دسته‌بندی کدگذارهای تصویر در بستر شبکه‌های مخابراتی

مهدی تیموری[۱]، استادیار؛ نفیسه حسینی[۲]، کارشناسی ارشد

۱- دانشکده علوم و فنون نوین- دانشگاه تهران- تهران- ایران- mehditeimouri@ut.ac.ir
۲- دانشکده علوم و فنون نوین- دانشگاه تهران- تهران- ایران- nafiseh_hoseini@ut.ac.ir

چکیده: امروزه با گسترش شبکه‌های مخابراتی، درخواست ارسال داده‌های چندرسانه‌ای به‌طور محسوسی افزایش یافته است؛ بنابراین اطلاع از نوع داده ارسال‌شده به‌منظور کنترل ارتباطات و جلوگیری از انتقال داده‌های مخرب موضوع مهمی است. یک سیستم شناسایی نوعی برای شناسایی نوع داده کدشده ارسالی معمولاً از دسته‌بندی در میان مجموعه مشخصی از کدگذارها استفاده می‌کند. این نوع دسته‌بندی معمولاً بر اساس ویژگی‌های مرتبط استخراج‌شده از جریان بیتی دریافتی صورت می‌پذیرد. بیشتر پژوهش‌های موجود تعداد کمی کدگذار تصویر را در مسئلت دسته‌بندی خود وارد کرده‌اند. در این مقاله روشی کارا برای دسته‌بندی بین ده نوع کدگذار تصویر مختلف پیشنهاد شده است. این روش بر مبنای ترکیب و توسعه روش‌های موجود پیشنهاد شده است. بر طبق نتایج شبیه‌سازی، سیستم پیشنهادی ده نوع کدگذار تصویر مختلف را با متوسط صحت ۸۸/۹۰ درصد به‌درستی از یکدیگر تشخیص می‌دهد. در میان این کدک‌ها، GIF با دقت ۹۹/۳ درصد و BMP با دقت ۹۲/۵ درصد دارای بالاترین درصد تشخیص درست از دیگر کدگذارهای تصویری هستند. از طرف دیگر FLIF و WEBP به‌ترتیب با دقت ۸۳/۳ و ۸۳/۶ کمترین دقت تشخیص درست را دارند.

واژه‌های کلیدی: کدگذارهای تصویر، دسته‌بندی، شناسایی نوع فایل.

Classification of Image Codecs in Telecommunication Networks

Mehdi Teimouri[1], Assistant Professor; Nafiseh Hoseini[2], M.Sc. Graduated Student

1- Faculty of New Science and Technology, University of Tehran, Tehran, Iran, Email: mehditeimouri@ut.ac.ir
2- Faculty of New Science and Technology, University of Tehran, Tehran, Iran, Email: nafiseh_hoseini@ut.ac.ir

Abstract: Nowadays, with the spread of communication networks, the demand to transmit multimedia data has significantly increased. So, the knowledge about data type which is transmitted through the network is an important issue for monitoring communications and preventing transmission of malicious data. A typical identification system attempts to identify the type of transmitted coded data through classification within a predefined set. The classification is usually based on some relevant features extracted from the received bit stream. Most of the researches in this field consider a few kinds of image codec in their classification problem. In this paper, an efficient identification system is proposed for classification within ten different images codecs. The proposed system is based on combination and extension of existing methods. According to simulation results, image codecs are classified with average accuracy of 88.90%. Among various codecs, GIF and BMP have the highest accuracy of 99.3% and 92.5%, respectively. On the other hand, FLIF and WEBP have the lowest accuracy 83.3% and 83.6%, respectively.

Keywords: Image codecs, classification, file type identification.

تاریخ ارسال مقاله: ۱۳۹۶/۰۲/۲۵
تاریخ اصلاح مقاله: ۱۳۹۶/۰۷/۲۱ و ۱۳۹۶/۰۹/۲۰
تاریخ پذیرش مقاله: ۱۳۹۶/۱۰/۰۱
نام نویسنده مسئول: مهدی تیموری
نشانی نویسنده مسئول: ایران – تهران – خیابان کارگر شمالی – دانشکده علوم و فنون نوین دانشگاه تهران.

۱ -مقدمه

شناسایی کور شبکه‌های مخابراتی یکی از حوزه‌های تحقیقاتی است که در سال‌های اخیر توجه زیادی به آن شده است [۱، ۲]. با گسترش دنیای ارتباطات کامپیوتری و نیاز به ارسال انواع فایل‌ها از طریق شبکه‌های ارتباطی، موضوع شناسایی و دسته‌بندی نوع بسته‌های اطلاعاتی مبادله شده در شبکه‌ها توجه بسیاری از محققان را به خود معطوف کرده است. جریان بیت‌های ارسالی در بستر شبکه ارتباطی ممکن است حاوی قسمتی از یک فایل متنی، تصویر، صوت و یا یک فایل ترکیبی نظیر پاورپوینت یا هر چیز دیگری باشد. بدیهی است که نظارت بر جریان داده‌های مبادله شده و شناسایی نوع فایل‌های ارسال‌شده جهت کنترل ارتباطات در شبکه‌های مخابراتی امری ضروری است. در حقیقت، عدم آگاهی از داده‌های ارسالی در بستر شبکه‌های مخابراتی پیامدهای بسیاری همچون توزیع داده‌های مخرب و ایجاد اختلالات را در بر خواهد داشت [۳].

یکی از انواع داده ارسالی در شبکه‌های مخابراتی داده‌های تصویری هستند. از آنجا که این داده‌ها با استفاده از کدگذارهای مختلفی مانند JPEG، BMP و غیره کد می‌شوند، قبل از رسیدن به محتوای تصویری این داده‌ها نیاز است که ابتدا نوع کدگذار به‌کاررفته در کدگذاری داده تعیین شود. چنین مسئله‌ای را معمولاً در قالب یک مسئله دسته‌بندی[۱] تعریف می‌کنند و شناسایی[۲] در حقیقت به معنی دسته‌بندی داده‌های دریافتی در میان تعداد مشخصی کدگذار است.

به‌طورکلی می‌توان تحقیقات موجود در زمینه دسته‌بندی و شناسایی نوع داده را به سه دسته مختلف دسته‌بندی بر پایه پسوند فایل، دسته‌بندی بر پایه امضای فایل [۴] و دسته‌بندی بر پایه تحلیل محتوای فایل [۱۳-۵]، تقسیم کرد. شناسایی بر مبنای پسوند عملاً در شبکه‌های مخابراتی کاربرد ندارد. علاوه بر این، شناسایی از روی امضای فایل با تطبیق بایت‌های ویژه که در سرآیند یا پایان دهنده هر فایل وجود دارد، انجام می‌پذیرد که کاربردی بودن چنین روشی را برای داده‌های دریافتی در بستر شبکه‌های مخابراتی زیر سؤال می‌برد. در روش‌های دسته‌بندی مبتنی بر محتوا از ویژگی‌های آماری مختلف استخراج‌شده از داده استفاده می‌گردد. این روش‌ها نیاز به داشتن کل فایل یا داده ندارند و صرفاً با داشتن بخشی از داده می‌توانند مورداستفاده قرار گیرند. چنین روش‌هایی مناسب استفاده در شبکه‌های مخابراتی می‌باشند.

در هیچ‌یک از تحقیقات پیشین حوزه شناسایی قطعات فایل، دسته‌بندی مخصوص برای کدگذارهای تصویری انجام نشده است. در حقیقت، بیشتر این تحقیقات تعداد کمی کدگذار تصویر را در کنار انواع داده‌های دیگر مانند فایل‌های متنی و غیره دسته‌بندی می‌نمایند. در این مقاله، به‌طور ویژه موضوع دسته‌بندی کدگذارهای تصویری را مورد بررسی قرار می‌دهیم. در این راستا، روشی کارا برای دسته‌بندی بین ده نوع کدگذار تصویر مختلف پیشنهاد شده است. همان‌طور که خواهیم دید، این روش که بر مبنای ترکیب و توسعه روش‌های موجود پیشنهاد شده است دارای عملکرد بسیار مناسبی است.

ساختار این مقاله در ادامه به این شرح است. در بخش دوم، مدل سیستم و تعریف دقیق مسئله ارائه می‌شود. در بخش سوم، تحقیقات پیشین ارائه شده و مورد بررسی قرار می‌گیرد. در بخش چهارم، روش پیشنهادی برای دسته‌بندی کدگذارهای تصویر ارائه می‌شود. روش پیشنهادی در این بخش بر مبنای شبیه‌سازی روش‌های پیشین ارائه می‌گردد؛ بنابراین در این بخش، نتیجه شبیه‌سازی روش پیشنهادی با روش‌های موجود مورد مقایسه و بررسی قرار می‌گیرد. در انتها و در بخش پنجم، نتیجه‌گیری ارائه می‌گردد.

۲ -مدل سیستم

یک شبکه ارسال تصویر را به‌صورت آنچه در شکل ۱ نمایش داده شده است در نظر بگیرید. همان‌طور که ملاحظه می‌شود، تعدادی فایل تصویری که با کدگذارهای متفاوت تصویر کدگذاری شده‌اند برای ارسال در شبکه آماده شده‌اند. هر یک این فایل‌های تصویری ممکن است با استفاده از کدگذار منحصر به خود نظیر PNG، GIF، JPEG و غیره کدگذاری شده باشند. هر فایل تصویر در قالب تعدادی بسته بایتی ارسال می‌شود. به‌عنوان مثال یک تصویر ورودی را که با کدگذار JPEG فشرده‌شده است در نظر بگیرید. این تصویر به k بسته n بایتی $J1,J2,\ldots Jk$ تبدیل‌شده و در شبکه ارسال می‌گردد. در حالت کلی، بسته‌های تصویری کدگذارهای مختلف به‌صورت غیرمنظم در بستر شبکه دریافت خواهند شد. در این حالت ممکن است قطعات بسته‌های تصویری ارسالی به‌طور متوالی قرار نگیرند و میان قطعات هر فایل قطعه‌ای از فایل دیگری که توسط همان کدگذار یا کدگذارهای دیگر کد شده است قرار بگیرد. به‌عنوان مثال، مطابق شکل ۱ فرض کنید جریان بسته‌های ارسال‌شده در شبکه به‌صورت...J1,P1,G1, باشد که در آن J1 بسته اول فایل JPEG1 و P1 بسته اول از فایل PNG1 و G1 بسته اول از فایل GIF1 می‌باشند. این جریان بایتی بسته‌ها توسط یک سیستم شناسایی دریافت شده و به‌منظور دسته‌بندی مورد تحلیل قرار می‌گیرند. سیستم شناسایی (دسته‌بندی) تصویر باید با استفاده از استخراج ویژگی‌های مناسب از جریان بایتی بسته‌های تصویری دریافت شده قادر به تشخیص نوع کدگذار هر بسته باشد.

مدل شکل ۱ یک مدل ساده‌شده از سیستم واقعی مراقبت (یا شنود) است که در آن ابتدا شکل موج فرکانس رادیویی توسط گیرنده دریافت، تحلیل و دمدوله می‌شود. همچنین اضافاتی مانند سرآیند بسته‌ها از داده جدا می‌شوند و در انتها داده‌های خام به دسته‌بند تحویل داده می‌شوند. به‌عبارت دیگر، فرض می‌کنیم سربارهای لایه‌های قبل از روی داده‌ها برداشته شده و با داده‌های تصویری خالص روبرو هستیم. علاوه بر این، فرض می‌کنیم خطاهای احتمالی در داده‌های دریافتی توسط کد کانال تصحیح شده است.

یک گام مهم برای شبیه‌سازی و بررسی عملکرد چنین سیستمی، جمع‌آوری مجموعه داده‌ای مناسب است. در این راستا، ابتدا ۱۰۰ تصویر خام که با استفاده از دوربین‌های حرفه‌ای و باکیفیت بالا گرفته

شده‌اند، جمع‌آوری شده‌اند. پس از جمع‌آوری مجموعه تصاویر خام، الگوریتم‌های متفاوت کدگذاری و فشرده‌سازی تصویر بر روی هر یک از این تصاویر اعمال شده است. این الگوریتم‌ها شامل ۱۰ نوع کدک متفاوت JPEG، PNG، GIF، BMP، TIFF، FLIF، BPG، WEBP، JPEGXR (به‌اختصار JXR) و JPEG2000 (به‌اختصار J2K) می‌باشد. از هر فایل تصویری کدشده شش بسته تصادفی بدون هم‌پوشانی و به طول ۱۰۲۴ بایت انتخاب می‌کنیم؛ بنابراین، در مجموع ۶۰۰۰ بسته ۱۰۲۴ بایتی داریم که به‌شکل یکنواخت توسط یکی از ۱۰ کدگذار مختلف کد شده‌اند. در ضمن، قبل از استخراج بسته‌ها، از هر فایل تصویری معادل یک‌هشتم طول آن را از ابتدا و انتهای فایل حذف می‌کنیم تا مطمئن باشیم دادهٔ تولیدشده شامل سرآیند و پایان دهنده نخواهد بود.

با استخراج تعدادی ویژگی مناسب، یک مجموعه داده شامل ۶۰۰۰ نمونه خواهیم داشت (در بخش بعد در مورد ویژگی‌ها بحث خواهیم کرد). هر نمونه از این مجموعه متناظر با یکی از ۶۰۰۰ بسته تولیدی است. ذکر این نکته ضروری است که برای استخراج ویژگی‌های طولانی‌ترین زیرتوالی و زیررشته مشترک، از قبل به‌ازای هر کدگذار ۵ فایل (معادل با ۳۰ بسته) مجزا و جدا از مجموعه داده به‌عنوان نماینده تولید نموده‌ایم. نیمی از مجموعه داده به‌عنوان مجموعه آموزش[۳] و نیمی دیگر به‌عنوان مجموعه آزمون[۴] در نظر گرفته می‌شوند (این انتخاب به‌صورت تصادفی انجام شده است). از مجموعه آموزش برای به‌دست آوردن بهترین پارامترهای دسته‌بندی با اعتبارسنجی متقابل[۵] ۱۰-فولد[۶] استفاده شده است. دسته‌بند مورداستفاده ماشین بردار پشتیبان (SVM) با هسته رادیال است. جهت مقایسه، از درخت تصمیم[۷] نیز به‌عنوان یک گزینه مناسب برای دسته‌بند استفاده خواهیم نمود. پس از استخراج پارامترهای مناسب دسته‌بند، با استفاده از کل داده آموزش دسته‌بند را آموزش می‌دهیم. مجموعه آزمون نیز به‌هیچ‌عنوان برای به‌دست آوردن پارامترها و ساخت مدل اولیه استفاده نشده و از آن صرفاً برای آزمایش دسته‌بند آموزش‌دیده استفاده می‌شود.

۳ -تحقیقات پیشین

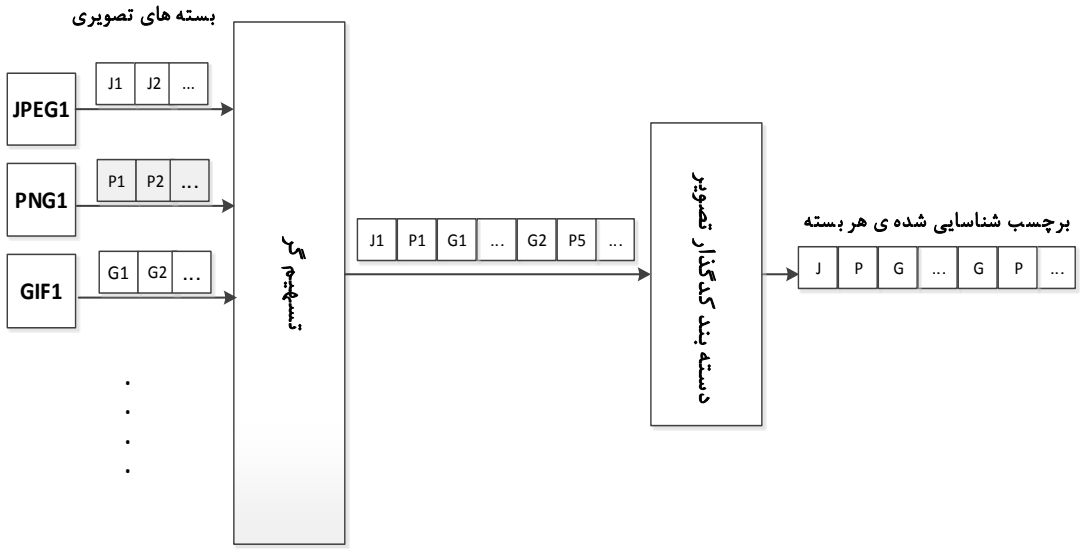
برای استخراج ویژگی گزینه‌های زیادی داریم. یکی از این ویژگی‌ها آنتروپی می‌باشد. آنتروپی یک تصویر متوسط اطلاعات موجود در آن تصویر به‌ازای هر بایت است. کلهون[۸] و کولز[۹] در سال ۲۰۰۸ شناسایی دوبه‌دوی تعدادی از تصاویر و فایل‌ها را مورد بررسی قراردادند. آن‌ها از آنتروپی به‌عنوان یک ویژگی آماری استفاده کردند و با استفاده از آن تصاویر با قالب JPEG و BMP را با بالاترین صحت (۸۲ درصد) از هم تفکیک کردند [۱۱]. در سال ۲۰۱۳ نیز بیب[۱۰] و همکاران برای دسته‌بندی بین ۳۰ نوع فرمت فایل متفاوت از ویژگی آنتروپی به‌عنوان یکی از ویژگی‌ها استفاده کردند [۱۴].

دو ویژگی از ویژگی‌های آماری دیگر میانگین و نما (مد) است. ویژگی‌های میانگین همبستگی بایت[۱۱] و انحراف مطلق میانگین[۱۲] (MAD) نیز دو ویژگی مهم در شناسایی هستند. شاخص پراکندگی واریانس (یا انحراف معیار) نیز از ویژگی‌های مهم در تشخیص است. شاخص آماری خودهمبستگی برای یافتن الگوهای تکراری به‌کار می‌رود. کلهون و کولز از همبستگی متقابل[۱۳] بین بایت‌های تشکیل‌دهنده نوع فایل‌ها به نتایج جالبی رسیدند. آن‌ها با محاسبه همبستگی بین مقادیر بایت‌های m ام و $m+1$ ام با صحت ۸۶ درصد نوع فایل‌های JPEG و GIF و با صحت ۸۱ درصد نوع فایل‌های JPEG و BMP را از هم تفکیک کردند [۱۱].

یکی دیگر از ویژگی‌های آماری که در شناسایی نوع فایل‌ها به‌کاربرده می‌شود کدهای اسکی است و معمولاً به سه حالت کد اسکی سطح پایین، کد اسکی سطح متوسط و کد اسکی سطح بالا به‌کار می‌رود [۱۱, ۱۴]. کلهون و کولز از این سه ویژگی در بردار ویژگی‌های خود استفاده نمودند و توانستند دو نوع فایل PDF و JPEG و همچنین PDF و GIF را به کمک کد اسکی سطح متوسط و با صحت ۸۶ درصد از هم تشخیص دهند. همچنین با کمک کد اسکی سطح پایین میزان تشخیص زوج فایل‌های PDF و JPEG از یکدیگر به ۹۹ درصد رسید. آن‌ها همچنین در تشخیص زوج کدک‌های تصویری JPEG و GIF به کمک کد اسکی سطح پایین، به صحت ۸۹ درصد رسیدند [۱۱]. بیب و همکاران نیز برای دسته‌بندی بین ۳۰ نوع فایل متفاوت از این ویژگی‌ها استفاده نمودند [۱۴].

دو ویژگی مورداستفاده دیگر ویژگی‌های طولانی‌ترین زیررشته مشترک[۱۴] و طولانی‌ترین زیرتوالی مشترک[۱۵] است. کلهون و کولز با استفاده از طولانی‌ترین زیرتوالی مشترک زوج فایل‌های JPEG و PDF را با صحت ۹۴ درصد و با استفاده از طولانی‌ترین زیررشته مشترک همین زوج فایل را با صحت ۹۳ درصد از یکدیگر تفکیک نمودند [۱۱]. فاصله همینگ[۱۶] و وزن همینگ[۱۷] ازجمله ویژگی‌هایی هستند که روی مجموعه‌ای از بیت‌ها اعمال می‌شوند. فاصله همینگ بین دو رشته بیت که دارای طول برابر هستند معادل با تعداد بیت‌های متفاوت در آن دو رشته است [۱۵]. وزن همینگ یک رشته نیز به تعداد یک‌های موجود در آن رشته گفته می‌شود. یکی از ویژگی‌های آماری مورداستفاده در شناسایی نوع فایل‌ها کشیدگی[۱۸] یا برجستگی می‌باشد. کشیدگی معیاری از تیزی منحنی توزیع بایت‌ها حول نقطه حداکثر یا به‌عبارت‌دیگر میزان ارتفاع منحنی توزیع است و برابر گشتاور چهارم نرمال شده می‌باشد [۱۶, ۱۷]. چولگی[۱۹] ویژگی آماری دیگری است که بیانگر میزان تقارن منحنی توزیع احتمالی بایت‌ها حول میانگین و برابر گشتاور سوم نرمال شده می‌باشد. در حقیقت چولگی بیانگر وجود یا عدم وجود تقارن در تابع توزیع بایت‌ها است. اگر بایت‌های تشکیل‌دهنده یک بسته نسبت به میانگین کاملاً متقارن باشند چولگی

برابر با صفر خواهد شد. اگر چولگی مقداری مثبت داشت توزیع بایت‌ها نامتقارن بوده و به سمت مقادیر بالاتر متمایل است و اگر چولگی دارای مقدار منفی باشد توزیع بایت‌ها نامتقارن بوده و به سمت مقادیر کوچک‌تر متمایل می‌باشند [۱۸].



شکل ۱: مدل کلی شناسایی و دسته‌بندی کدگذارهای تصویر در شبکه‌های مخابراتی.

یکی از ویژگی‌های آماری مورداستفاده در شناسایی و دسته‌بندی نوع فایل‌ها n-گرم‌ها می‌باشند. هر n-گرم زیررشته‌ای به طول n از یک‌رشته بیت یا بایت است. فراوانی یک n-گرم در مجموعه‌ای از بیت‌ها یا بایت‌ها ویژگی مناسبی است که توجه بسیاری از محققان را در شناسایی و دسته‌بندی فایل‌ها به خود جلب کرده است. از انواع n گرم‌ها می‌توان به ۱-گرم[۲۰]، ۲-گرم[۲۱]، ۳-گرم[۲۲] و... اشاره کرد. به‌عنوان‌مثال در ۱-گرم میان بیت‌های تشکیل‌دهنده یک نوع فایل مقدار فراوانی بیت‌های ۰ و ۱ را محاسبه کرده و تعداد رخداد آن‌ها را نسبت به تعداد کل بیت‌ها به‌دست می‌آوریم. به همین ترتیب در ۲-گرم به دنبال فراوانی مقادیر تمامی زیررشته‌های بیتی به طول ۲ یعنی ۰۰ و ۰۱ و ۱۰ و ۱۱ در مجموعه بیت‌ها خواهیم بود از n-گرم‌ها در حوزه بایت نیز استفاده می‌کنند و به آن توزیع فراوانی بایت[۲۳] می‌گویند به‌عنوان مثال، ۱-گرم در بایت‌ها یک بردار ۲۵۶ تایی است که فراوانی اعداد ۰ تا ۲۵۵ را در مقادیر بایت‌ها به‌دست می‌آورد. بسیاری از افراد در حوزه شناسایی فایل از این روش استفاده کرده‌اند. در سال ۲۰۰۵ وانگ[۲۴] و همکاران از تحلیل n-گرم‌ها برای شناسایی نوع فایل‌ها استفاده کردند و به نتایج قابل قبولی دست یافتند [۱۹]. بیب و همکاران نیز از ترکیب ۱-گرم و ۲-گرم در بردار ویژگی‌های خود استفاده نمود [۱۴]. دیواکاران[۲۵] و همکاران در سال ۲۰۱۶ نیز استفاده از n-گرم‌ها را مدنظر قراردادند. بااین‌حال، برای جلوگیری از بزرگ شدن بردار ویژگی، تنها از ۱-گرم در بردار ویژگی‌هایشان استفاده نمود [۲۰].

۴ نتایج شبیه‌سازی و روش پیشنهادی

همان‌طور که در بخش قبل اشاره شد کلهون و کولز دسته‌بندی دوبه‌دوی فایل‌ها و تصاویر را مورد بررسی قراردادند. آن‌ها هرکدام از ویژگی‌های آنتروپی، همبستگی، انحراف معیار، کد اسکی سطح متوسط، سطح بالا و سطح پایین، طولانی‌ترین زیررشته مشترک و طولانی‌ترین زیرتوالی مشترک را به‌صورت مجزا بر مجموعه داده‌های خود اعمال کرده و مقایسه‌ای دودویی بین ۴ نوع فایل JPEG، PDF، BMP و GIF انجام دادند [۱۱]. همان‌گونه که ملاحظه می‌شود، آن‌ها تنها ۳ نوع از کدک‌های تصویر را و آن‌هم به‌صورت دوبه‌دو از یکدیگر تفکیک می‌نمایند. بیب و همکاران نیز برای دسته‌بندی بین ۳۰ نوع فایل متفاوت از ویژگی‌های ۱-گرم، ۲-گرم، میانگین، آنتروپی، انحراف معیار، انحراف مطلق میانگین، کشیدگی، چولگی، میانگین همبستگی بایت، کد اسکی سطح پایین، بالا و متوسط استفاده کردند. آن‌ها در مجموعه فایل‌های خود ۵ نوع فایل تصویری JPEG، GIF، BMP، PNG و TIFF را مدنظر قراردادند [۱۴]. دیواکاران و همکاران نیز در سال ۲۰۱۶ برای دسته‌بندی بین ۹ نوع فایل متفاوت، ویژگی‌های ۱-گرم، میانگین، آنتروپی، انحراف معیار، کشیدگی و چولگی را بر روی مجموعه داده‌های خود اعمال کردند [۲۰]. در روش مورداستفاده توسط آن‌ها تنها دو نوع فایل تصویری JPEG و PNG در میان ۹ نوع فایل وجود داشت. در این بخش، روش استخراج مشخصه خود را که ترکیب و توسعه‌ای بر دو روش دیواکاران و بیب است معرفی می‌نماییم و عملکرد آن را با هر یک از این دو روش مقایسه خواهیم نمود.

از آنجا که روش پیشنهادی ترکیب و توسعه‌ای بر دو روش دیواکاران و بیب است، در ابتدا لازم است نحوه عملکرد این دو روش موجود را مورد بررسی و تحلیل قرار دهیم. در ادامه، با رفع ضعف‌های

این دو روش و ترکیب آن‌ها روش پیشنهادی خود را ارائه خواهیم داد. با استفاده از روش استخراج مشخصه دیواکاران، متوسط صحت دسته‌بندی ۱۰ نوع داده تصویری برای درخت تصمیم برابر ۸۱/۷ درصد و برای ماشین بردار پشتیبان برابر ۷۵/۰ درصد می‌باشد. در جدول ۱، ماتریس درهم‌ریختگی[۲۶] روش دیواکاران با استفاده از دسته‌بند درخت تصمیم نمایش داده شده است. در این ماتریس هر سطر متناظر با یکی از کدک‌های ورودی دسته‌بند و هر ستون متناظر با یکی از تشخیص‌های محتمل دسته‌بند است.

با استفاده از روش استخراج مشخصه بیب، متوسط صحت دسته‌بندی ۱۰ نوع داده تصویری برای درخت تصمیم برابر ۸۲/۵ درصد و برای ماشین بردار پشتیبان برابر ۷۷/۳ درصد می‌باشد. در جدول ۲ ماتریس درهم‌ریختگی روش بیب با استفاده از دسته‌بند درخت تصمیم نمایش داده شده است. با مقایسه ماتریس‌های درهم‌ریختگی دو روش می‌توان گفت که هرچند متوسط صحت در روش بیب به ۸۲/۵ افزایش پیدا کرده است، اما روش دیواکاران در شناسایی برخی از انواع کدک مانند JPEG یا BPG بهتر عمل می‌کند.

با ملاحظه و مقایسه نتیجه دو روش بیب و دیواکاران می‌توان دید که در میان داده‌های تصویری موجود، تشخیص اشتباه بین دو نوع کدک TIFF و BMP زیاد رخ می‌دهد. ازاین‌روی اگر بتوانیم ویژگی‌های جدیدی به مجموعه ویژگی‌ها بیفزاییم که توسط آن‌ها امکان تشخیص دقیق‌تری بین دو نوع کدک TIFF و BMP داشته باشیم، می‌توان درصد تفکیک را تا حدودی بهبود داد. دسته‌ای از ویژگی‌های مناسب برای این کار ویژگی‌های طولانی‌ترین زیرتوالی و زیررشته مشترک هستند. با افزودن این ویژگی‌ها می‌توان تا حد زیادی اشتباه بین سایر کلاس‌ها را نیز کاهش داد. برای هر نمونه طولانی‌ترین زیررشته مشترک و طولانی‌ترین زیر ترتیب مشترک با هر یک از ۳۰ نماینده‌ها هر کلاس کدک محاسبه می‌گردد. سپس میانگین نرمال‌شده مقدار طولانی‌ترین زیررشته مشترک بسته موردنظر با تمام ۳۰ نماینده هر کدگذار را به‌عنوان یک ویژگی در نظر می‌گیریم. همین کار را برای طولانی‌ترین زیرتوالی مشترک انجام می‌دهیم. با این کار برای هر بسته ۲۰ ویژگی جدید به‌دست می‌آید.

با استفاده از اجتماع مجموعه ویژگی‌های بیب و دیواکاران به همراه ۲۰ ویژگی پیشنهادی طولانی‌ترین زیررشته و زیرتوالی مشترک متوسط صحت دسته‌بندی ۱۰ نوع داده تصویری با بهره گرفتن از درخت تصمیم ۸۷/۸ درصد و با استفاده از ماشین بردار پشتیبان ۸۴/۰ درصد به‌دست می‌آید. در جدول ۳ ماتریس درهم‌ریختگی روش پیشنهادی (که آن را روش پیشنهادی اول می‌نامیم) با استفاده از دسته‌بند درخت تصمیم نمایش داده شده است.

جدول ۱: ماتریس درهم‌ریختگی روش دیواکاران با استفاده از دسته‌بند درخت تصمیم

نوع تصویر	JPEG	BPG	TIFF	WEBP	BMP	PNG	FLIF	GIF	J2K	JXR
JPEG	۸۲/۱	۱/۳	۰	۳/۶	۰	۳/۶	۲/۳	۱/۳	۲/۶	۳
BPG	۲/۳	۸۱/۶	۰	۳/۳	۰	۱/۶	۳/۳	۱	۳/۳	۰
TIFF	۰	۰	۸۳/۶	۰	۱۶/۳	.	۰	.	۰	۰
WEBP	۱/۳	۵/۶	۰	۷۳/۶	۰	۲	۷/۳	۰	۷/۶	۲/۳
BMP	۰	۰	۱۲/۳	۰	۸۷/۶	۰	۰	۰	۰	۰
PNG	۱/۶	۴/۳	۰	۲	۰	۸۱/۳	۱/۳	۰/۶	۳/۶	۵
FLIF	۲/۶	۸/۶	۰	۸/۶	۰	۰	۶۹/۱	۰	۸/۳	۶
GIF	۰	۰	۰	۰	۰	۰/۶	۰	۹۷/۷	۰	۱/۶
J2K	۰/۳	۲/۳	۰	۵/۶	۰	۰	۶	۰/۶	۸۳/۴	۱/۶
JXR	۴	۰/۳	۰	۴/۳	۰	۳/۳	۱	۴/۶	۵	۷۶/۵

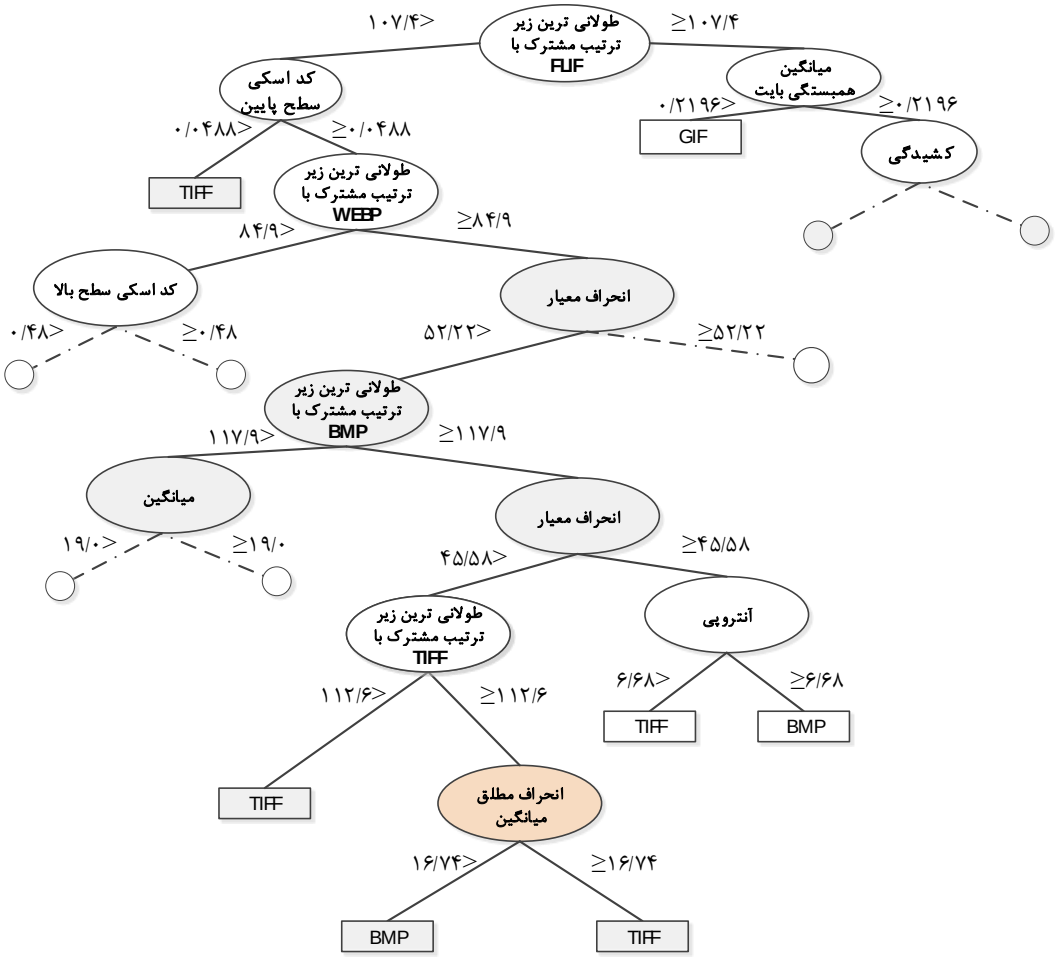
جدول ۲: ماتریس درهم‌ریختگی روش بیب با استفاده از دسته‌بند درخت تصمیم

نوع تصویر	JPEG	BPG	TIFF	WEBP	BMP	PNG	FLIF	GIF	J2K	JXR
JPEG	۸۱/۷	۳/۳	۰	۰	۰	۲/۶	۳/۳	۰/۶	۳/۳	۵
BPG	۰/۶	۸۰/۴	۰	۴	۰	۴	۱/۳	۰/۳	۵/۶	۱/۶
TIFF	۰	۰	۸۳/۸	۰	۱۶	۰/۳	۰	.	۰	۰
WEBP	۲	۶/۶	۰	۷۵/۷	۰	۰	۷	۰	۴/۶	۴/۳
BMP	۰	۰	۸/۳	۰	۹۱/۸	۰	۰	۰/۳	۰	۰
PNG	۳/۶	۲/۶	۰	۱/۶	۰	۸۳/۵	۴	۰	۱/۶	۲/۶
FLIF	۱/۳	۷	۰	۳/۳	۰	۱/۳	۷۴/۱	۰	۱۰	۲/۳
GIF	۰/۳	۰	۰	۰	۰	۰/۶	۰	۹۷/۴	۰	۱/۶
J2K	۰/۶	۵/۳	۰	۳/۶	۰	۰	۵	۰/۶	۸۱/۲	۳/۳

JXR	۲/۶	۳/۳	۰	۲	۰/۳	۵	۳/۳	۵/۶	۱/۳	۷۵/۵

با مقایسه ماتریس‌های درهم‌ریختگی می‌توان گفت درصورتی‌که از دسته‌بند درخت تصمیم استفاده کنیم اضافه کردن ویژگی طولانی‌ترین زیررشته و زیرتوالی مشترک روی مجموعه داده‌های ما نه‌تنها باعث بهبود دقت در تفکیک کدک‌های BMP و TIFF می‌شود بلکه در دقت تشخیص کدک‌های دیگر نیز به‌طور چشم‌گیری مؤثر خواهد بود. این روش بیشترین تأثیر را با افزایش ۱۰ درصدی روی دقت دسته‌بندی FLIF گذاشته و آن را از ۷۴/۱ به ۸۴/۲ تبدیل کرده است. به‌طورکلی متوسط صحت دسته‌بندی کدک‌های تصویری با به‌کارگیری دسته‌بند درخت تصمیم در مقایسه با روش بیب از ۸۲/۵۳ به ۸۷/۵۳ افزایش پیدا کرده است.

شکل ۲ بخشی از درخت تصمیم روش پیشنهادی اول را نشان می‌دهد. همان‌طور که ملاحظه می‌شود، ویژگی انحراف مطلق میانگین تنها در سطوح انتهایی درخت ایفای نقش کرده و درنتیجه تأثیر چندانی در شناسایی نوع فایل‌های تصویری نخواهد داشت. با تحلیل نتایج آزمایش‌های کلهون در تفکیک دوبه‌دوی مجموعه داده‌هایش در [۱۱] می‌توان دید که ویژگی همبستگی (وابستگی بین دو بایت متوالی) ویژگی مفیدی است که در مجموعه ویژگی‌های روش پیشنهادی اول مورداستفاده قرار نگرفته است. در روش پیشنهادی دوم ویژگی همبستگی را به مجموعه ویژگی‌های خود اضافه کرده و ویژگی انحراف مطلق میانگین را از مجموعه ویژگی‌ها حذف می‌کنیم. جدول ۴ ماتریس درهم‌ریختگی روش پیشنهادی دوم با استفاده از دسته‌بند درخت تصمیم را نشان می‌دهد. با تحلیل مقادیر به‌دست‌آمده مشهود است که این روش در تشخیص درست JPEG نقش مؤثری را ایفا کرده و نسبت به‌روش پیشنهادی اول حدود ۸ درصد در دقت تشخیص JPEG از سایر کدک‌ها افزایش ایجاد کرده است. همچنین BPG، GIF، TIFF و JXR نیز با افزایش دقت تشخیص همراه هستند. با استفاده از این روش، متوسط صحت دسته‌بندی ۱۰ نوع داده تصویری حدود ۱/۱ درصد افزایش‌یافته و از ۸۷/۸ به ۸۹/۹ رسیده است. شکل ۳ درخت تصمیم روش پیشنهادی دوم را نشان می‌دهد با مقایسه این شکل و شکل ۲ می‌توان اهمیت ویژگی همبستگی را نسبت به انحراف مطلق میانگین ملاحظه نمود. این ویژگی در سطوح بالاتر ایفای نقش کرده و درنتیجه در تفکیک انواع داده‌های تصویری نقش مؤثرتری خواهد داشت.



شکل ۲: بخشی از درخت تصمیم روش پیشنهادی اول

جدول ۳: ماتریس درهم‌ریختگی روش پیشنهادی اول با استفاده از دسته‌بند درخت تصمیم

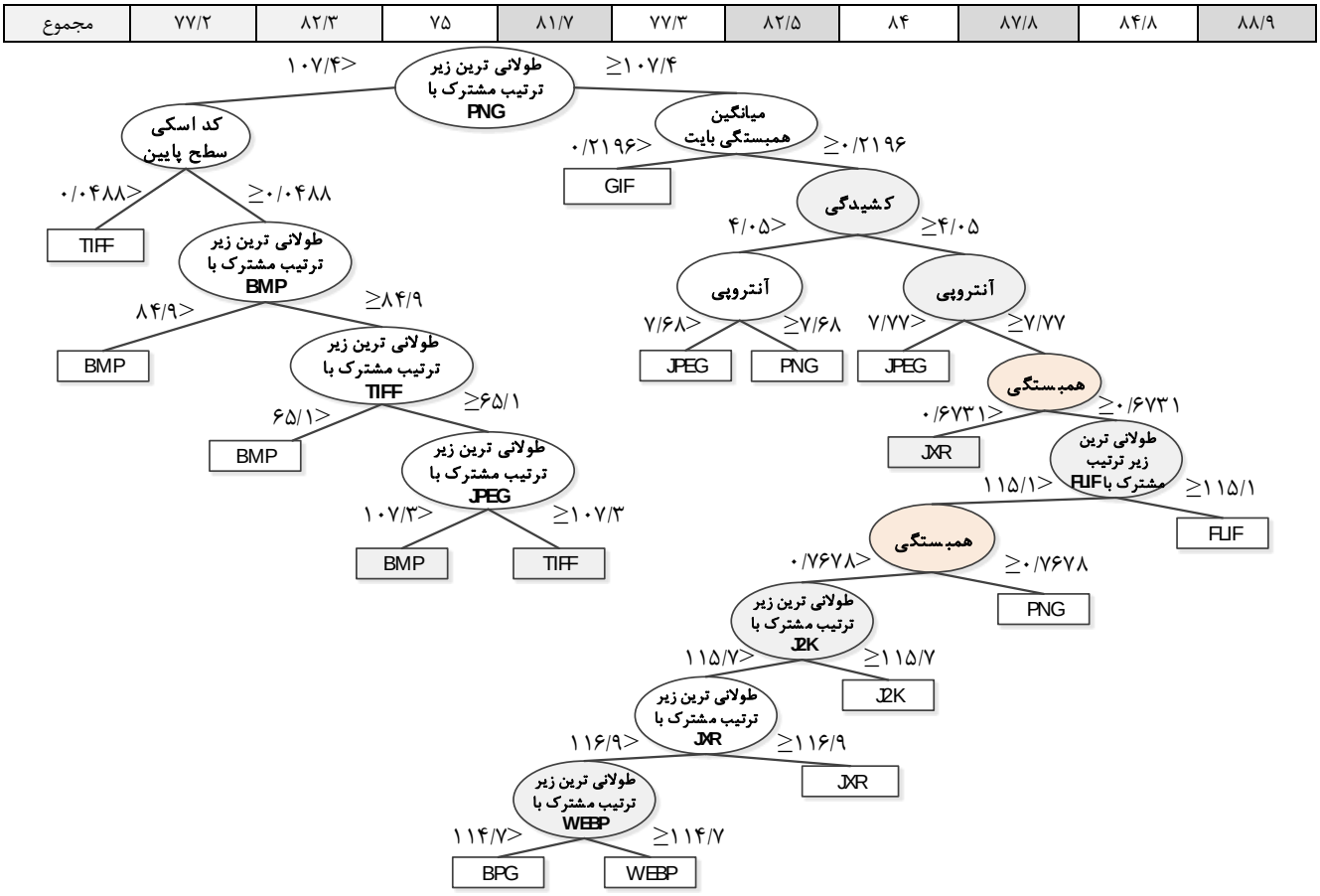
نوع تصویر	JPEG	BPG	TIFF	WEBP	BMP	PNG	FLIF	GIF	J2K	JXR
JPEG	۸۲/۷	۰/۶	۰	۴	۰	۱/۳	۴	۲	۱/۳	۴
BPG	۰/۶	۸۶/۲	۰	۴/۶	۰	۰/۶	۱/۳	۰	۵	۱/۳
TIFF	۰	۰	۸۶	۰	۱۴	۰	۰	۰	۰	۰
WEBP	۰/۶	۱/۳	۰	۸۴/۳	۰	۱/۳	۵	۰	۲/۶	۴/۶
BMP	۰	۰	۶/۳	۰	۹۲/۶	۰	۰	۰	۰	۰
PNG	۰	۲/۶	۰	۲/۶	۰	۸۹/۷	۰/۶	۰/۶	۰/۶	۲/۶
FLIF	۲/۳	۳/۳	۰	۴/۳	۰	۰	۸۴/۲	۰	۲/۳	۳/۳
GIF	۰	۰/۶	۰	۰	۰	۰	۰	۹۸/۷	۰	۰/۶
J2K	۰	۲	۰	۰	۰	۰/۶	۲	۰	۹۰/۳	۱/۶
JXR	۰/۶	۲/۶	۰	۲/۶	۰	۲	۰/۶	۲	۵/۶	۸۳/۷

جدول ۴: ماتریس درهم‌ریختگی روش پیشنهادی دوم با استفاده از دسته‌بند درخت تصمیم

نوع تصویر	JPEG	BPG	TIFF	WEBP	BMP	PNG	FLIF	GIF	J2K	JXR
JPEG	۹۰/۴	۲/۶	۰	۰	۰	۱/۳	۳	۰	۰	۲/۶
BPG	۰	۸۹/۱	۰	۲/۶	۰	۰/۶	۲/۶	۰	۲/۶	۲
TIFF	۰/۳	۰/۳	۸۷/۱	۰	۱۲/۳	۰	۰	۰	۰	۰
WEBP	۰	۲/۶	۰	۸۳/۶	۰	۰	۳/۳	۲/۳	۲/۶	۵/۳
BMP	۰	۰	۷/۳	۰	۹۲/۵	۰	۰	۰	۰	۰
PNG	۱/۶	۲/۶	۰	۳/۳	۰	۸۸/۳	۱/۳	۰/۶	۰	۲
FLIF	۴/۶	۲	۰	۳/۶	۰	۰/۶	۸۳/۳	۰	۰	۵/۳
GIF	۰/۶	۰	۰	۰	۰	۰	۰	۹۹/۳	۰	۰
J2K	۰	۲/۶	۰	۳/۶	۰	۰/۶	۲/۶	۰	۸۹/۶	۰/۶
JXR	۰/۶	۱/۳	۰	۵/۳	۰	۳/۳	۰/۳	۰	۰/۶	۸۵/۶

جدول ۵: دقت دسته‌بندی کدک‌های تصویری مختلف با روش‌ها و دسته‌بندهای متفاوت

	روش کلهون		روش دیواکاران		روش نیکل بیب		روش پیشنهادی اول		روش پیشنهادی دوم	
	ماشین بردار پشتیبان	درخت تصمیم	ماشین بردار پشتیبان	درخت تصمیم	ماشین بردار پشتیبان	درخت تصمیم	ماشین بردار پشتیبان	درخت تصمیم	ماشین بردار پشتیبان	درخت تصمیم
JPEG	۸۵/۱	۸۰/۲	۸۴/۲	۸۲/۱	۸۶/۳	۸۱/۷	۸۹/۰	۸۲/۷	۹۱	۹۰/۴
BPG	۶۷/۲	۸۰/۷	۶۸/۳	۸۱/۸	۶۶/۶	۸۰/۴	۷۹/۶	۸۶/۲	۸۶/۶	۸۹/۱
TIFF	۷۷/۲	۸۳/۵	۷۶/۶	۸۳/۸	۷۸	۸۳/۸	۷۴	۸۶	۷۴/۳	۸۷/۱
WEBP	۶۵/۸	۷۴/۲	۶۵	۷۳/۶	۶۵/۶	۷۵/۷	۸۱/۶	۸۴/۳	۸۵/۳	۸۳/۶
BMP	۹۸/۱	۹۰/۱	۹۷/۳	۸۷/۵	۹۸/۳	۹۱/۸	۹۹/۳	۹۲/۶	۸۲/۳	۹۲/۵
PNG	۷۳/۷	۸۲/۱	۶۲/۵	۸۱/۸	۷۶	۸۳/۵	۸۰/۳	۸۹/۷	۸۰/۶	۸۸/۳
FLIF	۵۹/۸	۷۳/۶	۵۶/۶	۶۹/۱	۶۰/۳	۷۴/۱	۷۵/۶	۸۴/۲	۸۴/۶	۸۳/۳
GIF	۹۷/۵	۹۷/۲	۹۵	۹۷/۷	۹۸/۳	۹۷/۴	۹۶/۳	۹۸/۷	۹۵/۳	۹۹/۳
J2K	۷۴/۹	۸۴/۳	۷۳/۳	۸۳/۴	۷۰/۶	۸۱/۲	۸۲/۶	۹۰/۳	۸۳	۸۹/۶
JXR	۷۲/۳	۷۶/۸	۷۱/۳	۷۶/۵	۷۲/۶	۷۵/۵	۸۱/۳	۸۳/۷	۸۵	۸۵/۶



شکل ۳: درخت تصمیم روش پیشنهادی دوم

جدول ۵ مقایسه‌ای بین دقت دسته‌بندی کدک‌ها در روش‌های مختلف با استفاده از دسته‌بندهای درخت تصمیم و ماشین بردار پشتیبان را نشان می‌دهد. همان‌طور که قبلاً اشاره شد و از این جدول هم پیداست، روش پیشنهادی دوم با استفاده از دسته‌بند درخت تصمیم و متوسط صحت ۸۸/۹ درصد دارای بهترین دقت تفکیک بین ۱۰ نوع داده تصویری موجود در مجموعه داده‌های ما می‌باشد.

۵ -نتیجه‌گیری

در این مقاله روشی برای دسته‌بندی بین کدک‌های مختلف تصویر پیشنهاد شد. در این راستا، ابتدا به جزئیات روش‌های استخراج مشخص استفاده‌شده توسط بیب [۱۴]، دیواکاران [۲۰] و کلهون [۱۱] پرداخته شد. در ادامه، ویژگی‌های مورداستفاده در روش بیب و دیواکاران را به‌طور جداگانه بر مجموعه داده‌های تولیدی به کار گرفتیم. با بررسی نتایج آزمایش‌های کلهون ۲۰ ویژگی طولانی‌ترین زیررشته و زیرتوالی مشترک را مناسب یافتیم و آن‌ها را به اجتماع مجموعه ویژگی‌های بیب و دیواکاران افزودیم. با این روش پیشنهادی اول متوسط صحت دسته‌بندی به ۸۷/۸ درصد افزایش پیدا کرد. با تحلیل درخت تصمیم روش پیشنهادی اول به این نتیجه رسیدیم که ویژگی انحراف مطلق میانگین که در روش نیکل بیب به‌کاررفته است، تأثیر چندانی بر دسته‌بندی ندارد. به همین علت از ویژگی همبستگی مورداستفاده توسط کلهون بهره جستیم و ویژگی انحراف مطلق میانگین را از مجموعه ویژگی‌ها حذف کردیم. با این کار و با بهره‌گیری از درخت تصمیم، دقت تشخیص نوع داده‌های تصویری BPG، TIFF، JPEG، GIF و JXR توسط روش پیشنهادی دوم تا حدودی بهبود یافت و به متوسط صحت ۸۸/۹ درصد رسیدیم.

در این مقاله، دسته‌بندی بین ۱۰ کدگذار تصویر انجام شد. به‌عنوان یک مسئله باز و عملی‌تر در آینده، می‌توان تعداد این کدگذارها را افزایش داد. همچنین پارامترهای کدگذاری متنوع‌تری را برای این کدگذارها انتخاب نمود.

مراجع

[۱] مهدی تیموری، «آشکارسازی سیگنال لینک ۱۶»، *مجله مهندسی برق دانشگاه تبریز*، جلد ۴۶، شماره ۴، ص. ۸۴-۸۷، ۱۳۹۵.

[۲] مهدی تیموری، حمیدرضا کاکایی مطلق و مرتضی حدادی، «شناسایی کور کدهای ضربی BCH»، *مجله مهندسی برق دانشگاه تبریز*، جلد ۴۷، شماره ۱، ص. ۴۹-۵۴، ۱۳۹۶.

[3] E. S. Pilli, R. C. Joshi, and R. Niyogi, "Network forensic frameworks: Survey and research challenges," *digital investigation*, vol. 7, pp. 14-27, 2010.

[4] G. Kessler. File Signature Table, 16 December 2017, https://www.garykessler.net/library/file_sigs.html.

[5] M. McDaniel and M. H. Heydari, "Content based file type detection algorithms," *in Proceedings of the 36th International Conference on Annual Hawaii International Conference on,System Sciences*, pp. 1-10, 2003.
[6] W.-J. Li, K. Wang, S. J. Stolfo, and B. Herzog, "Fileprints: Identifying file types by n-gram analysis," *in Information Assurance Workshop, 2005. IAW'05. Proceedings from the Sixth Annual IEEE SMC*, pp. 64-71, 2005.
[7] M. Karresand and N. Shahmehri, "Oscar—file type identification of binary data in disk clusters and ram pages," *in Security and privacy in dynamic environments, ed: Springer*, pp. 413-424, 2006.
[8] M. Karresand and N. Shahmehri, "File type identification of data fragments by their binary structure," *in Information Assurance Workshop, 2006 IEEE*, pp. 140-147, 2006.
[9] S. J. Moody and R. F. Erbacher, "Sádi-statistical analysis for data type identification," *in Third International Workshop onSystematic Approaches to Digital Forensic Engineering*, SADFE'08., pp. 41-54, 2008.
[10] C. J. Veenman, "Statistical disk cluster classification for file carving," *in Information Assurance and Security, 2007. IAS 2007. Third International Symposium on*, pp. 393-398, 2007.
[11] W. C. Calhoun and D. Coles, "Predicting the types of file fragments," *Digital Investigation*, vol. 5, pp. S14-S20, 2008.
[12] M. C. Amirani, M. Toorani, and A. Beheshti, "A new approach to content-based file type detection," *in Computers and Communications,. ISCC 2008. IEEE Symposium on*, pp. 1103-1108, 2008.
[13] K. Nguyen, D. Tran, W. Ma, and D. Sharma, "A New Approach to Executable File Fragment Detection in Network Forensics," *in Network and System Security, ed: Springer*, pp. 510-517, 2014.
[14] N. L. Beebe, L. A. Maddox, L. Liu, and M. Sun, "Sceadan: using concatenated n-gram vectors for improved file and data type classification," *IEEE Transactions on Information Forensics and Security*, vol. 8, pp. 1519-1530, 2013.
[15] M. Norouzi, D. J. Fleet, and R. R. Salakhutdinov, "Hamming distance metric learning," *in Advances in neural information processing systems*, pp. 1061-1069, 2012.
[16] D. Ruppert, "What is kurtosis? An influence function approach," *The American Statistician*, vol. 41, pp. 1-5, 1987.
[17] L. T. DeCarlo, "On the meaning and use of kurtosis," *Psychological methods*, vol. 2, p. 292, 1997.
[18] R. A. Groeneveld and G. Meeden, "Measuring skewness and kurtosis," *The Statistician*, pp. 391-399, 1984.
[19] W.-J. Li, K. Wang, S. J. Stolfo, and B. Herzog, "Fileprints: Identifying file types by n-gram analysis," *in Proceedings from the Sixth Annual IEEE SMC Information Assurance Workshop*, pp. 64-71, 2005.
[20] D. M. Divakaran, Y. S. Liau, and V. L. Thing, "Accurate in-network file-type classification," *in Proceedings of the Singapore Cyber-Security Conference (SG-CRC) 2016: Cyber-Security by Design*, p. 139-146, 2016.

زیرنویس‌ها

[1] Classification
[2] Identification
[3] Train
[4] Test
[5] Cross validation
[6] Fold
[7] Decision tree
[8] Calhoun
[9] Coles
[10] Beebe
[11] Average contiguity between bytes
[12] Mean Absolut Deviation
[13] Cross correlation
[14] Longest Common SubString
[15] Longest Common SubSequence
[16] Hamming Distance
[17] Hamming Weight
[18] Kurtosis
[19] Skewness
[20] Unigram
[21] Bigram
[22] Trigram
[23] Byte frequency distribution(BFD)
[24] Wang
[25] Divakaran
[26] Confusion Matrix