ویژگی‌ها استفاده گردیده در حالی که در روش پیشنهادی به منظور ارزیابی شدت ارتباط میان ویژگی‌ها برای اولین بار از مفهوم مرکزیت یال استفاده شده است. استفاده از این مفهوم سبب می‌شود که شدت ارتباط بین دو ویژگی با دقت بیشتری شناسایی شود. به این صورت که از هر زیرگراف کامل بیشینه ویژگی که دارای بیشترین امتیاز فیشر یا لاپلاسین است کاندید شده و شدت ارتباط آن با مجموعه ویژگی‌های انتخاب‌شده از آن زیرگراف کامل با استفاده از مفهوم مرکزیت یال سنجیده می‌شود و در صورتی که مقادیر محاسبه‌شده برای مرکزیت یال از یک آستانه δ کمتر باشند آن ویژگی انتخاب خواهد شد. در ضمن از آنجا که مرکزیت یال در هر زیرگراف کامل بیشینه تنها برای ویژگی کاندید و مجموعه ویژگی‌های انتخاب‌شده از آن زیرگراف کامل محاسبه می‌شود و همچنین تعداد گره‌های موجود در هر زیرگراف کامل هم نسبت به تعداد کل ویژگی‌ها خیلی کمتر است، در نتیجه از نظر پیچیدگی زمانی چندان زمان‌بر نخواهد بود. در نتیجه ویژگی‌های انتخاب‌شده دارای افزونگی کمتری نسبت به سایر روش‌ها که از شباهت مستقیم استفاده می‌کنند هستند.

برای دستیابی به هدف دوم یعنی انتخاب ویژگی‌هایی که دارای بیشترین ارتباط با دسته هدف باشند از معیار امتیاز فیشر استفاده شده است. در واقع پس از انتخاب یک زیرگراف کامل بیشینه با استفاده از این معیار، ویژگی‌هایی از آن زیرگراف کامل انتخاب می‌شوند که دارای ارتباط بیشتری با دسته هدف باشند. شکل ۱ شبه‌کد روش پیشنهادی را نشان می‌دهد.

همچنین در زیرمجموعه ویژگی نهایی از هر زیرگراف کامل بیشینه حداقل یک ویژگی وجود دارد. بنابراین ویژگی‌های نهایی از تمام فضای ویژگی‌ها انتخاب شده و به‌خوبی قادر خواهند بود که اطلاعات لازم درباره

جدول ۲: مشخصات مجموعه‌های داده‌ای استفاده‌شده.

مجموعه داده‌ای	تعداد ویژگی	تعداد کلاس
SPAM	۶۸۶	۲
CC-PEV	۵۴۸	۲

دسته هدف را ارائه دهند. همچنین در انتخاب ویژگی‌های نماینده هر زیرگراف کامل بیشینه، از مفهوم مرکزیت یال و امتیاز فیشر استفاده شده است. استفاده از مرکزیت یال سبب انتخاب ویژگی‌های با حداقل افزونگی از هر زیرگراف کامل بیشینه می‌شود.

از طرف دیگر برای حذف ویژگی‌های نامرتبط از میزان مناسب‌بودن هر ویژگی استفاده شده است. استفاده از معیار مناسب‌بودن هر ویژگی در کنار مفهوم مرکزیت یال سبب می‌شود که ویژگی‌های انتخابی از هر زیرگراف کامل بیشینه هم از لحاظ افزونگی و ارتباط نماینده خوبی برای ویژگی‌های آن زیرگراف کامل بیشینه باشند و هم خود به‌صورت فردی ویژگی‌هایی کارا و مناسب باشند.

مقادیر پارامترهای θ و δ بر اساس روش سعی و خطا و انجام آزمایش‌های متعدد انتخاب شده‌اند، به این صورت که مقادیر متفاوتی برای این پارامترها در نظر گرفته شده و بر اساس آن مقادیر دقت روش اندازه‌گیری شد و سپس مقداری که دارای بیشترین دقت بود انتخاب گردیده است.

٤- ارزیابی روش پیشنهادی

در این بخش، عملکرد روش پیشنهادی CDF برای مسئله پنهان‌شکنی تصویر با استفاده از انتخاب ویژگی مورد ارزیابی قرار می‌گیرد. به این منظور، روش پیشنهادی با تعدادی دیگر از الگوریتم‌های انتخاب ویژگی مقایسه می‌شود و عملکرد آنها برای تشخیص پنهان‌شکنی مورد بررسی قرار می‌گیرد. این روش‌های انتخاب ویژگی عبارتند از:

- روش انتخاب ویژگی افزونگی- ارتباط (RRFS)[1] [۲۸]: یک روش کارا و چندمتغیره مبتنی بر راهکار فیلتر برای انتخاب ویژگی است که تلاش می‌کند یک زیرمجموعه از ویژگی‌ها را که دارای کمترین افزونگی با هم و همچنین بیشترین ارتباط با کلاس هدف باشند انتخاب کند. این روش قادر به انتخاب ویژگی در حالت بدون ناظر نیز است.
- روش حداقل افزونگی- حداکثر ارتباط (MRMR)[2] [۱۴]: یکی از معروف‌ترین روش‌های چندمتغیره مبتنی بر راهکار فیلتر است که در فرایند انتخاب ویژگی هم افزونگی ویژگی‌ها و هم ارتباط آنها را با کلاس هدف در نظر می‌گیرد.
- امتیاز فیشر (FS)[3] [۱۸]: یک روش انتخاب ویژگی تک‌متغیره مبتنی بر راهکار فیلتر است که هدف آن انتخاب یک زیرمجموعه ویژگی است که در آن زیرمجموعه، فاصله بین الگوهای موجود در یک کلاس مشابه تا حد ممکن کمترین و فاصله بین الگوهای موجود در کلاس‌های مختلف تا حد ممکن زیاد باشد.
- روش GCNC [۳۱]: روش فیلتر مبتنی بر تئوری گراف چندمتغیره است که از ترکیب دو مفهوم خوشه‌بندی گراف و مرکزیت گره استفاده کرده است.

1. Relevance-Redundancy Feature Selection
2. Minimal Redundancy-Maximal Relevance
3. Fisher Score

برای نشان‌دادن قابلیت تعمیم روش پیشنهادی در دسته‌بندهای مختلف، در آزمایش‌ها از دو دسته‌بند ماشین بردار پشتیبان (SVM)[4] و بیز ساده (NB)[5] استفاده شده است. برای پیاده‌سازی دسته‌بندهای ذکرشده از ابزار وکا [۳۲] که شامل طیف وسیعی از الگوریتم‌های یادگیری ماشین است استفاده گردیده است. در آزمایش‌های انجام‌گرفته بر روی روش پیشنهادی، مجموعه‌های داده‌ای به‌طور تصادفی به داده‌های آموزشی (۶۶٪ از کل مجموعه داده‌ای) و داده‌های آزمایشی تقسیم می‌شوند. با توجه به تقسیم تصادفی داده‌ها به مجموعه‌های آموزشی و آزمایشی، هر روش پیشنهادی ۱۰ بار اجرا شده و از میانگین نتایج به‌دست‌آمده به‌عنوان معیار ارزیابی استفاده شده است.

٤-١ مجموعه داده‌ای

برای ارزیابی در این پژوهش، ابتدا با استفاده از روش‌های پنهان‌نگاری non-zero DCT Coefficient embedding و Nsf۵:۰.۱ و Nsf۵:۰.۳ بر روی مجموعه‌ای متشکل از ۱۰۰۰۰ تصویر مربوط به مجموعه داده SUN [۳۳] تصاویر پوشانه را ساخته‌ایم. سپس به منظور استفاده از روش CDF، از دو مجموعه داده‌ای Spam feature و CC-PEV [۴] برای ارزیابی عملکرد روش پیشنهادی و مقایسه عملکرد آن با سایر روش‌های انتخاب ویژگی استفاده شده است. مجموعه داده‌ای SPAM شامل ۶۸۶ ویژگی است و همچنین مجموعه داده‌ای CC-PEV شامل ۵۴۸ ویژگی است که در مجموع ۱۲۳۴ ویژگی استخراج گردیده و در جدول ۲ مشخصات مجموعه داده‌ای استفاده‌شده آمده است.

٤-٢ نتایج عملی

در این بخش به بررسی دقت دسته‌بندی در بین روش‌های مختلف انتخاب ویژگی پرداخته می‌شود. در تمام جدول‌های این بخش از میانگین دقت دسته‌بندی در ۱۰ اجرای مستقل برای مقایسه روش‌های مختلف استفاده شده است. جدول‌های ۳ تا ۶ به مقایسه دقت دسته‌بندی در روش پیشنهادی با سایر روش‌های انتخاب ویژگی بر روی دسته‌بندهای SVM و NB می‌پردازد. همان طور که نتایج گزارش‌شده در جدول‌ها نشان می‌دهد روش پیشنهادی در مقایسه با سایر روش‌های انتخاب ویژگی از دقت بالاتری برخوردار است.

در هر سطر میانگین دقت دسته‌بندی برای هر روش انتخاب ویژگی در ۱۰ اجرای مستقل و روش جاسازی و همچنین دسته‌بند مورد استفاده در هر جدول ذکر شده است. همچنین تعداد ویژگی‌های انتخاب‌شده به ترتیب از ۵۰۰ تا ۵۰ در نظر گرفته شده و دقت روش‌ها با آن تعداد ویژگی و با استفاده از دسته‌بند یکسان محاسبه شده است. با توجه به جدول‌ها و نتایج حاصل مشاهده می‌شود که روش پیشنهادی در اکثر موارد با تعداد ویژگی‌های یکسان انتخاب‌شده با سایر روش‌ها دارای دقت بالاتری نسبت به آنها است.

٤-٣ تحلیل آماری روش پیشنهادی

به منظور مقایسه میان روش پیشنهادی و روش‌های ذکرشده دیگر و همچنین تحلیل آماری نتایج از آزمون فریدمن [۳۴] استفاده شده است. آزمون فریدمن یک آزمون ناپارامتری است که برای مقایسه روش‌های مختلف بر روی دیتاست‌های متفاوت مورد استفاده قرار می‌گیرد به این

4. Support Vector Machines
5. Naive Bayes

جدول ۳: مقایسه کارایی روش پیشنهادی با سایر روش‌های انتخاب ویژگی بدون ناظر بر روی دسته‌بند SVM با استفاده از روش جاسازی NON-ZERO DCT COEFFICIENT با NSF ۵ : ۰/۱.

روش‌های انتخاب ویژگی	Evaluation criteria					
	SVM					
	دقت	تعداد ویژگی	دقت	تعداد ویژگی	دقت	تعداد ویژگی
DSFFC	۶۵٫۸۴		۶۱٫۷۵		۵۹٫۸۸	
RRFS	۶۳٫۶۸		۶۱٫۸۹		۶۰٫۲۳	
GCNC	۶۷٫۳۲	۵۰۰	۶۳٫۸۸	۴۰۰	۶۰٫۴۷	۳۰۰
FS	۶۶٫۴۸		۶۲٫۰۵		۶۱٫۸۴	
MCEC	**۶۸٫۲۹**		**۶۶٫۱۲**		**۶۳٫۲۱**	
DSFFC	۵۶٫۶۹		۵۳٫۶۳		۵۱٫۷۱	
RRFS	۵۷٫۶۰		۵۳٫۷۱		۵۲٫۷۱	
GCNC	۵۸٫۲۱	۲۰۰	۵۶٫۱۱	۱۰۰	۵۲٫۴۶	۵۰
FS	۵۷٫۹۳		۵۵٫۷۵		۵۱٫۸۹	
MCEC	**۶۰٫۰۳**		**۵۶٫۸۷**		**۵۳٫۰۷**	

جدول ۴: مقایسه کارایی روش پیشنهادی با سایر روش‌های انتخاب ویژگی بدون ناظر بر روی دسته‌بند SVM با استفاده از روش جاسازی NON-ZERO DCT COEFFICIENT با NSF ۵ : ۰/۳.

روش‌های انتخاب ویژگی	Evaluation criteria					
	SVM					
	دقت	تعداد ویژگی	دقت	تعداد ویژگی	دقت	تعداد ویژگی
DSFFC	۹۲٫۸۸		۸۹٫۷۵		۸۸٫۸۸	
RRFS	۹۱٫۶۸		۸۸٫۸۹		۸۹٫۲۳	
GCNC	۹۳٫۳۲	۵۰۰	۹۱٫۸۸	۴۰۰	۹۰٫۰۷	۳۰۰
FS	۹۲٫۴۸		۸۹٫۰۵		۸۸٫۸۴	
MCEC	**۹۶٫۶۵**		**۹۵٫۸۴**		**۹۲٫۶۸**	
DSFFC	۸۶٫۶۹		۷۵٫۶۳		۶۹٫۷۱	
RRFS	۸۵٫۶۰		۷۶٫۷۱		۷۰٫۷۱	
GCNC	۸۸٫۲۱	۲۰۰	۸۰٫۱۱	۱۰۰	۷۷٫۴۶	۵۰
FS	۸۵٫۰۳		۷۸٫۵۴		۷۳٫۹۱	
MCEC	**۸۸٫۵۲**		**۸۳٫۷۸**		**۷۸٫۵۲**	

جدول ۵: مقایسه کارایی روش پیشنهادی با سایر روش‌های انتخاب ویژگی بدون ناظر بر روی دسته‌بند NB. با استفاده از روش جاسازی NON-ZERO DCT COEFFICIENT با NSF ۵ : ۰/۱.

روش‌های انتخاب ویژگی	Evaluation criteria					
	Naive Bayes					
	دقت	تعداد ویژگی	دقت	تعداد ویژگی	دقت	تعداد ویژگی
DSFFC	۶۳٫۴۸		۶۱٫۶۶		۵۹٫۸۸	
RRFS	۶۳٫۰۵		۶۱٫۱۵		۵۹٫۲۳	
GCNC	۶۵٫۵۴	۵۰۰	۶۴٫۸۸	۴۰۰	۶۰٫۰۷	۳۰۰
FS	۶۵٫۶۲		۶۲٫۸۰		۶۰٫۹۰	
MCEC	**۶۶٫۲۹**		**۶۴٫۱۲**		**۶۳٫۷۱**	
DSFFC	۵۶٫۱۹		۵۳٫۰۳		۵۱٫۵۸	
RRFS	۵۶٫۹۱		۵۲٫۹۵		۵۲٫۲۵	
GCNC	۵۸٫۰۵	۲۰۰	۵۵٫۸۴	۱۰۰	۵۲٫۱۶	۵۰
FS	۵۸٫۶۹		۵۴٫۱۸		۵۱٫۸۹	
MCEC	**۵۹٫۸۵**		**۵۵٫۱۲**		**۵۲٫۵۵**	

صورت که هر روش بر روی هر دیتاست رتبه‌بندی می‌شود. در آزمون فریدمن فرض H_0 مبتنی بر یکسان‌بودن میانگین رتبه‌ها در بین گروه‌هاست. ردشدن فرض صفر به این معنی است که در بین گروه‌ها حداقل دو گروه با هم اختلاف معناداری دارند. در تحلیل نتایج آزمون فریدمن می‌توان گفت که چنانچه سطح معنی‌داری کمتر از میزان خطا باشد، وجود تفاوت بین حداقل یک زوج از نمونه‌ها استنباط می‌شود. از آنجا که این آزمون‌ها معمولاً در سطح خطای ۵٪ در نظر گرفته می‌شود، برای رسیدن به این نتیجه باید سطح معنی‌داری کمتر از ۰/۰۵ باشد. جداول ۷ و ۸ نتایج تست آماری فریدمن را برای روش ارائه‌شده در مقایسه با سایر روش‌های انتخاب ویژگی ذکرشده نمایش می‌دهد. در جدول ۷ میانگین رتبه‌های محاسبه‌شده برای هر روش توسط هر دسته‌بندی‌کننده نمایش داده شده که روش پیشنهادی دارای کمترین میانگین رتبه است.

روشی که بهترین کارایی را داشته باشد رتبه یک و دومین بهترین

جدول ۶: مقایسه کارایی روش پیشنهادی با سایر روش‌های انتخاب ویژگی بدون ناظر بر روی دسته‌بند NB. با استفاده از روش جاسازی NSF 5 : 0/3 NON-ZERO DCT COEFFICIENT.

روش‌های انتخاب ویژگی	Evaluation criteria					
	Naive Bayes					
	دقت	تعداد ویژگی	دقت	تعداد ویژگی	دقت	تعداد ویژگی
DSFFC	۹۰/۶۳		۸۶/۱۵		۸۷/۹۲	
RRFS	۹۱/۱۶		۸۷/۲۰		۸۷/۱۸	
GCNC	۹۲/۳۲	۵۰۰	۹۱/۸۸	۴۰۰	۸۸/۹۲	۳۰۰
FS	۹۰/۴۸		۸۹/۸۵		۸۶/۱۹	
MCEC	**۹۴/۰۲**		**۹۱/۷۶**		**۹۰/۵۵**	
DSFFC	۸۴/۰۸		۷۶/۸۰		۷۰/۷۶	
RRFS	۸۳/۸۲		۷۷/۴۱		۷۱/۵۸	
GCNC	۸۵/۷۶	۲۰۰	۷۹/۰۹	۱۰۰	۷۴/۶۶	۵۰
FS	۸۳/۴۵		۷۷/۵۴		۷۳/۳۲	
MCEC	**۸۶/۹۲**		**۸۰/۲۷**		**۷۵/۱۶**	

جدول ۷: میانگین رتبه‌های به دست آمده توسط آزمون فریدمن با استفاده از دسته‌بندهای SVM و NB.

method	Mean rank	
	SVM	NB
DSFFC	۴/۳۳	۴/۳۳
RRFS	۴	۴
GCNC	۲/۱۶	۲/۱۶
FS	۳/۴۱	۳/۳۳
MCEC	**۱**	**۱/۲۵**

کارایی رتبه ۲ و همین طور به تعداد روش‌ها رتبه خواهیم داشت. جدول ۸ نشان می‌دهد که آزمون فریدمن یک P_value به مقدار ۰/۰۰۱ برای دقت روش پیشنهادی و بر روی دسته‌بندی کننده SVM گزارش داده است. چون این مقدار کمتر از ۰/۰۵ است می‌توانیم ادعا کنیم که نتایج روش پیشنهادی دارای اختلاف معنی‌داری با سایر روش‌های بدون ناظر ذکرشده است. همچنین سایر P_valueهای به دست آمده توسط آزمون فریدمن برای دسته‌بندی‌کننده NB در جدول ۸ نیز نشان‌دهنده این ادعا است.

۵- نتیجه‌گیری و راهکارهای آتی

یکی از چالش‌های عمده در مسئله پنهان‌شکنی تصاویر تعداد زیاد ویژگی‌های استخراج‌شده برای این کار است. مجموعه‌های داده‌ای با ابعاد بالا از دو جهت باعث کاهش عملکرد پنهان‌شکنی می‌شود. از یک طرف با افزایش ابعاد داده‌ها، حجم محاسبات افزایش پیدا می‌کند و از طرف دیگر مدلی که بر اساس داده‌های با ابعاد بالا ساخته می‌شود دارای قابلیت تعمیم پایینی است و احتمال بیش‌برازش افزایش پیدا می‌کند.

در این مقاله تلاش شده که با استفاده از یک روش انتخاب ویژگی مبتنی بر خوشه‌بندی گراف، ویژگی‌های دارای افزونگی و یا نامرتبط حذف شده و در نتیجه دقت الگوریتم پنهان‌شکنی افزایش داده شود. این روش شامل ۳ مرحله بازنمایی گرافی مسئله، خوشه‌بندی ویژگی‌ها و جستجوی زیرمجموعه بهینه بر مبنای ارزش هر ویژگی است.

در روش پیشنهادی استفاده از مفهوم مسئله زیرگراف کامل بیشینه وزن‌دار و ترکیب آن با معیار مرکزیت یال سبب ارائه یک روش انتخاب ویژگی کارا شد که در هر دو حالت بدون ناظر و باناظر قادر به انتخاب زیرمجموعه بهینه بود. از آنجا که گراف ویژگی‌ها یک گراف وزن‌دار است و در روش پیشنهادی تمرکز بر روی روابط میان ویژگی‌هاست مسئله

جدول ۸: نتایج به دست آمده برای روش پیشنهادی توسط آزمون فریدمن.

Friedman test	SVM	NB
Chi-Square	۱۶/۶۵۰	۱۵/۶۰۰
df	۳	۳
Asymp.Sig.	۰/۰۰۱	۰/۰۰۱۳۶۹

زیرگراف کامل بیشینه وزن‌دار یک راهکار بسیار مناسب جهت خوشه‌بندی ویژگی‌هاست زیرا در یک زیرگراف کامل بیشینه مجموعه ویژگی‌ها کاملاً به هم مرتبط هستند و بیشترین شباهت را با یکدیگر دارند. در این روش پیشنهادی، پس از یافتن زیرگراف کامل بیشینه وزن‌دار در یک فرایند تکراری، با استفاده از معیار مرکزیت یال و در نظر گرفتن مناسب‌بودن هر ویژگی، ویژگی‌های تأثیرگذار و دارای حداقل افزونگی از هر زیرگراف کامل بیشینه شناسایی شده و به‌عنوان ویژگی‌های نهایی انتخاب می‌شوند. به عبارت دیگر در این روش، از هر خوشه ویژگی‌هایی که تأثیرگذاری بیشتری داشته به‌عنوان نماینده آن خوشه برای زیرمجموعه ویژگی نهایی انتخاب می‌شوند. برای محاسبه تأثیرگذاری هر ویژگی از میزان مناسب‌بودن آن ویژگی استفاده شد. در نتیجه، زیرمجموعه ویژگی نهایی، شامل ویژگی‌هایی است که به‌خوبی قادر به نمایندگی از ویژگی‌های اولیه هستند. در روش پیشنهادی، در فرایند جستجوی زیرمجموعه بهینه از هیچ الگوریتم یادگیری استفاده نمی‌شود. بنابراین این روش مبتنی بر راهکار فیلتر است و در مقایسه با سایر روش‌های انتخاب ویژگی دارای دقت بالاتری بوده و تنها مشکل روش بار محاسباتی پیداکردن زیرگراف‌های کامل است که چون عمل انتخاب ویژگی تنها یک بار در فرایند پنهان‌شکنی اعمال می‌شود با توجه به بهبود دقت نسبت به روش‌های دیگر قابل چشم‌پوشی است. از جمله راهکارهایی که برای آینده می‌توان پیشنهاد داد این که چگونه بار محاسباتی پیداکردن زیرگراف‌های کامل را کاهش داد و همچنین استفاده از دیگر مفاهیم تئوری گراف و الگوریتم‌های شبکه‌های اجتماعی مانند مرکزیت گره و تشخیص جوامع در روش‌های انتخاب ویژگی می‌باشد.

مراجع

[1] S. M. Badr, G. Ismaial, and A. H. Khalil, "A review on steganalysis techniques: from image format point of view," *International J. of Computer Applications*, vol. 102, no. 4, pp. 11-19, Sep. 2014.

[2] V. Bhasin, P. Bedi, and A. Singhal, "Feature selection for steganalysis based on modified stochastic diffusion search using

Fisher score," in *Proc. Int. Conf. on Advances in Computing, Communications and Informatics, ICACCI'14*, pp. 2323-2330, New Delhi, India, 24-27 Sept. 2014.
[3] Y. Miche, B. Roue, A. Lendasse, and P. Bas, "A feature selection methodology for steganalysis," in B. Gunsel, A. K. Jain, A. M. Tekalp, B. Sankur (Eds.) Multimedia Content Representation, *Classification and Security: Int. Workshop, MRCS 2006,* Springer, Berlin, Heidelberg, pp. 49-56, Sep. 2006.
[4] T. Pevn, P. Bas, and J. Fridrich, "Steganalysis by subtractive pixel adjacency matrix," *Trans. Info. For. Sec.*, vol. 5, no. 2, pp. 215-224, Jun. 2010.
[5] A. Westfeld and A. Pfitzmann, *Attacks on Steganographic Systems*, Information Hiding, 2000.
[6] J. Fridrich, M. Goljan, and R. Du, "Reliable detection of LSB steganography in color and grayscale images," in *Proc. of the Workshop on Multimedia and Security*, *MM&Sec'01*, pp. 27-30, Ottawa, ON, Canada, 5-5 Oct 2001.
[7] S. Dumitrescu, X. Wu, and Z. Wang, "Detection of LSB steganography via sample pair analysis," *Information Hiding*, vol. 51, no. 7, pp. 1995-2007, Jul. 2003.
[8] A. Westfeld, Detecting Low Embedding Rates," in: F. A. P. Petitcolas (Ed.) *Information Hiding: 5th International Workshop*, IH 2002, Springer Berlin Heidelberg, Berlin, 2003.
[9] S. M. S. Tarzjani and S. Ghaemmaghami, "Detection of LSB replacement and LSB matching steganography using Gray level run length matrix," in *Proc. of 5th Int. Conf. on Intelligent Information Hiding and Multimedia Signal Processing*, pp. 787-790, Kyoto, Japan, 12-14 Sept. 2009.
[10] M. Goljan, J. Fridrich, and T. Holotyak, "New blind steganalysis and its implications," *Proc. SPIE 6072, Security, Steganography, and Watermarking of Multimedia Contents VIII,* pp. 1-13, 2006.
[11] J. Harmsen and W. Pearlman, "Steganalysis of additive-noise modelable information hiding," in *Proc. SPIE Security Watermarking Multimedia Contents*, vol. 5020, pp. 131-142, 2003.
[12] Z. Zhang and E. R. Hancock, "Hypergraph based information-theoretic feature selection," *Pattern Recognition Letters*, vol. 33, no. 15, pp. 1991-1999, 1 Nov. 2012.
[13] P. Zhu, W. Zhu, Q. Hu, C. Zhang, and W. Zuo, "Subspace clustering guided unsupervised feature selection," *Pattern Recognition*, vol. 66, pp. 364-374, Jun. 2017.
[14] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226-1238, Aug. 2005.
[15] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," *Proc. of the 14th Int. Conf. on Neural Information Processing Systems: Natural and Synthetic, NIPS'01*, pp. 585-592, Vancouver, BC, Canada, 3-8 Dec. 2002.
[16] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Machine Intell*, vol. 22, no. 8, pp. 888-905, Aug. 2000.
[17] F. Chung, "Spectral graph theory," in Regional Conf. Series in Mathematics American Mathematical Society, vol. 92, pp. 1-212, 1997.
[18] Q. Gu, Z. Li, and J. Han, "Generalized fisher score for feature selection," in *Proc. of the Int. Conf. on Uncertainty in Artificial Intelligence*, pp. 266-273, Barcelona, Spain, 14-17 Jul. 2011.
[19] X. He, D. Cai, and P. Niyogi1, "Laplacian score for feature selection," in *Proc. of the 18th Int. Conf. on Neural Information Processing System*, *NIPS'05,* pp. 507-514, Vancouver, BC, Canada, 5-8 Dec. 2005.
[20] H. Cheng, W. Deng, C. Fu, Y. Wang, and Z. Qin, "Graph-based semi-supervised feature selection with application to automatic spam image identification," In: Yu Y., Yu Z., Zhao J. (eds) *Computer Science for Environmental Engineering and EcoInformatics. CSEEE 2011. Communications in Computer and Information Science,* vol. 159. Springer, Berlin, pp. 259-264, 2011.
[21] Q. Song, J. Ni, and G. Wang, "A fast clustering-based feature subset selection algorithm for high-dimensional data," *IEEE Trans. on Knowledge and Data Engineering*, vol. 25, no. 1, pp. 1-14, Jan. 2013.
[22] S. Bandyopadhyay, T. Bhadra, P. Mitra, and U. Maulik, "Integration of dense subgraph finding with feature clustering for unsupervised feature selection," *Pattern Recognition Letters*, vol. 40, pp. 104-112, 15 Apr. 2014.
[23] S. Tabakhi and P. Moradi, "Relevance-redundancy feature selection based on ant colony optimization," *Pattern Recognition*, vol. 48, no. 9, pp. 2798-2811, Sept. 2015.
[24] P. Moradi and M. Rostami, "A graph theoretic approach for unsupervised feature selection," *Engineering Applications of Artificial Intelligence*, vol. 44, pp. 33-45, Sept. 2015.
[25] J. Che, Y. Yang, L. Li, X. Bai, S. Zhang, and C. Deng, "Maximum relevance minimum common redundancy feature selection for nonlinear data," *Information Sciences*, vol. 409-410, pp. 68-86, Oct. 2017.
[26] M. Ghaemi and M. R. Feizi-Derakhshi, "Feature selection using forest optimization algorithm," *Pattern Recognition*, vol. 60, pp. 121-129, Dec. 2016.
[27] A. K. Farahat, A. Ghodsi, and M. S. Kamel, "Efficient greedy feature selection for unsupervised learning," *Knowledge and Information Systems*, vol. 35, no. 2, pp. 285-310, May 2013.
[28] A. J. Ferreira and M. A. T. Figueiredo, "An unsupervised approach to feature discretization and selection," *Pattern Recognition*, vol. 45, no. 9, pp. 3048-3060, Sept. 2012.
[29] Q. Wu, J. K. Hao, and F. Glover, "Multi-neighborhood tabu search for the maximum weight clique problem," *Ann Oper Res*, vol. 196, no. 1, pp. 611-634, Jul. 2012.
[30] U. Brandes, "On variants of shortest-path betweenness centrality and their generic computation," *Social Networks*, vol. 30, no. 2, pp. 136-145, May 2008.
[31] M. Mandal and A. Mukhopadhyay, "Unsupervised non-redundant feature selection: a graph-theoretic approach," in *Proc. of the Int. Conf. on Frontiers of Intelligent Computing: Theory and Applications, FICTA'13*, pp. 373-380, 2013.
[32] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, The WEKA data mining software, Available from: <http://www.cs.waikato.ac.nz/ml/weka>.
[33] J. H. J. Xiao, K. Ehinger, A. Oliva, and A. Torralba, "SUN database: large-scale scene recognition from abbey to zoo," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 3485 -3492, San Francisco, CA, USA , 13-18 Jun. 2010.
[34] M. Friedman, "A comparison of alternative tests of significance for the problem of m rankings," *Annals of Math. Statistics*, vol. 11, no. 1, pp. 86-92, 1940.

سعید آزادی‌فر در سال ۱۳۹۱ مدرک کارشناسی مهندسی کامپیوتر خود را از دانشگاه رازی کرمانشاه و در سال ۱۳۹۴ مدرک کارشناسی ارشد مهندسی کامپیوتر خود را از دانشگاه اصفهان دریافت نمود. از سال ۱۳۹۵ نام‌برده به عنوان دانشجوی دکتری مهندسی کامپیوتر گرایش هوش مصنوعی در دانشگاه صنعتی خواجه نصیرالدین طوسی مشغول به تحصیل است همچنین از سال ۱۳۹۶ به عنوان مدرس با دانشگاه آزاد اسلامی واحد تهران جنوب همکاری می‌نماید. زمینه‌های تحقیقاتی مورد علاقه ایشان عبارتند از: داده‌کاوی، بیوانفورماتیک، یادگیری ماشین و پردازش تصویر.

سیدحسین خواسته در سال ۱۳۸۳ مدرک کارشناسی مهندسی برق خود را از دانشگاه صنعتی شریف و در سال‌های ۱۳۸۵ و ۱۳۹۱ مدرک کارشناسی ارشد و دکترای مهندسی کامپیوتر خود را از دانشگاه صنعتی شریف یافت نمود. دکتر خواسته از سال ۱۳۸۷ در دانشکده مهندسی کامپیوتر دانشگاه صنعتی خواجه نصیرالدین طوسی در تهران مشغول به فعالیت گردید و اینک نیز عضو هیأت علمی این دانشکده می‌باشد. زمینه‌های تحقیقاتی مورد علاقه نام‌برده متنوع بوده و شامل موضوعاتی مانند یادگیری ماشین، تحلیل و داده‌کاوی در داده‌های حجیم، سیستم‌های مبتنی بر زنجیره بلوکی، سیستم‌های توزیعی و رباتیک می‌باشد.

محمدهادی ادریسی در سال ۱۳۹۲ مدرک کارشناسی مهندسی کامپیوتر خود را از دانشگاه شیراز و در سال ۱۳۹۴ مدرک کارشناسی ارشد مهندسی کامپیوتر گرایش هوش مصنوعی خود را از دانشگاه اصفهان دریافت نمود. زمینه‌های تحقیقاتی مورد علاقه ایشان عبارتند از: یادگیری ماشین، بینایی ماشین و پردازش تصویر.