رده بندی بدافزار اندروید با استفاده از الگوریتم کلاستربندی K-Means

چکیده

بدافزار برای کسب دسترسی یا آسیب رساندن به یک سیستم کامپیوتر بدون هشدار به کاربر طراحی شده بود. به علاوه، مهاجم از بدافزار برای ارتکاب جرم یا شیادی بهره می برد. این مقاله رویکرد رده بندی بدافزار اندروید را بر مبنای الگوریتم کلاستربندی K-Means ارائه کرده است. ما مدل پیشنهادی را بر حسب دقت با استفاده از الگوریتم هاییادگیری ماشین ارزیابی می کنیم. دو مجموعه داده ای برای نمایش تمرین الگوریتم های کلاستربندی K-Meand انتخاب شدند که پایگاه داده Virus Total و Malgenome بودند. ما بدافزار اندروید را در به سه رده تقسیم کردیم که ransomware، scareware و goodware هستند. نه ویژگی برای هر نوع مجموعه داده ای در نظر گرفته شدند شامل Lock Detected (قفل شناسایی شده)، Text Detected (متن شناسایی شده)، Encryption Detected (رمزگذاری شناسایی شده)، Treat (تهدید)، Porn (شهوانی)، Law (قانون)، Copyright (حق نشر) و Moneypak. ما از نرم افزار IBM SPSS Statistic برای رده بندی داده ها و از ابزارهای WEKA برای ارزیابی کلاستر ساخته شده استفاده کردیم. الگوریتم کلاستربندی K-Means پیشنهادی نشانگر نتیجه امیدوار کننده ای با دقت با در هنگام آزمایش با استفاده از الگوریتم Random Forest بود.

1- مقدمه

بدافزار برای کسب دسترسی یا آسیب رساندن به یک کامپیوتر بدون اطلاع کاربر توسعه یافته است. موارد متعددی برای بدافزار وجود دارند که شامل spyware، Key looggerها یا ویروس هایی هستند که روی پردازشگر داده سازمان تأثیر می گذارند. بدافزار به رشد ادامه می دهد و تکامل می یابد تا از آنتی ویروس ها و دیگر سطوح محافظت عبور کند و تیم امنیت نمی تواند آن را کنترل کند. بیش از 4000 حمله ransomware هر روزه از سال 2016 رخ داده است. این افزایشی 300 درصدی نسبت به سال 2015 است که در آن 1000 حمله ransomware در هر روز مشاهده می شد. مجرمین از طریق بدافزار می تواند تعداد

زیادی از قربانیان را به طور یک مرتبه ای با اتوماتیک کردن این حملات و بسط دسترسی آلودگی های خودشان به چندین سیستم به ازای هر قربانی آلوده نمایند. این مورد می تواند باعث آسیب بیشتر و زمان افت بالقوه ای شود که فشار بیشتری روی قربانیان برای حل سریع تر مشکل وارد می کند. معمولاً افراد داده های مهم را روی دستگاه های الکترونیک نظیر لپ تاپ و موبایل بدون هیچ گونه پشتیبان گیری نگه می دارند. وقتی دستگاه های الکترونیکی آلوده می شوند یا توسط بدافزار اندروید مورد حمله قرار می گیرند، بازگرداندن داده ها کار مشکلی می شود.

دو نوع بدافزار اندروید وجود دارد که Ransomware و Scareware هستند. Scareware در صنعتی میلیارد دلاری در سال 2016 منفجر شد که جو تب طلایی برای مجرمین سایبری، با تقاضا و تأمین متغیرهای ransomware جدید و پلتفرم های ارسال ایجاد شود. Ransomware از طریق ایمیل اسپمی کار می کند که حاوی پیوست آلوده است. پیوست آلوده از کاربر می خواهد که ضمیمه را که ظاهری متقاعد کننده دارد، باز کند. پس از آلوده شدن، ransomware از کاربر در دسترسی به سیستم جلوگیری می کند یا دسترسی او را محدود می کند یا صفحه کامپیوتر را قفل می کند یا روی فایلی کدگذاری می کند که با یک رمز عبور حروف چینی شده است. سپس، پیام ransom نمایش داده می شود که به کاربر دستور می دهد پول غرامت را از طریق سیستم پرداختی نظیر Ukash یا Paysafecard بپردازد تا مجدداً دسترسی پیدا کند. برعکس، scareware به عنوان نرم افزا آنتی ویروس جعلی شناخته شده که به معمول ترین روش ها برای فریب و استفاده از پول قربانی تبدیل شده است. مایکروسافت scareware را در ایالات متحده در سال 2011 به تعداد 52 میلیون مرتبه شناسایی کرده است. برنامه scareware درست مانند برنامه های قانونی امنیتی است. scareware ادعا کرده که تعداد زیادی از تهدیدهای ناموجود روی کامپیوتر را شناسایی کرده و سپس قربانی را وادار می کند تا پولی برای نسخه کامل نرم افزار بپردازد تا تمامی خطرات را از بین ببرد.

این مقاله روی رده بندی بدافزار اندروید با استفاده از الگوریتم کلاستربندی K-Means که روی دو مجموعه داده ای استخراج شده از شناساگر ransom.mobi استخراج شده است تمرکز می نماید. مجموعه داده Virus Total از 907 نمونه تشکیل شده است، در حالی که مجموعه داده Malgenome تعداد 1255 نمونه دارد. هر دو مجموعه داده ای نه نوع ویژگی دارند که شامل Detected (قفل شناسایی شده)، Text

Detected (متن شناسایی شده)، Encryption Detected (رمزگذاری شناسایی شده)، Treat (تهدید)، Porn (شهوانی)، Law (قانون)، Copyright (حق نشر) و Moneypak می باشند. پس از آن، رده بدافزار اندروید که با استفاده از الگوریتم کلاستربندی K-Means ساخته شده است با استفاده از الگوریتم جنگل تصادفی (Random Forest) تجزیه و تحلیل خواهد شد. اهداف این مقاله به شرح ذیل هستند:

الف) طراحی یک مدل رده بندی بدافزار اندروید بر مبنای رویکرد رفتار.

ب) رده بندی بدافزار اندروید با استفاده از الگوریتم کلاستربندی K-Means.

ج) ارزیابی مدل پیشنهادی بر حسب دقت با استفاده از الگوریتم های یادگیری ماشین.

ادامه این مقاله به شرح زیر سازمان یافته است:

در بخش 2 توضیحی در مورد کار مرتبط رو رده بندی بدافزار اندروید و تکنیک کلاستربندی K-Means ارائه شده است. در فصل 3 مدل رده بندی پیشنهادی برای رده بندی بدافزار اندروید که در آن هر پیش بینی کلاستر به عناصر کلاستر تبدیل می شود، ارائه می گردد. کلاستر ساخته شده از الگوریتم کلاستربندی مبتنی بر قاعده پس از آن برای آموزش الگوریتم رده بندی کننده استفاده شده است. بخش 4 متدلوژی ها و نتایج تجربی ارزیابی تحلیل عملکرد را نشان می دهد. در نهایت در فصل 5 نتیجه گیری کار ارائه شده و تحقیق آینده برجسته می شود.

2- کار مرتبط

بدافزار می تواند در اشکال مختلفی از کد، اسکریپت ها، محتوای فعال و دیگر نرم افزارها ظاهر شود. بدافزار یک عبارت جهانی است که برای اشاره به چندین نوع نرم افزار متخاصم شامل ویروس های کامپیوتری، ransomware، کرم ها، اسب های تروجان، rootkitها، key loggerها، dilerها، spyware، adware و دیگر برنامه های آسیب رسان به کار برده شده است.

2-1- رویکرد رده بندی بدافزار اندروید

چندین تکنیک ضد ransomware در سال های اخیر برای شناسایی و جلوگیری از افزایش تعداد حملات ransomware، به صورت که در جدول 1 نشان داده شده است، پیشنهاد شده اند. در کل، بسیاری از محققان، الگوریتم کلاستربندی را برای رده بندی بدافزار اندروید به کار می برند. جدول 1 تحلیل قیاسی رویکرد

رده بندی بدافزار اندروید را نشان می دهد. کار انجام شده توسط Wu و همکارانش DroidMat را برای شناسایی بدافزار اندروید با استفاده از ویژگی های مبتنی بر رفتار ارائه کرده است. DroidMat اطلاعات ایستا را از هر فایل مانیفست برنامه استخراج می کند و API Calls مربوط به مجوزها می باشد. الگوریتم K-means برای پیشبرد قابلیت مدلسازی بدافزار به کار برده شده است. پس، تعداد کلاسترها بواسطه روش تجزیه مقدار منفرد (kNN) برای رده بندی برنامه به عنوان برنامه بی خطر یا آلوده تعیین شده است. آن ها برای رسیدن به دقت 97/87٪ تست شده روی مجموعه داده Contagio Mobile کار می کنند.

پژوهش انجام شده توسط Burguera و همکارانش تحلیلی پویا از رفتار برنامه ای برای تشخیص بدافزار در پلتفرم اندروید ارائه کرد (Crowdroid). Crowdroid در چارچوبی برای گردآوری اثرات از کاربران واقعی بر مبنای انبوه سپاری تنظیم شده است. آن ها به دقت 100٪ تست شده روی دو نوع مجموعه داده ای دست یافتند: بدافزار مصنوعی برای اهداف آزمایش ساخته شد و بدافزار واقعی از Virus Total. اما، آزمایش برای مقدار کمی از داده ها تست شد. کار دیگر انجام شده توسط Aung و همکاران چارچوبی برای رده بندی برنامه های اندروید با استفاده از تکنیک های یادگیری ماشین پیاده سازی کرده است. این سیستم ویژگی های مبتنی بر مجوز متفاوت و رخدادهای کسب شده از برنامه های اندروید را مانیتور می کند. آن ها روی 200 نمونه مجموعه داده ها با استفاده از رده بندی کننده های یادگیری ماشین برای طبقه بندی این مورد تست شدند که آیا برنامه بی خطر است یا بدافزار است. کار ما متفاوت از پژوهش Aung و همکارانش می باشد از آن نظر که ما مجموعه داده های Virus Total و Malgenome را با استفاده از الگوریتم K-means به سه رده تقسیم می کنیم؛ ransomware، scareware یا goodware. علاوه بر این کار انجام شده توسط Schlesinger و همکاران از داده های زنده با ویژگی مبتنی بر مجوز استفاده کرده که ما در اینجا از ویژگی مبتنی بر رفتار استفاده کردیم. پس، ما ویروس را با استفاده از الگوریتم کلاستربندی K-Means گروه بندی می نماییم. ما الگوریتم جدول تصادفی را انتخاب کردیم، چون مناسب ترین الگوریتم برای هر دو مجموعه داده ای است

جدول 1 : تحلیل قیاسی روی رویکرد رده بندی بدافزار اندروید.

کار انجام شده	ویژگی ها	الگوریتم	مجموعه داده	/نتیجه

توسط				
DroidMat	مجوزها، بکارگیری مولفه ها، پیام های عمدی گذرنده و تماس های API	K-Means و KNN	Contagio Mobile	٪97/87
Croedroid	بدافزار اندروید مبتنی بر رفتار	K-Means	Virus Total	٪100
مبتنی بر مجوز	مجوز و رخداد	K-Means و جنگل تصادفی	Android Application	٪91/75

2-2- رده بندی داده ها

در کل دو نوع داده رده بندی داده ای وجود دارد؛ یادگیری نظارت نشده و نظارت شده. یادگیری نظارت نشده مدلی با نتایج درست در حین آموزش ارائه نمی کند. بنابراین مبنای ویژگی های آماری آن ها را می توان فقط به عنوان کلاستر استفاده کرد. کلاستر می تاند حتی در صورتی انجام گردد که تنها برای عده کمی از نمایندگان اشیای کلاس های مطلوب در دسترس باشد. از طرف دیگر، یادگیری نظارت شده داده های ورودی آموزش با نتایج مطلوب ارائه می کند. نتایج درست مشخص شده اند و به عنوان ورودی در عرض فرایند یادگیری به مدل داده شده اند. ساختار مجموعه آموزش مناسب، اعتبارسنجی و تست بسیار حیاتی و مهم است. این روش ها معمولاً سریع و دقیق هستند. به علاوه، باید قادر به تعمیمی باشد که نتایج صحیح در زمانی ارائه می کند که داده های جدیدی در ورودی بدون دانش قبلی در مورد هدف داده شده اند.

3- مدل رده بندی

این بخش توضیحی در مورد مدل رده بندی با استفاده از الگوریتم کلاستربندی K-Means ارائه می نماید.

3-1- مدل رده بندی بدافزار اندروید

پنج فاز برای رده بندی داده های نظارت نشده ای نیاز هستند که داده های خام، پیش پردازش، استخراج ویژگی، الگوریتم کلاستربندی و الگوریتم رده بندی هستند، وجود دارد که درست به همان صورتی است که در شکل 1 نشان داده شده است. ما بدافزار اندروید را به سه نوع تقسیم می کنیم که ransomware، scareware و goodware می باشند. دو مجموعه داده ای از شناساگر ransom.mobi استخراج شدند؛ Virus Total و Malgenome. این مجموعه داده ها، داده های نظارت نشده ای هستند که برای تحلیل داده های توصیفی

برای یافتن الگوهای پنهان یا گروهی از داده ها استفاده شده اند. پیش پردازش داده ها تکنیک های کاوش داده ای هستند که داده های خام را به قالب قابل درک تبدیل می کنند. برای تکمیل این فاز، داده های خام باید متحمل یک سری مرحله پیش پردازش در جدول 2 شوند.

جدول 2 – مراحل در پیش پردازش

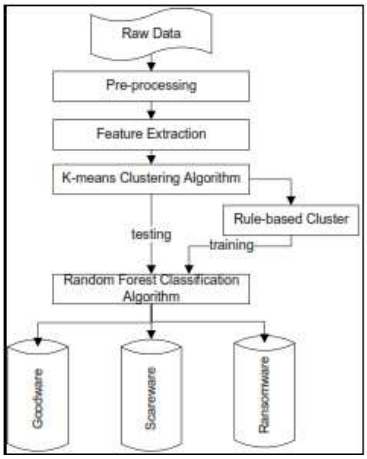
مراحل پیش پردازش	شرح
تمیز کردن داده ها	بخش مقادیر از دست رفته را پر می کند، داده های نویزدار را هموار کرده یا تناقضات موجود در داده ها را حل می کند.
یکپارچه سازی داده ها	تناقض درون داده ها در زمانی حل خواهد شد که داده ها با نمود مختلفی کنار یکدیگر قرار داده شده باشند.
تغییر داده ها	داده ها نرمال سازی شده اند، گردآوری شدند و تعمیم یافته اند.
تقلیل داده ها	نمود تقلیل یافته ای از داده ها در منبع داده ها ارائه می نماید.
تفکیک داده ها	شامل کاهش تعدادی از طیف های مقداری وقفه های مشخصه می باشد.

2-3- استخراج ویژگی

در ابتدا، مجموعه داده های Virus Total و Malgenome از وب سایت شناساگر ransom.mobi دانلود شدند. ما 907 و 1255 نمونه مجموعه داده به ترتیب از Virus Total و Malgenome انتخاب کردیم. فایل برای استخراج ویژگی های لازم بدافزار اندروید در قالب xls. نافشرده سازی شد. هر دو مجموعه داده منتخب نه ویژگی داشتند که برای رده بندی بدافزار اندروید طبق رده آن استفاده شده است. ویژگی ها به شرح زیر هستند:

Lock Detected، Text Detected، Text Score، Encryption Detected، Thread، Porn، Law، Copyright و Moneypak. اما رفتار ویروس ransomware می تواند بر مبنای سه ویژگی شامل Locking Detector، Encryption Detector و Threatening Text Detector نمایه سازی شود، درست همان طور که در جدول 3 نشان داده شد است. سپس، ما مجموعه داده ها را در قالب فایل (arff.) از ویژگی های استخراج شده ایجاد کردیم. در نهایت، مجموعه داده ها را با استفاده از الگوریتم رده بندی جنگل تصادفی آزمایش کردیم تا تشخیص دهیم که بدافزار اندروید Ransomware، Scareware یا

Goodware است، چون قوی تر است. الگوریتم جنگل تصادفی تلفیقی از پیش بینی کننده های درخت است به گونه ای که هر درخت وابسته به مقادیر بردار تصادفی نمونه گیری شده به صورت مستقل و با همان توزیع برای تمامی درخت ها در جنگل می باشد.



شکل 1 – مدل رده بندی بدافزار اندروید

جدول 3: ویژگی های بدافزار اندروید مبتنی بر رفتار

تکنیک	شرح
شناساگر قفل	درخواست حق مدیریت دستگاه و سپس قفل کردن دستگاه را می خواهد. کادر گفتگوی هشدار صفحه کامل یا فعالیتی منطبق می کند. رخداد فشار دادن کلید را به گونه ای وسیله قرار می دهد که "پنجره قفل" نمی تواند توسط قربانیان از بین برود.
شناساگر کدگذاری	کلید کدگذاری سخت رمزگذاری شده است. کلیدی به ارای هر دستگاه تولید می کند.
شناساگر متن تهدیدآمیز	به خانواده هایی که در انگلیس زندگی می کنند هشدار می دهد که برای پرداخت ها از Moneypak استفاده کنند، در حالی که خانواده هایی که در روسیه هستند کارت اعتباری را می پذیرند.

3-3- الگوریتم کلاستربندی K-Means

در این مقاله، ما مجموعه داده ها را با استفاده از الگوریتم کلاستربندی K-Means رده بندی می کنیم. الگوریتم کلاستربندی K-Means یک تکنیک داده کاوی است که می توان از آن برای مرتب سازی مجموعه داده ها در سه رده استفاده کرد که ransomware، scareware و goodware می باشند.

$$J(v) = \sum_{i=1}^{\tilde{c}} \sum_{j=1}^{\tilde{c}_i} (\|X_i - V_j\|)^2$$

که در آن $\|Xi - Vj\|$ فاصله اقلیدسی میان Xi و Vj است. در حالی که C_i تعداد نقاط داده ای در iآمین کلاستر است و C مرکز کلاستر است. برای ارزیابی دقت این روش، داده های کلاستربندی شده با داده های برچسب گذاری شده مقایسه شده اند تا مشخص شود که آیا موارد به درستی کلاستربندی شده اند یا خیر. با توجه به دانش قبلی که مجموعه فعلی حاوی سه نوع بدافزار اندروید می باشد، الگوریتم K-Means در سه کلاستر به کار برده شده است. مقدار درست یا بهینه k به سادگی تعیین نشده است.

$X= \{X_1, X_2, X_3, \dots\dots\dots\dots\dots\dots X_n\}$ را برابر مجموعه ای از نقاط داده ای قرار دهید و اجازه دهید $X= \{V_1, V_2, \dots V_c\}$ مجموعه ای از مراکز باشد. سپس، به صورت تصادفی مراکز کلاستر c را انتخاب کنید. سپس، فاصل میان هر نقطه داده ای و مراکز کلاستر را محاسبه کنید. پس از آن، داده ها را با مراکز کلاستری تخصیص دهید که فاصله آن ها از کلاستر مینیمم تمامی مراکز کلاستر است. متعاقباً، مجدداً مرکز کلاستر جدید را با استفاده از فرمول $V_i = (\frac{1}{c_i}) \sum_{j=1}^{c_i} X_1$ محاسبه کنید که در آن C_i نشانگر تعداد نقاط داده ای در iامین کلاستر است. در نهایت، فاصله میان هر نقطه و مراکز کلاستر جدید به دست آمده را مجدداً محاسبه کنید. اگر هیچ نقطه ای مجدداً تخصیص نیافته است متوقف شوید؛ در غیر این صورت مجدداً فرایند تخصیص نقاط داده ای به مرکز کلاستری را تکرار کنید که فاصله آن ها از کلاستر مینیمم تمامی مراکز کلاستر است.

4-3- کلاستربندی مبتنی بر قاعده

جدول 4 کلاستربندی مبتنی بر قاعده را نشان می دهد که برای رده بندی مجموعه داده ای Virus Total و Malgenome متعلق به کلاسترهای منتخب استفاده شده است. دو ویژگی وجود دارند که باید در نظر گرفتند شوند که شامل Lock Detected و Encryption Detected می باشند. اگر هر دو ویژگی صحیح (true) باشند، داده ها ransomware هستند. به علاوه، اگر مقدار ویژگی Lock Detected صحیح باشد و Encryption Detected غلط (false) باشد، بنابراین داده ها scareware هستند. اما اگر هر دو ویژگی مقدار غلط (false) داشته باشند، داده ها goodware می باشند.

جدول 4 – کلاستربندی مبتنی بر قواعد

ویژگی ها	بدافزار اندروید		
	Ransomware	Scareware	Goodware
Lock Detected	True	True	False
Encryption Detected	True	False	False

4- تحلیل عملکرد

این بخش تنظیمات آزمایشی و مقیاس عملکرد استفاده شده برای رده بندی بدافزار اندروید را ارائه می نماید.

4-1- تنظیمات آزمایشی

در ابتدا آزمایش با گردآوری مجموعه داده ها از شناساگر ransom.mobi آغاز شد. ما از دو نوع مجموعه داده ای از Virus Total و Malgenome استفاده کردیم. سپس، نمونه ها از هر مجموعه داده ای استخراج شده و در فایل csv. ذخیره شدند. پس از آن، فرایند کلاستربندی K-Means با استفاده از نرم افزار آماری IBM SPSS Statistic انجام شد تا بدافزار اندروید به سه رده کلاستربندی شود که ransomware، scareware و goodware هستند. سپس، کلاسترهای پیش بینی شده ای که بر مبنای الگوریتم کلاستربندی K-Means استفاده شدند در یک فایل csv. ذخیره شدند. چون مجموعه داده ها داده های نظارت نشده بودند، ما کلاستر پیش بینی شده را در کلاستربندی مبتنی بر قاعده اجرا کردیم تا بدافزار اندروید را یا به صورت Ransomware، scareware یا goodware رده بندی نماییم. پس از آن، داده های نظارت شده را با کلاستر پیش بینی شده با استفاده از اندازه نسبت 60:40 تقسیم کردیم که در آن 60٪ از مجموعه داده ها به عنوان مجموعه آموزش استفاده خواهند شد، در حالی که 40٪ به عنوان مجموعه آزمایش استفاده خواهند شد. در نهایت رویکرد کلاستربندی پیشنهادی را با استفاده از الگوریتم رده بندی جنگل تصادفی روی Waikato Environment برای ابزارهای تحلیل اطلاعات (WEKA) آزمایش کردیم. پس از پردازش مدل با استفاده از سیستم آموزش، مدل پیش بینی هایی بر خلاق مجموعه تست ارائه خواهد داد. مجموعه آزمایش حاوی مقادیری است که برای ویژگی ای شناخته شده هستند که باید پیش بینی شوند. بنابراین تعیین این مورد که آیا تخمین مدل صحیح است یا نه ساده می باشد. وقتی مدل پرورش یافت و تست شد، باید عملکرد مدل ارزیابی شود.

الگوریتم: کلاستربندی مبتنی بر قاعده

```
INPUT: Class
BEGIN
1: FOR (each incoming DATA) DO
2: IF (LockDetected == TRUE && EncryptionDetected == TRUE) THEN
3: GIVE value Ransomware
4: ELSEIF (LockDetected == TRUE && EncryptionDetected == FALSE) THEN
5: GIVE value Scareware
6: ELSEIF (LockDetected == FALSE && EncryptionDetected == FALSE) THEN
7: GIVE value Goodware
8: ENDIF
9: ENDFOR
10: END Rule-Based Clustering
```

شکل 2 - الگوریتم کلاستربندی مبتنی بر قاعده

4-2- الگوریتم کلاستربندی مبتنی بر قاعده

شکل 2 الگوریتم کلاستربندی مبتنی بر قاعده را نشان می دهد. ورودی های الگوریتم کلاستربندی مبتنی بر قاعده مقدار ویژگی های LockDetected و EncryptionDetected هستند. در مرحله 2 تا 8، هر داده وارد شده مقدار LockDetected و EncryptionDetected را برای تمامی مجموعه های داده ای کاوش خواهد کرد تا مشخص کند که آیا رده داده ای ransomware است، scareware است یا goodware. اگر هر دو مقدار LockDetected و EncryptionDetected برابر true باشد، رده داده ای را ransomware در نظر بگیرید. اگر مقدار LockDetected برابر true و مقدار EncryptionDetected برابر false است، رده داده ای را scareware در نظر می گیریم. در نهایت اگر هم LockDetected و هم EncryptionDetected برابر False هستند، رده داده ها را Goodware در نظر می گیریم.

4-3- سنجش عملکرد

برای ارزیابی کارایی الگوریتم کلاستربندی پیشنهادی، ما از چهار مقیاس عملکرد زیر استفاده کرده ایم. این مقیاس ها به شرح زیر هستند:

الف) دقت (Acc): چه تعداد رده بدافزار به درستی توسط الگوریتم کلاستربندی مبتنی بر قاعده پیش بینی شده اند؟

ب) نرخ خطا (Err rate): چه تعداد رده بدافزار اندروید به اشتباه بواسطه الگوریتم کلاستربندی مبتنی بر قاعده پیش بینی شده اند؟

ج) منفی های غلط (FN): چه تعداد رده بدافزار اندروید توسط الگوریتم کلاستربندی مبتنی بر قاعده تشخیص داده نشده اند؟

د) مثبت های غلط (FP): چه تعداد رده بدافزار به اشتباه رده بندی شده اند؟

مقیاس های دقت برای محاسبه تعداد بدافزار اندرویدی که به درستی رده بندی شده و از الگوریتم پیشنهادی استفاده می کنند، بسیار قابل توجه هستند. اگر مقدار دقت بالا باشد، عملکد الگوریتم پیشنهادی برای رده بدافزار اندروید بسیار کارآمد است. به علاوه هم سنجه های FN و هم سنجه های FP در ارزیابی کارایی رویکردهای تعدیل امنیتی بسیار مهم هستند. به عنوان نمونه FP می تواند مقادیر منفی قابل توجهی در سودمندی الگوریتم تشخیص و محافظت داشته باشد. این به دلیل آن است که آزمودن آن ها مستلزم زمان و منبع می باشد. اگر نرخ FP بالا باشد، کاربر می تواند آن ها را نادیده بگیرد. سنجه نرخ خطا برای بازبیتی موضوعات تناسب بیش از حد مهم است. مشخصه عملیات گیرنده (ROC) از طرف دیگر مقیس قطعیت الگوریتم با رده بندی ایجاد شده است.

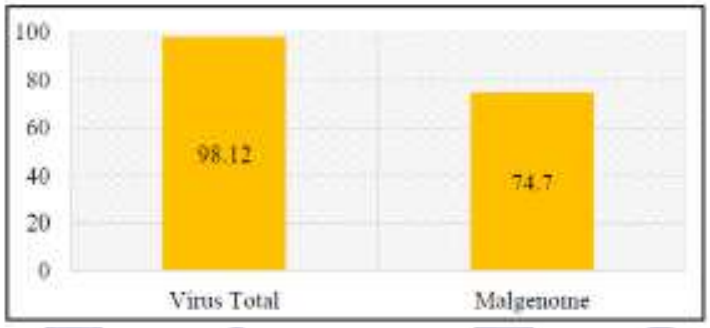
4-4- نتایج و بحث

در این بحش حروجی رده بندی الگوریتم های کلاستربندی K-Means روی ویژگی های استخراج شده ارائه می گردد. ما مجموعه داده ای را با استفاده از الگوریتم رده بندی جنگل تصادفی مقادیر اندازه نسبت 60:40 بر حسب مقدار دقت، خطای مطلق میانگین، مشخصه عملیات گیرنده (ROC) و نرخ مثبت صحیح (TP) و مثبت غلط (FP) آزمایش کردیم.

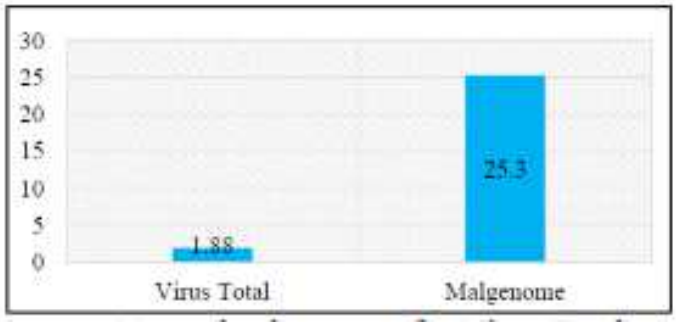
4-4-1- مقدار دقت. شکل 3 مقدار دقت را برای مجموعه داده های Virus Total و Malgenome نشان می دهد که با استفاده از الگوریتم رده بندی جنگل تصادفی تست شده اند. مجموعه داده Virus Total به بالاترین دقت دست یافت که 98/12٪ است. از طرف دیگر، مجموعه داده Malgenome تنها به مقدار دقت 74/70٪ دست یافت. این نشان می دهد که مجموعه داده Virus Total دقت بیشتری دارد و گروه

دقیقی از بدافزار اندروید با الگوریتم کلاستربندی K-Means در مقایسه با مجموعه داده Malgenome استفاده شده است.

2-4-4- خطای مطلق میانگین. شکل 4 خطای مطلق میانگین هر دو مجموعه داده را نشان می دهد. خطای مطلق میانگین بای ارزیابی این مورد استفاده شده که پیش بینی رده چقدر به خروجی نزدیک است. مجموعه داده ای Malgenome نرخ خطای بالاتری با مقدار 25/30٪ در مقایسه با مجموعه داده Virus Total با مقدار 1/88٪ دارد. Virus Total پایین ترین نرخ خطا را دارد، چون ویژگی های منتخب در مجموعه داده شرایط مثبت را برطرف می نمایند. رده بندی مناسبی از ویژگی روی تولید نتیجه تأثیر خواهد گذاشت. بنابراین، یک رده بدافزار اندروید خوب در مجموعه داده مشارکت می کند تا نرخ خطای، نرخ مثبت صحیح (TP) و مثبت غلط (FP) پایینی داشته باشد.



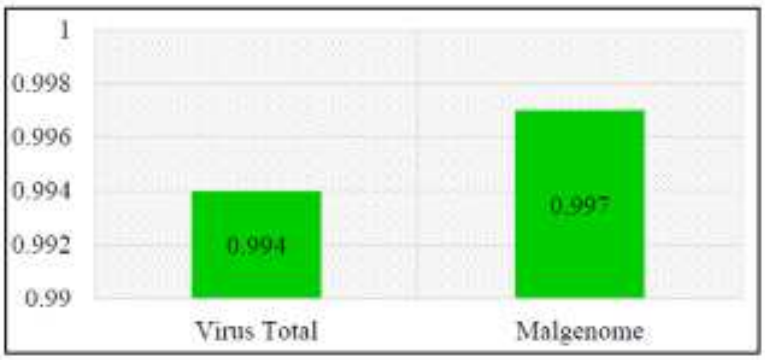
شکل 3 – مقدار دقت برای مجموعه داده های Virus Total و Malgenome



شکل 4 – خطای مطلق میانگین برای مجموعه داده های Virus Total و Malgenome

3-4-4- مشخصه عملیاتی گیرنده (ROC). شکل 5 مقدار مشخصه عملیاتی گیرنده (ROC) ا برای مجموعه داده های Virus Total و Malgenome نشان می دهد. بهترین نتیجه ROC زمانی است که

مقدار ROC نزدیک یک است. مجموعه داده Malgenome بالاترین مقدار ROC به میزان 0/997 در مقایسه با مجموعه داده Virus Total با مقدار 0/994 دارد. هر دو مجموعه داده پراکندگی اندکی در حدود 0/003 از خود نشان می دهند.



شکل 5 – مشخصه عملیاتی گیرنده (ROC)

4-4-4- نرخ مثبت صحیح (TP) و مثبت غلط (FP). جدول 5 نرخ مثبت صحیح (TP) و مثبت غلط (FP) مجموعه داده های Virus Total و Malgenome را نشان می دهد. برای به دست آوردن بهترین نتیجه، تحلیل مجموعه داده ها باید به بالاترین نرخ TP و پایین ترین نرخ FP برسد. مقدار TP نشان می دهد که مجموعه داده به درستی رده بدافزار اندروید خود را تقسیم بندی کرده است. نرخ TP برای مجموعه داده های Virus Total و Malgenome به ترتیب 0/981 و 0/747 می باشد. برای نرخ FP، Malgenome بالاترین مقادیر را با 0/739 نشان می دهد در حالی که برای Virus Total این مقدار 0/004 می باشد.

جدول 5 – نرخ TP و FP برای مجموعه داده های Virus Total و Malgenome

Malgenome		Virus Total	
FP	TP	FP	TP
0/739	0/747	0/004	0/981

5- نتیجه گیری

بدافزار اندروید مشکل جدیدی است که این روزها با آن مواجهیم و حل این مشکل اثبات کرده که بسیار چالش برانگیز است. در این مقاله، ما رویکرد رده بندی بدافزار اندروید را بر مبنای الگوریتم کلاستربندی K-Means با

استفاده از ویژگی های Encryption ،Text Score ،Text Detected ،Lock Detected ،Detected ،Thread ،Porn ،Law ،Copyright و Moneypak به عنوان بردار ویژگی ها ارائه نموده ایم. الگوریتم پیشنهادی پس از آن با دو مجموعه داده تست شده است؛ یعنی مجموعه داده های Virus Total و Malgenome. سپس، از الگوریتم کلاستربندی مبتنی بر قاهده برای گروه بندی بدافزار اندروید به Ransomware، Scareware یا Goodware استفاده کردیم. الگوریتم کلاستربندی مبتنی بر قاعده پیشنهادی نتایج بهترین در هنگام تست روی مجموعه داده Virus Total با بالاترین دقت و پایین ترین خطای مطلق میانگین به میزان 98/12٪ و 1/88٪ به ترتیب نشان می دهد. اما مجموعه داده Malgenome مقدار ROC ای داشت که با مقدار 0/003 اندکی بالاتر از مجموعه داده Virus Total بود. در کل، مجموعه داده Virus Total در هنگام تست با استفاده از رویکرد پیشنهادی با بالاترین مقدار TP و پایین ترین مقدار FP عملکرد خوبی داشت. ما برنامه داریم دیگر ویژگی های بدافزار اندروید و مجموعه داده های تمرین شده روی الگوریتم کلاستربندی مبتنی بر قانون را برای بهبود دقت کلاستربندی بررسی می کنیم.

References

[1] Europol's European Cybercrime Centre, "Police Ransomware Threat Assessment," no. February, 2014.

[2] A. Ajjan, "Ransomware: Next-Generation Fake Antivirus | SophosLabs Technical Paper," 2013. [Online]. Available: https://www.sophos.com/en-us/why-sophos/our-people/technical-papers/ransomware-next-generation-fake-antivirus.aspx.

[3] N. Andronio, S. Zanero, and F. Maggi, "HELDROID: Dissecting and detecting mobile ransomware," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9404, pp. 382–404, 2015.

[4] F. A. Narudin, A. Feizollah, N. B. Anuar, and A. Gani, "Evaluation of machine learning classifiers for mobile malware detection," *Soft Comput.*, pp. 1–15, 2014.

[5] S. Cesare and Y. Xiang, "Classification of malware using structured control flow," *Conf. Res. Pract. Inf. Technol. Ser.*, vol. 107, pp. 61–70, 2010.

[6] Q. Liao, "Ransomware : a Growing Threat To Smes How Ransomware Works ?," no. 2004, pp. 360–366.

[7] T. Rains, "Scareware : Don ' t Let Scammers Scare You," no. May, p. 2012, 2012.

[8] M. Schlesinger and V. Hlavác, "Supervised and unsupervised learning.," *Artif. Intell.*, no. April, 2011.

[9] A. S. Raza Ali, Usman Ghani, "Data Clustering and Its ApplicAtions," 2016. [Online]. Available: http://members.tripod.com/asim_saeed/paper.htm. [Accessed: 19-May-2016].

[10] Sophos, "Stopping Fake Antivirus: How to Keep Scareware Off Your Network," 2011.

[12] I. Burguera, U. Zurutuza, and S. Nadjm-Tehrani, "Crowdroid: Behavior-Based Malware Detection System for Android," *Proc. 1st ACM Work. Secur. Priv. smartphones Mob. devices - SPSM '11*, p. 15, 2011.
[13] Z. Aung and W. Zaw, "Permission-Based Android Malware Detection," *Int. J. Sci. Technol. Res.*, vol. 2, no. 3, pp. 228–234, 2013.
[14] D. D. Hosfelt, "Automated detection and classification of cryptographic algorithms in binary programs through machine learning," 2015.
[15] Y. Mishina, R. Murata, Y. Yamauchi, T. Yamashita, and H. Fujiyoshi, "Boosted random forest," *IEICE Trans. Inf. Syst.*, vol. E98D, no. 9, pp. 1630–1636, 2015.
[16] Dino Sejdinovic, "Statistical Data Mining and Machine Learning," no. 1998, 2006.
[17] R. R. Bouckaert, E. Frank, M. Hall, R. Kirkby, P. Reutemann, A. Seewald, and D. S., "WEKA Manual for Version 3-6-13," 2015.
[18] Computer Crime and Intellectual Property Section (CCIPS). "How To Protect Your Network From Ransomware. [online]. Available: https://www.justice.gov/criminal-ccips/file/872771/download
[19] J. Crowe.," Ransomware Trends and Forecasts" 2017.[online]. Available: https://blog.barkly.com/new-ransomware-trends-2017
[20] K. Chen, "Algorithm On Clustering, Orientering and Conflict-Free Coloring," 2007.