محبه علمی بژو، شی «علوم و فناوری به ی مدافند نوبن» سال پنجم، شماره ۲، تابستان ۱۳۹۳؛ ص ۱۱۹–۱۰۷

شناسایی حملات برنامههای کاربردی تحت وب با استفاده از ترکیب دستهبندهای تک کلاسی

حسین شیرازی (*، امینه جمالی فرد ، سیدمحمدرضا فرشچی ۳

۱- دانشیار ۲- دانشجوی کارشناسیارشد مهندسی کامپیوتر، دانشکده فرماندهی و کنترل، دانشگاه صنعتی مالک اشتر ۳- مربی، دانشکده آمار و علوم رایانه، دانشکده اقتصاد، دانشگاه علامه طباطبایی (دریافت: ۹۲/۱۱/۰۲، پذیرش: ۹۳/۰۵/۰۳)

چکیدہ

بخش مهمی از آمادگی دفاعی کشور در شرایط تهدیدات نامتقارن، اتخاذ راهبردهای دفاعی غیرعامل است. به دلیل گستردگی کاربرد و حساسیت سامانههای تحت وب و با توجه به رشد روزافزون تهدیدات امنیتی، این سامانهها به یکی از آسیب پذیرترین اهداف دشـمن تبـدیل شـدهانـد. کشـف حملات سایبری به مراکز ثقل کشور را میتوان یکی از روشهای بالا بردن آستانه مقاومت ملی دانست. تشخیص ناهنجاری سامانههای تحت وب رویکردی است که بر کشف حملات جدید و ناشناخته تأکید دارد. در این مقاله روشی برای تشخیص ناهنجاری در برنامههای کاربردی تحت وب استفاده از ترکیب دستهبندهای تک کلاسی پیشنهاد شده است. در مرحله آموزش بردارهای ویژگی استخراج شده مرتبط با هـر درخواست HTTP، وارد سامانه شده و نمونه شبیهسازی شده درخواست عادی توسط هر دستهبند یادگیری میشود. سپس با اسـتفاده از روشهای مختلف ترکیب دستهبندهای تک کلاسی، بار دیگر نمونه شبیهسازی شده درخواست عادی توسط هر دستهبند یادگیری میشود. سپس با اسـتفاده از روشهای مختلف ترکیب استراتژیهای مختلف ترکیب، جهت تصمیم گیری گروهی استفاده شده است. نتایج ارزیابیهای کمی و کیفی روش پیشنهادی بر روی پایگاهداد دستهبندهای تک کلاسی، بار دیگر نمونه شبیهسازی شده درخواست عادی THTP به سامانه یادگیری منتقل میشود. برای ترکیب دستهبندها از استراتژیهای مختلف ترکیب، جهت تصمیم گیری گروهی استفاده شده است. نتایج ارزیابیهای کمی و کیفی روش پیشنهادی بر روی پایگاهداده پیشنهادی در استفاده از تصمیم گیری گروهی، معیارهای کارایی سامانه تشخیص ناهنجاری را به خوبی بهبود بخشیده است.

کلید واژهها: امنیت سایبری، سامانههای تحت وب، دستهبندهای تک کلاسی، تصمیم گیری گروهی، عملگر S-OWA.

Detection of Attacks against Web Applications Using Combination of One-Class Classifiers H. Shirazi^{*}, A. Jamalyfard, S. M. R. Farshchi

Mirazi , A. Jamalytard, S. M. K. Farst Malek-Ashtar University of Technology (Received: 22/01/2014; Accepted: 25/07/2014)

Abstract

The passive defence strategies are used to protect the national security in the asymmetric defence conditions. The web application is one of the most widely used tools in the World Wide Web. Because of its dynamic nature, it is vulnerable to serious security risks. The discovery of cyber-attacks can be seen as a method of enhancing national resistance. Anomaly based intrusion detection is an approach that focuses on the new and unknown attacks. A method for anomaly detection in web applications using a combination of one-class classifiers is proposed. In the preprocessing phase, normal HTTP traffic is logged and features vector is extracted from each HTTP request. The proposed method consists of two steps; in the training phase, the extracted features vectors associated with each request enter the system and the model of normal requests, using combination of one-class classifiers, is learned. In the detection phase, anomaly detection operation is performed on the features vector of each HTTP request using the learned model of the training phase. S-OWA operator and other combination methods are used to combine the one-class classifiers. The data used for training and test are from CSIC2012 dataset. The detection and false alarm rates obtained from experiments, shows better results than those obtained by other methods.

Keywords: Cyber Security, Web-Applications, Combination of One-Class Classifiers, S-OWA Operator.

* Corresponding Author E-mail: shirazi@mut.ac.ir

۱. مقدمه

اینترنت را می توان جدیدترین سلاح معاصر و به صورت بالقوه تأثیر گذارترین و ویرانگرترین آنها دانست. امروزه کشورهای جهان تلاش می کنند تا افسار این شبکه سرکش را جهت رام کردن و تحت نظر گرفتن آن به دست گیرند. شاید یکے از بارزترین و مهمترین تهديدات اينترنتي صهيونيستي، توليد ويروس استاكس نت' به منظور اختلال در مراکز هستهای ایران بود که بر اساس گزارشهای مطبوعاتی این ویروس ها پس از تولید توسط شرکت های مهندسی، ماهها در نیروگاههای هستهای دیمونا در صحرای نقب ٔ مورد آزمایش قرار گرفتند. این سلاح سایبری نمونهای قابل توجه از اقدامات و گامهای دشمن در تغییر مفاهیم جنگهای سنتی قلم داد می شود، جنگهایی که از جنگ با توپ و مواد منفجره به جنگ سایبری تبدیل شدهاند [۱]. زمانی که درباره تهدیدات امنیتی برنامههای وب سخن به ميان مي آيد، تهاجم عليه سايتها، سرقت اطلاعات کارتهای اعتباری، حملات منع سرویس به وب سایتها در جهت مستأصل ارائه خدمات و سرویسهای تعریف شده آنان، ویروسها، تروجانها، كرمها و...، در ذهن تداعى مىشود. مىبايست بپذيريم که با توجه به ماهیت برنامههای وب تهدیدات امنیتی متعددی متوجه آنهاست.

دفاع غیرعامل در واقع مجموعه تمهیدات، اقدامات و طرحهایی است که با استفاده از ابزار، شرایط و حتیالمقدور بدون نیاز به نیروی انسانی به صورت خود اتکا صورت گیرد، چنین اقداماتی از یک سو توان دفاعی مجموعه را در زمان بحران افزایش داده و از سوی دیگر پیامدهای بحران را کهش میدهد و امکان بازسازی مناطق آسیبدیده را با کمترین هزینه فراهم می سازد. در حقیقت طرحهای پدافند غیرعامل قبل از انجام مراحل تهاجم و در زمان صلح تهیه و اجرا می شوند. با توجه به فرصتی که در زمان صلح جهت تهیه چنین طرحهایی فراهم می شود ضروری است این قبیل تمهیدات در متن طراحیها لحاظ گردند [۲].

تأمین امنیت برنامههای کاربردی تحت وب را اقدامی دفاعی ضروری در عصر حاضر است. جهت تأمین امنیت منابع، سامانههای تشخیص نفوذ سعی دارند حملات احتمالی به دادهها و منابع محاسباتی سامانههای تحت وب را تشخیص دهد. دو دیدگاه کلی در مسئله تشخیص نفوذ وجود دارد: دیدگاه مبتنی بر امضاء و دیدگاه مبتنی بر تشخیص ناهنجاری. سامانههای مبتنی بر امضاء، خصوصیات و مشخصههای شناخته شده حملات از پیش شناخته شده و مشخص را به کار می برند و از روش ساده تشخیص مبتنی بر قانون استفاده می کنند. این سامانهها به سادگی پیاده سازی می شوند ولی نیازمند دانش اولیه انواع حملات هستند و نمی توانند حملات جدیدی که قبلاً مشاهده نکردهاند را شناسایی کنند، می توان به سادگی آنها را با

نقشههای حمله جدید مورد هجوم قرار داد و بیشتر در حوزههای عمومی و تجاری مورد اقبال و توجه هستند. سامانههای مبتنی بر تشخیص ناهنجاری از هوش مصنوعی، فنون یادگیری ماشینی و داده کاوی برای پردازش اطلاعات تولید شده توسط حسگرهای شبکه، برای کشف رخدادهای غیرعادی شبکه استفاده می کنند. نمونه شبیه سازی شده از رفتارهای عادی سامانه تحت شرایط کنترل شده مرحله آموزش و در بخش برون خط تشخیص ناهنجاری است. هنگامی که این نمونه شبیه سازی شده موضیت شبکه متناوباً با این نمونه شبیه سازی شده می و انحرافها از شرایط طبیعی کشف شوند. این بخش قسمت برخط تشخیص ناهنجاری را تشکیل می دهد.

وظیفه اصلی یک سامانه تشخیص ناهنجاری برنامههای کاربردی تحت وب، تشخیص رفتارهای غیرعادی با توجه به رخدادهای این برنامهها و رفتارهای درخواستهای HTTP میباشد. فرض کنید برنامه کاربردی ما با تعدادی درخواست متخاصم روبروست که با به-کارگیری رویدادهای حملههای مختلف که برخی از آنها برای ما ناشناخته است، سعی در مختل کردن عملیات آن دارند. در این صورت مسئلهای که با آن روبرو هستیم این است که چگونه می وان به دنباله درخواستهای HTTP برنامه کاربردی تحت وب در طول زمان برچسب عادی یا غیرعادی زد. حل این مسئله وظیفه اصلی سامانههای تشخیص نفوذ مرسوم می باشد.

در روشهای مبتنی بر یادگیری ماشین برای تشخیص ناهنجاری، ابتدا رفتار عادی با استفاده از دستهبندهای تککلاسی یادگیری میشود و سپس هر انحرافی از این رفتار عادی به عنوان ناهنجاری در نظر گرفته میشود استفاده از بهترین دستهبندهای تککلاسی یا ترکیبی از آنها برای یادگیری رفتار عادی همواره به عنوان راه حلی پیشروی محققین می باشد.

در این مقاله روشی برای تشخیص درخواستهای HTTP ناهنجار در برنامههای کاربردی تحت وب ارائه می شود که از ترکیبی از دستهبندهای تک کلاسه استفاده می نماید. نرخ تشخیص^۲ و نرخ هشدار نادرست^۴ این روش در مقایسه با سایر روش های ارائه شده، شناسایی مناسب با خطای اندکی را نشان می دهند. ادامه مقاله بدین ترتیب است: در بخش دوم پیشینه تحقیق را مطالعه می شود. در بخش سوم به بررسی برخی تعاریف اولیه مورد ارجاع در مقاله، پرداخته می شود. در بخش چهارم روش پیشنهادی مقاله در تشخیص ناهنجاری در برنامههای کاربردی تحت وب را ارائه داده و در روش منتی بر ترکیب چند دستهبند تک کلاسی بهره گرفته می شود. بدین منظور ابتدا چهار دستهبند تک کلاسی را به صورت مستقل به کار منظور ابتدا چهار دستهبند تک کلاسی را به صورت مستقل به کار

¹ Stuxnet ² Negev Dessert

³ Detection Rate

⁴ False Alarm Rate

عملگر S-OWA را برای ترکیب دستبندها بـ ه کـار بـرده مـیشـود. سپس در بخش پنجم روش پیشنهادی را مـورد ارزیـابی قـرار داده و نتایج حاصل را بیان میشود. در بخش ششم به نتیجهگیـری و بیـان پژوهشهای آینده پرداخته شده است.

۲. پیشینه تحقیق

نخستین سامانه تشخیص نفوذ مبتنی بر تشخیص ناهنجاری، از تخمين پارامتر بيزين درخواست HTTP براي تشخيص ناهنجاري برنامههای کاربردی تحت وب استفاده کرد.کروگل و ویگنا، روشهایی ارائه دادند که روی تحلیل یارامترهای درخواست. ای HTTP تمرکز دارند و اساساً شامل ترکیبی از نمونه شبیهسازی شده های تشخیص مختلف میباشند. این نمونه های شبیه سازی شده روی طول ویژگی ها، توزيع كاراكترى ويژگىها، استنتاج ساختارى، حضور يا عدم حضور ویژگیها و ترتیب ویژگیهای درخواست HTTP تمرکز دارند [۳]. وانگ و استوفلو سامانه تشخیص ناهنجاری شبکه پیشنهاد دادند که از فاصله Mahalanobis به عنوان راهی در تشخیص درخواستهای ناهنجار در مجموعه دادههایی با ویژگیهای چندگانه و از مقیاس دهی هـر متغیر بر مبنای انحراف معیار استاندارد و کوواریانس، استفاده می کند [۴]. وانگ و همکارانش تشخیص دهنده ناهنجاری محتوا بر اساس تحلیل n-gram پیشنهاد دادند که از فیلترهای bloom استفاده می کرد و مقاومت در برابر حملات مشابه و چندریخت را فراهم می کرد [۵]. اینگهام و همکارانش نشان دادند که چگونه می توان سامانهای ارائه داد که با استفاده از الگوریتم استنتاج DFA به همراه ابتکاریهای (کهش اتوماتا، خطر مثبت نادرست را کمینه کند. روش آنها الگوریتم آموزش دارای قابلیت کار با دادههای غیرایستا با طول دلخواه را فراهم می کند [۶]. سانگ و همکارانش ابزاری آماری مبتنی بر یادگیری ماشین برای دفاع در برابر حملات تزریق ارائه کردهاند. این رویکرد ترکیبی زنجیرههای مارکوف را برای نمونه شبیهسازی شده كردن درخواستهاى HTTP و استخراج الگوريتم آموزشي مرتبط به کار بردهاند [۷].

روشی برای امنیت برنامه های کاربردی تحت وب پیشنهاد شده که دیواره آتش برنامه کاربردی تحت وب^۲ نام گرفته است. این روش پس از پیش پردازش داده ها، در موتور تشخیص خود از الگوریتم درخت تصمیم C4.5 استفاده نموده است. علت استفاده از این الگوریتم، کاربرد گسترده آن در حیطه تشخیص نفوذ و موفقیت الگوریتم مبتنی بر درخت تصمیم در مسابقه تشخیص نفوذ مروفقیت الگوریتم شده است. WAF پیشنهادی برای دستهبندی درخواستهای TTP بیان با نمونه های عادی و حمله آموزش داده می شود. در مرحله آموزش، مامانه ارائه می شود. ساختار روش پیشنهادی در شکل (۱) آمده است. نخست، پیش پردازش برای استخراج ویژگی های و برچسب هر

بسته HTTP انجام می شود. سپس، یک سوم مجموع دادگان برای آموزش تشخیص دهنده با استفاده از الگوریتم C4.5 در نرمافزار WEKA؛ به کار گرفته می شود. نرمافزار درخت تصمیمی که موتور تشخیص WAF را بازنمایی می کند، به عنوان خروجی، بازمی گرداند [۳].



شکل NAF .۱ ارائه شده در مرجع [۳]

۳. تعاريف اوليه

آسیب پذیری های تحت وب به عنوان بخش عمده ای در زمینه امنیت سامانه های رایانه ای مطرح می باشند. به منظور شناسایی حملات شناخته شده تحت وب، سامانه های تشخیص نفوذ به تعداد زیادی از امضای حملات مجهز می شوند. متأسفانه، همگام سازی با استفاده از به روزرسانی تغییرات رخ داده در حملات اینترنتی امر بسیار سختی است. همچنین ممکن است آسیب پذیری با نصب برنامه های کاربردی تحت وب مشخصی رخ دهد. به همین دلیل سامانه های تشخیص نفوذ بهتر است با رویکرد تشخیص ناهنجاری پیاده سازی می شوند.

تشخیص ناهنجاری در حیطه مسائلی است که تلاش می شود در میان دادهها الگوهایی که با رفتار از پیش مورد انتظار مغایرت دارنـد، کشف شود. این الگوهای ناهمگون بیشتر مربوط بـه دادههای پـرت، مشاهدات ناسازگار، موارد استثنا، موارد انحرافی، ویژگیهای گمراه کننده و فعالیتهای مختل کننده در زمینههای کاربردی گوناگون می باشد. واژههای داده پرت⁷ و داده ناهنجار¹ بـه صورت متناوب در متون تخصصی این حیطه به جای یکدیگر به کار می روند که معادل هم می باشند [۸].

برنامه کاربردی تحت وب هر نوع برنامه کاربردی است که از مرورگر وب بـه عنـوان سـرویس گیرنـده اسـتفاده مـیکنـد. تمـامی پایگـاههـای موجـود بـر روی اینترنـت از پروتکـل HTTP اسـتفاده مینمایند. با این که پروتکل HTTP با استفاده از پروتکلهای دیگری

¹ Heuristics

² Web Application Firewall (WAF)

³ Outlier

⁴ Anomaly

نظیر IP و TCP مأموریت خود را انجام میدهد، ولی ایس پروتکل HTTP است که به عنوان زبان مشترک ارتباطی بین سرویس گیرنده و سرویس دهنده وب به رسمیت شناخته شده و از آن استفاده می شود. در واقع مرور گر وب صدای خود را با استفاده از پروتکل HTTP به گوش سرویس دهنده وب می رساند و تقاضای سرویس مي کند [۹].

درخواست HTTP مجموعه ای از خطوط متنی است (با CRLF از یکدیگر جدا شدهاند) که به سرویس دهنده وب ارسال میشود و شامل خط درخواست'، قسمتهای سرپیام درخواست' و بدنه درخواست میباشد. خط درخواست از سه بخش تشکیل شده که با فاصله از یکدیگر جدا شدهاند. این سه بخش نام روشی که می بایست اعمال شود، مسیر محلی منبع درخواست و نسخه پروتکلی که مورد استفاده قرار می گیرد را مشخص می کند. نخستین کلمهای که در درخواست HTTP ظاهر می شود کلمه method است. بیشتر درخواست.های HTTP از نوع روش GET هستند ولی انواع روش.های دیگری همانند POST و HEAD نیز وجود دارند. بعد از method، مسیر منبع (URI) ذکر می شود که عموماً یک فایل، یک فهرست در سامانه فایل یا ترکیبی از هر دو است. آخرین بخش، نسخه پروتکل استفاده شده توسط سرویس گیرنـده را مشـخص مـی کنـد (عمومـاً .(HTTP/1.1 L HTTP/1.0

خط در خواست به طور معمول به شکل زیر است:

GET / path/to/file/index. Html HTTP/1.1

در ادامه خط درخواست اولیه در درخواست HTTP، قسمتهای سرپیام درخواست وجود دارند که اطلاعاتی درباره درخواست هستند. خطوط قسمت سرپیام به فرمت سرپیام عادی هستند: یک خط برای هر سرپیام به صورت "مقدار: نام سرپیام"[†] که با CRLF خاتمه مییابد. در 1.0 HTTP به طور معمول ۱۶ سرپیام وجود دارد، با این وجود هیچ یک اجباری نیستند. HTTP 1.1 با ۴۶ سرپیام مشخص میشود، که تنها سرپیام Host در درخواست اجباری است. سرپیامهای درخواست مجموعهای از خطوط اختیاری هستند که اطلاعاتی اضافی درباره درخواست، سرویس گیرنده و یا هـر دو ارائـه میدهند (جستجوگر، سیستم عامل و غیره). هر یک از این خطوط از نامی تشکیل شده که نوع سرپیام را مشخص می کند و با (:) و مقدار سرپيام دنبال مي شود [١٠].

۴. الگوی شبیهسازی شده پیشنهادی

۴–۱. ساختار

در حوزه تشخیص ناهنجاری در برنامه های کاربردی تحت وب که مورد بحث ما در این پژوهش میباشد، تنها مجموعهدادگان

درخواستهای HTTP عادی در دسترس است و در مرحله آموزش ما می خواهیم رفتار عادی در خواستهای HTTP را شبیه سازی کنیم تا در مرحله تشخیص، درخواستهای ورودی به سامانه پیشنهادی ما با این الگوی شبیهسازی شده عادی مقایسه شوند. برای تهیه الگوی نمونه شبیهسازی شده عادی از دستهبندهای تککلاسی استفاده می شود. در واقع هر کدام از دستهبندهای تک کلاسی را به صورت مستقل، یک سامانه تشخیص ناهنجاری برای درخواستهای HTTP در نظر گرفته و مراحل آموزش و تشخیص را برای هر یک انجام شده است.

۲-۴. استخراج ویژگی

یکی از مهم ترین ملزومات برای ارائه یک سامانه تشخیص ناهنجاری برنامههای کاربردی تحت وب بر مبنای پروتکل HTTP، شاخت رفتار عادی این پروتکل است تا بتوان ناهنجاریها و حملهها را که به عنوان انحراف از حالت عادی تعریف می شود، تشخیص داد. گام نخست در شناخت رفتار عادی پروتکل HTTP، توصیف دقیق و جامع ویژگیها و رفتار آن میباشد. این توصیف اغلب با تعریف ویژگی صورت می گیرد و به تبع آن رفتار عادی به عنوان قیدی روی مقدار ویژگیهای تعریف شده یا رابطهای مابین ویژگیها تعریف می شود. برای تعریف ویژگیهایی از درخواستهای HTTP که در مسئله تشخيص ناهنجاري تعيين كننده باشند، نيازمند شناخت حملات و نحوه تأثير آنها روی بخشهای مختلف این درخواستها است. با استفاده از دانش خبره در حملات وب، ۲۸ ویژگی مؤثر در تشخیص ناهنجاری شناسایی شده است [۳ و ۱۱] (جدول (۱)).

| نام ویژگی | نام ویژگی |
|------------------------------------|---------------------------------|
| طول سرپيام "Accept-Charset" | طول قسمت Path |
| طول Header | طول سرپيام "Accept" |
| طول سرپيام"Accept-Encoding " | طول درخواست |
| طول سرپيام "Accept-Charset" | طول سرپيام "Cookie" |
| طول سرپيام "Accept-Language" | طول سرپيام "Content-Type" |
| طول سرپيام "Content-Length " | طول سرپيام "Referrer " |
| طول Host | شناسه متد |
| طول سرپيام "User-Agent" | تعداد کاراکترهای خاص در Header |
| تعداد آرگومانهای درخواست | تعداد کاراکترهای دیگر در Header |
| تعداد اعداد در Header | تعداد حروف در path |
| تعداد حروف در Header | تعداد کاراکترهای خاص در Path |
| تعداد اعداد در قسمت Path | Min ASCII char در Min ASCII |
| تعداد کاراکترهای دیگر در قسمت Path | Max ASCII char در Request |
| تعداد Cookie ها | تعداد بایتهای متمایز |
| | |

جدول ۱. ویژگیهای درخواست HTTP [۳].

هر کدام از این ویژگیها به نوعی در حملات متداول تحت وب، تحت تأثير فعاليتهاى مخرب مهاجمان قرار گرفتهاند. بعضى ویژگیها به طول درخواست، طول قسمت Path یا Header بستگی دارد زیرا طول قسمتها برای تشخیص حملات سرریز بافر اهمیت

¹ Request Line

² Request Header Fields

³ Request Body ⁴ Header-Name: Value

دارند. همچنین مشاهده شده که کاراکترهای غیرالفبایی-غیرعددی در بسیاری از حملات سرریز مشاهده شدهاند. با این حال، چهار گونه کاراکتر در این فهرست لحاظ شده است: حروف، اعداد، کاراکترهای غیرالفبای-غیرعددی و سایر کاراکترها. کاراکترهای غیرعددی-غیرالفبایی معنای خاصی در تعدادی از زبانهای برنامهنویسی دارند و این کاراکترها در جدول (۱)، کاراکترهای خاص نامیده شدهاند.

۴–۳. دستەبندھای تککلاسی

در فرآیند آموزش در روشهای دستهبندی دو یا چند کلاسی، داده های مربوط به همه کلاسها موجود است. در صورتی که در تشخیص ناهنجاری تک کلاسی، هنگام توصیف رفتار عادی هیچ مجموعهداده حملهای وجود ندارد و در فرآیند آموزش فقط دادههای مربوط به یک کلاس (کلاس رفتار عادی که به طور عامتر کلاس هدف نامیده میشود) موجود است [17]. در این گونه مسائل مجبور به استفاده از دستهبندیهای تک کلاسی بوده تا بتوان مشخصات یک کلاس موجود را یادگیری نماییم.

در روشهای دستهبندی تککلاسی دو مؤلفه آصلی باید مشخص شود. مؤلفه اول عبارت است از اندازه گیری مقدار فاصله ((x) یا شباهت ((p(x)) یک شیء x در فضای ویژ گی به کلاس هدف (کلاس رفتار عادی) و مؤلفه دوم عبارت است از حد آستانه روی مقدار فاصله یا شباهت. در فرآیند تشخیص، یک شیء جدید x برچسب عادی میخورد اگر فاصله آن از کلاس عادی کوچکتر از حد آستانه باشد (θ>(x)) یا شباهت آن به کلاس عادی بزرگتر از حد آستانه باشد (θ(x)(2)). دستهبندیهای تککلاسی را با توجه به روشی که در حل مسئله دستهبندی تککلاسی به کار می گیرند و نمونه شبیهسازی شده که از کلاس هدف ارائه میکنند در سه گروه قرار میدهد، روشهای مبتنی بر مرز (مانند SVD و SVD)، روشهای مبتنی بر چگالی (FOA) و روشهای مبتنی بر دوبارهسازی^۱ (مانند SOM). (مانند SOM) و روشهای مبتنی بر دوبارهسازی^۱ (مانند SOM).

در این مقالـه بـرای یـادگیری رفتـار عـادی درخواسـت HTTP از دستهبندیهای تککلاسی زیر استفاده می شود:

۱- دستهبندی تک کلاسی SVDD : یک ابر کره را بر دادههای کلاس موجود (به عنوان کلاس هدف) محاط می کند. محدوده ابر کره توسط اشیائی از کلاس هدف تعیین می شود. این اشیاء بردارهای پشتیبان نامیده می شوند. در SVDD فاصله شیء x از کلاس هدف طبق رابطه زیر محاسبه می شود:

$$D_{SVDD}(x) = k(x, x) - 2 \sum_{i} \alpha_{i} * k(x, x_{i})$$

+
$$\sum_{i,i} \alpha_{i} \alpha_{j} * k(x_{i}, x_{j}) D_{SVDD}(x)$$
(1)

که در آن، k نشان دهنده تابع هسته، $x_i e_i g_i$ ضریب لاگرانژ منتسب به بردار پشتیبان x_i است.

۲- اختلاط نمونههای شبیهسازی شدههای گوسی (MOG)³: اختلاط نمونه شبیهسازی شدههای گوسی یک ترکیب خطی از توزیع نرمال است که تابع چگالی آن طبق رابطه زیر به دست میآید:

$$P_{MOG}(x) = \frac{1}{N_{MOG}} \sum_{i} \alpha_i P_N(x; \mu_i, \Sigma_i)$$
(Y)

که در آن، ۵٫ ضریب اختلاط است. MOG بایاس کمتری نسبت به یک تابع توزیع نرمال دارد و در عوض به دادههای بیشتری برای آموزش نیاز دارد. در صورتی که MOG با دادههای کمتری آموزش داده میشود واریانس بیشتری از خود نشان میدهد. وقتی تعداد نمونه شبیه سازی شدههای گوسی، N_{MOG}، مشخص باشد، میانگین و کوواریانس هر کدام از نمونه شبیه سازی شدههای گوسی با روش بیشینه سازی امید ریاضی^۳ تعیین میشود.

۳- تصمیم چگالی پارزن^³ (PDE): تخمین چگالی پارزن روشی برای تخمین چگالی احتمال یک متغیر تصادفی میباشد. برای هر شیء x، چگالی تخمینی از رابطه زیر به دست میآید:

$$P_{PDE}(x) = \frac{1}{N} \sum_{i} K_h(x - x_i) \tag{(7)}$$

که تابع هسته مورد استفاده (اغلب گوسی)، N تعداد اشیاء موجود در مجموعهداده آموزش، i، x_i امین شیء موجود در مجموعهداده آموزش و h عرض هسته است که با آموزش و با استفاده از روش بیشینه مقدار احتمال⁶ تعیین می شود [۱۲].

۴- ماشین بردار پشتیبان (SVM): این روش ابرصفحه هایی با حداکثر حاشیه را به دست میآورد که دسته های داده ها را از هم جدا کنند. هدف، پیدا کردن بهترین خط (ابر صفحه) که دو دسته را از هم جدا کند. در حالت دو بعدی معادله این خط به صورت زیر است:

 $w_1 X_1 + w_2 X_2 + b = 0$

در حالت n بعدی خواهیم داشت:

(۴)

$$\sum_{i=0}^{n} w_i x_i + b = 0 \tag{(a)}$$

نمونه شبیهسازی شده سامانه تشخیص ناهنجاری پیشنهادی برای هر درخواست HTTP در شکل (۲) نشان داده شده است.

برای تشخیص حمله، با استفاده از یادگیری ماشین و با رویکرد تشخیص ناهنجاری ابتدا نمونه شبیه سازی شده رفتار عادی برنامه کاربردی تحت وب بر مبنای پروتکل HTTP یادگیری شده و سپس با اعمال درخواستهای HTTP، انحراف از حالت عادی اندازه گیری

² Mixture of Gaussian Model

³ Expectation-Maximization

⁴ Parzen Density Estimator

⁵ Maximum Likelihood

¹ Reconstruction

می شود. در روش پیشنهادی، مرحله آموزش دستهبندها به صورت برون خطی انجام می شود. به عبارت دیگر نمونه شبیه سازی شده رفتار عادی در خواست ها مبتنی بر پروتکل HTTP، توسط دستهبندهای تک کلاسی قبل از شروع به استفاده از برنامه کاربردی تحت وب یادگیری می شود و سپس از نمونه شبیه سازی شده های یادگیری شده در هنگام کارکرد برنامه کاربردی برای تشخیص ناهنجاری استفاده می شود. روش پیشنهادی، در مرحله یادگیری رفتار عادی برنامه کاربردی تحت وب، نیازی به مجموعه حاده حمله ندارد و صرفاً از مجموعه دادگان رفتار عادی برنامه کاربردی برای ساختن مرزهای تصمیم بهره می گیرد.



شکل ۲. آموزش و تشخیص سامانه تشخیص ناهنجاری.

ما در روش پیشنهادی خود، مسئله تشخیص ناهنجاری را به صورت مسئله تصمیم گیری گروهی دنبال می کنیم و روش های متداول ترکیب و نوعی روش میانگین مرتب شده وزندار^۱ (OWA) موسوم به AWA-۵، را در آنها به کار می گیریم. در حالت کلی، فرآیند تصمیم گیری گروهی عبارت است از حالتی که دو یا چند متخصص، هر کدام با عقاید و ویژگیهای منحصر بفرد خود سعی می کنند تا یک تصمیم مشترک بگیرند. مهم ترین مسئلهای که در تصمیم گیری گروهی مطرح می شود این است که چگونه نظرات متخصصین با هم ترکیب شود طوری که تصمیم گرفته شده در جهت ارضای معیار مشخصی باشد [۱۱].

دستهبندهای تک کلاسی به سختی میتوانند تمامی مشخصات^۲ داده را در نظر بگیرند. ترکیب دستهبندها به همین منظور مطرح میشود. محققان به طور مستمر به دنبال بهبود کارایی روشهای پیشنهادی در مسایل دستهبندی میباشند و ترکیب دسته بندها یکی از راههای دستیابی به این هدف است. ترکیب دستهبندها منجر به بهبود کارایی و استحکام دستهبندی در ازای افزایش پیچیدگی میشود. فرض کنید برای فرآیند یادگیری، یکی از دستهبندها با

توجه به قدرت آن دستهبند در تشخیص حمله های موجود انتخاب شود و در آینده حمله ای جدید در شبکه اعمال شود که در نقطه کور دستهبند مورد استفاده قرار داشته باشد، در نتیجه حمله تشخیص داده نخواهد شد. در صورتی که استفاده از چند دستهبند، به شرط اینکه دستهبندهای انتخاب شده رویکردهای یادگیری متفاوتی داشته باشند و مکمل یکدیگر باشند، احتمال مواجهه با حالت مذکور را کاهش میدهد و نقطه کور یک دستهبند با دستهبندهای دیگر پوشش داده می شود.

روشهای مختلفی مانند میانگین گیری، رأی اکثریت، انتخاب دستهبند کمینه، انتخاب دستهبند بیشینه، انتخاب دستهبند میانه و قالبهای تصمیم نظیر S-OWA برای ترکیب خروجی دستهبندها پیشنهاد شده است.

۴-۴. استراتژیهای ترکیب

(6)

مسئله بازشناسی الگویی را در نظر میگیریم که الگوی Z به یکی از m کلاس ممکن (w₁,...,w_m) تعلق گیرد. فرض می شود R دسته بند داریم که هر یک الگوی Z را با بردار اندازه گیری^۲ مشخصی بازنمایی می کنند. بردار اندازه گیری مرتبط با دسته بند i ام با x نشان داده می شود. در فضای اندازه گیری هر کلاس w با تابع چگالی احتمال شبیه سازی می شود و احتمال اولیه رخداد^۴ آن با (w_k) منایش داده می شود. فرض می شود نمونه های شبیه سازی شده متقابلاً منحص به فرد⁶ هستند که این بدان معنی است که در نهایت تنها یک نمونه شبیه سازی شده با هر الگو مرتبط است.

حال، با توجه به تئوری بیز، اندازهگیریهای داده شده x_i ،R، ،R، الگوی Z میبایست به کلاس wi=1...R الگوی Z میبایست به کلاس w_i تعلق بگیرد کـه احتمـال ثانویـه آن تعبیر به صورت زیر بیشینه شود:

Assign $z \to w_j$ if

$$P(w_j | x_1, \dots, x_R) = argmax_k P(w_k | x_1, \dots, x_R)$$

قانون تصمیم بیز فوق بیان می دارد برای استفاده صحیح از تمامی اطلاعات در دسترس به منظور تصمیم گیری، ضروری است احتمالات فرضیات متعددی، با در نظر گرفتن تمامی اندازه گیریها در یک زمان، محاسبه شود. این مطلب، عبارتی صحیح در مسئله کلاس بندی است ولی ممکن است راهکاری عملی نباشد. محاسبه توابع احتمال ثانویه به دانشی از آمار اندازه گیری در سطح بالاتری بستگی دارد که به صورت توابع چگالی احتمال پیوسته $(x_1 ... x_R w_k)$ هستند و به سختی قابل استنتاج می باشند. به همین دلیل می بایست قانون فوق را ساده سازی کرد و آن را به صورت محاسبات تصمیم گیری انجام شده توسط دسته بندهای تکی، که تنها شامل اطلاعات موجود در بردار _نx بیان کرد. مشاهده می شود که این امر نه تنها قانون تصمیم بیز

¹ Order Weighted Averaging

² Characteristics

³ Measurement Vector

⁴ Priori Probability of Occurrence

⁵ Mutually Exclusive

⁶ Joint Probability Density Function

را قابل حل می کند، بلکـه ترکیـب دسـتهبنـدها کـه در عمـل از آنهـا استفاده می شود را نیز ممکن می سازد. بهعلاوه این رویکرد حیطـهای را برای گسترش استراتژی های ترکیب دستهبندها، فراهم می آورد.

می توان از قانون تصمیم بیز شروع کرد و آن را با در نظر گرفتن فرضیات مشخصی، روشن تر کرد. ابتدا احتمال ثانویه (P(w_k l x₁, ..., x_R) را با استفاده از تئوری بیز بازنویسی میکنیم. خواهیم داشت:

$$P(w_k | x_1, ..., x_R) = \frac{p(x_1 ... x_R | w_k) P(w_k)}{p(x_1, ..., x_R)}$$
(Y)

که در آن، (p(x₁, ..., x_R) اندازهگیری غیر شرطی چگالی احتمال پیوسته است. این احتمال را میتوان به صورت توزیعهای اندازهگیری شرطی به صورت زیر نوشت:

$$p(x_1, \dots, x_R) = \sum_{j=1}^m P\left(x_1, \dots, x_R | w_j\right) P(w_j) \tag{A}$$

از این رو تنها میتوان روی عنصر شمارنده رابطه (۷) حساب کرد.

$$p(x_1, ..., x_R | w_k) = \prod_{i=1}^R p(x_i | w_k)$$
(9)

که در آن، (p(x_i l w_k)، نمایه فرآیند اندازه گیری i امین بازنمایی است. با جایگذاری (۸ و ۹) در (۷) خواهیم داشت:

$$P(w_k \mid x_1, \dots, x_R) = \frac{P(w_k) \prod_{i=1}^R p(x_i \mid w_k)}{\sum_{j=1}^m P(w_j) \prod_{i=1}^R p(x_i \mid w_j)}$$
(1.)

assign $Z \rightarrow w_j$ if

$$P^{-(R-1)}(w_{j}) \prod_{i=1}^{K} \left(P(w_{j} | x_{i}) \right)$$

= $\max_{k=1,..,m} P^{-(R-1)}(w_{k}) \prod_{i=1}^{R} P(w_{k} | x_{i})$ (11)

א بیان احتمالاتی ثانویه حاصل شده از دستهبندهای مورد بحث خواهیم داشت: assian $Z o w_i$ if

assign
$$Z \rightarrow W_j$$
 if

$$P^{-(R-1)}(W_j) \prod_{i=1}^R P(W_j | x_i)$$
(17)

 $\max_{k=1,...,m} P^{-(R-1)}(w_k) \prod_{i=1}^{i=1} P(w_k | x_i)$ قانون تصمیم گیری (۱۲) در ستنمایی^۲ فرضیه را بـا ترکیـب احتمـالات ثانویه تولید شده توسط دستهبندهای تکی را با اسـتفاده از قـانون حاصـل ضرب بیان میکند. این قانون برای ترکیب خروجی دسـتهبنـدها بسـیار

کارآمد است. به این صورت که در موتور تشخیص ترکیبی بـا نزدیـک بـه صفر کردن خروجی احتمـال یـک بازنمـایی، از آن جلـوگیری مـیکنـد. همانطور که در ادامـه ایـن بخـش خـواهیم دیـد، ایـن امـر در ترکیـب قانونهای تصمیمگیری نامطلوب است. زیرا تمامی دستهبندها میبایسـت برای هر شناسه کلاس مفروض گزینه رد یا قبول را تولید کنند.

قانون حاصل جمع ⁷: اگر بخواهیم قانون (۱۱) را بیشتر مورد بررسی قرار دهیم، در برخی از کاربردها میبایست این فرض را مورد نظر قرار داد که احتمالات ثانویه محاسبه شده توسط هر دستهبند به سادگی از احتمالات اولیه به دست نمی آید. یکی از دلایل این فرض این است که مشاهدات به دست آمده به دلیل نویز زیاد مبهم باشند. در این وضعیت میتوان فرض کرد که احتمالات ثانویه را به صورت زیر میتوان نشان داد:

$$P(w_k|x_i) = P(w_k(1+\delta_{ki})) \tag{17}$$

که در آن، 1 » δ_{ki} است. با جایگذاری (۱۲) در (۱۱) به عنوان احتمال ثانویه خواهیم داشت:

$$P^{-(R-1)}(w_k) \prod_{i=1}^{R} P(w_k | x_i) = P(w_k) \prod_{i=1}^{R} (1 + \delta_{ki})$$
(17)

اگر قانون حاصل ضرب را گسترش دهیم و از هر عبارت درجه دوم و بیشتر صرف نظر کنیم، میتوان سمت راست (۱۳) را به صورت زیر بازنویسی کرد:

$$P(w_k) \prod_{i=1}^{R} (1 + \delta_{ki}) = P(w_k) + P(w_k) \sum_{i=1}^{R} \delta_{ki}$$
(14)

داری (۱۱ و ۱۱) در (۱۱) خواهیم داشت:

assign $Z \rightarrow w_j$ if

$$(1-R)P(w_{j}) + \sum_{i=1}^{R} P(w_{j} | x_{i})$$

$$= \max_{k=1,\dots,m} \left[(1-R)P(w_{k}) + \sum_{i=1}^{R} P(w_{k} | x_{i}) \right]$$
(1 Δ)

قوانین تصمیم گیری (۱۱ و ۱۴) طرح اولیه برای ترکیب دستهبندها را نشان میدهند. بسیاری از استراتژیهای ترکیب را از این قوانین و با در نظر گرفتن (۱۶) میتوان پیشنهاد کرد:

$$\prod_{i=1}^{R} P(w_{k}|x_{i}) \leq \min_{i=1,\dots,R} P(w_{k}|x_{i})$$

$$\leq \frac{1}{R} \sum_{i=1}^{R} P(w_{k}|x_{i}) \leq \max_{i=1,\dots,R} P(w_{k}|x_{i})$$
(19)

رابطه (۱۶) پیشنهاد میکند که قوانین ترکیب حاصل جمع و حاصل ضرب را به وسیله باندهای پایینی و بالایی تقریب زده شود. همچنین می توان با سختسازی^۴ احتمالات ثانویه (P(w_k|x_i)، توابع دو مقداری Δ_{ki} را به صورت زیر تولید کرد:

¹ Conditionally Independent

² Likelihood

³ Sum Rule

⁴ Hardening



از این نتیجه می توان در خروجی ترکیب دسته بندها به جای تركيب احتمالات ثانويه استفاده كرد. با اين تقريب بقيه قوانين ترکیب را میتوان به دست آورد.

قانون میانه : با فرض احتمالات اولیه مساوی، قانون حاصل جمع در (۱۵) را میتوان برای محاسبه میانگین احتمال ثانویه برای هر کلاس دربرگیرنده خروجیهای همه دستهبندها، بهکار برد. به عنوان مثال:

assign
$$Z \to w_j$$
 if

$$\frac{1}{R} \sum_{i=1}^{R} P(w_j | x_i) = \max_{k=1,\dots,m} \frac{1}{R} \sum_{i=1}^{R} P(w_k | x_i) \qquad (1)$$

بنابراین، قانون الگو را به کلاسی متعلق میداند که میانگین احتمال ثانویه آن بیشینه شود. چنانچه خروجی یکی از دستهبندها احتمال ثانویه ای باشد که متعلق به کلاس ناهنجار است، روی میانگین تأثیر می گذارد و ممکن است سبب تصمیم گیری ترکیبی نادرستی شود. این مسئله که تخمین قوی از میانگین، میانه است، امری شناخته شده میباشد. به همین دلیل مناسبتر است که تصمیم ترکیبی بر مبنای استفاده از میانه احتمالات ثانویه باشد تا میانگین آنها. این امر قانون زیر را باعث میشود:

assign
$$Z \to w_j$$
 if

$$\max_{=1,...,R} P(w_j | x_i) = \max_{k=1,...,m} \max_{i=1,...,R} P(w_k | x_i)$$
(19)

قانون رأى اكثريت : با شروع از ١٥ با فرض مساوى بودن احتمال اولیه و با سختسازی احتمالات با توجه به (۱۷) خواهیم داشت:

assign $Z \rightarrow w_i$ if

$$\sum_{i=1}^{R} \Delta_{ji} = \max_{k=1,\dots,m} \sum_{i=1}^{R} \Delta_{ki}$$

$$(\Upsilon \cdot)$$

برای هر کلاس w_k حاصل جمع سمت راست تساوی ۲۰ به سادگی با شمارش رای دریافت شده برای این فرضیه از دستهبندهای تکی به دست میآید. کلاسی که بیشترین تعداد رأی را داشته باشد به عنوان رأى اكثريت انتخاب مي شود [١٣].

استفاده از روش تولید وزنهای S-OWA برای ترکیب دستهبندها: با استفاده از روش تولید وزنهای S-OWA میتوان نظرات دستهبندها با یکدیگر ترکیب کرد. ما برای ترکیب دستهبندهای تکی ذکر شده از عملگر S-OWA استفاده نمودیم و این روش را برای جمعبندی نظر دستهبندها درباره درخواستهای HTTP به کار بستەايم.





$$\frac{1}{2}\sum_{i=1}^{n}w_{i} = \frac{1}{2}$$
(71)

برای به دست آوردن تعریف جدید، این فرمول را تغییر میدهـیم و از تساوی زیر استفاده می کنیم:

$$\frac{1}{2}\sum_{i=1}^{n}w_{i} = \frac{1}{2}$$
(YY)

فرمول درجه orness را به صورت زیر بازنویسی میکنیم:

orness(W) =
$$\frac{1}{2} + \sum_{i=1}^{n} (\frac{(n-i)}{(n-1)} - \frac{1}{2})w_i$$

= $\frac{1}{2} + \sum_{i=1}^{n} \frac{(n-2i+1)}{2(n-1)}w_i$
orness (W) = $\frac{1}{2} + \sum_{i=1}^{n} q_i w_i$

حال در نظر می گیریم وضعیتی را که n زوج باشد، n = 2m ؛ همچنين k <= m ، i = k ن فاشت: i = n+1-k و

$$q_{k} = \frac{(n-2k+1)}{2(n-1)}$$

$$q_{n+1-k} = \frac{n-2n-2+2k+1}{n-1}$$
(YF)

$$= \frac{-n+2k-1}{2(n-1)} = -q_k$$

$$orness(W) = \frac{1}{2} \sum_{i=1}^{n} q_k (w_k - w_{n+1-k})$$
 (Y Δ)

اگر n فرد باشد، بنابراین n = 2m+1 و خواهیم داشت:

همچنين

orness(W) =
$$\frac{1}{2} \sum_{i=1}^{n} q_k (w_k - w_{n+1-k})$$
 (YF)

 $+ q_{m+1} w_{m+1}$

¹ Median Rule

² Majority Vote Rule

$$q_{m+1} = \frac{2m+1-2(m-1)+1}{2(n-1)} \tag{YV}$$

پس برای orness خواهیم داشت:

$$orness(W) = \frac{1}{2} \sum_{i=1}^{n} q_k (w_k - w_{n+1-k})$$
 (YA)

با استفاده از این عبارت مستقیماً روشی را برای ساختن عملگرهای S-OWA با وزنهایی با درجـه orness از پـیش تعیـین شـده بیـان میکنیم.

فرض می کنیم درجه orness با نام Ω از پیش داده شده باشد. می توانیم فرض کنیم که تمامی وزنهای به جز w_n و w_n مساوی باشند. با این فرض تابع orness به سادگی به صورت زیر درمی آید:

$$orness(W) = \frac{1}{2} + q_1 (w_1 - w_n) = \frac{1}{2} + \frac{1}{2} (w_1 - w_n)$$
(۲۹)

با درجه orness از پیش تعیین شده Ω میتوان به تعریف واضحی برای تفاوت میان اولین و آخرین وزن رسید:

$$w_1 - w_n = 2(\Omega - 0.05) \tag{(7.)}$$

می توان ₁ w و w_n را هر عددی در بازه بین صفر و یک انتخاب کرد به طوری که شرط فوق را برآورده نمایند. سـپس مجموع وزنهای باقیمانده میبایست بین صفر و یک باقی بماند. بنابراین داریم:

$$vi = \frac{1}{n} [1 - (w1 - wn)], i = 2, 3, ..., n - 1$$
 (71)

 $1. \Delta = 2(\Omega - 0.05)$ 2. Let $L = \frac{1}{n} (1 - |\Delta|)$ 3. for i = 2, ..., n - 1 $w_i = L$ $4. if \Delta > 0 then$ $w_1 = L + \Delta, \quad w_n = L$ $if \Delta \le 0 then$ $w_1 = L, \quad w_n = L + L$

با چنین فرآیند وزنهای S-OWA تولید می شود. در واقع ما پس از تولید وزنهای S-OWA تولید می شود. در واقع ما پس از تولید وزنهای تولید شده تولید وزنهای عملگر S-OWA را اختصاص می دهیم. برای ما در این پژوهش چنانچه $0 < \Delta$ باشد، $\Delta + L = k_1$ و $M_i = L + \lambda$ برای پژوهش چنانچه $0 < \Delta$ باشد، $\Delta + d = k_1$ و $N_i = 1, .. n$ و i = 1, .. n واقع خروجی حاصل از ترکیب خروجی دسته بندهاست خواهیم داشت:

$$F(a_1, \dots a_n) = \Delta Max_i[\alpha_i] + L \sum_{i=1}^n \alpha_i$$
 (°Y)

 $= \Delta Max_i[\alpha_i] + \frac{(1-\Delta)}{n} \sum_{i=1}^n \alpha_i$

بنابراين خواهيم داشت:

$$F(a_1, \dots a_n) = \Delta Max_i[\alpha_i]$$

$$+(1 - \Delta) Ave(a_1, \dots, a_n)$$
(TT)

فرمول فوق عملگر S-OWA نامیده شده است [۱۵]. اگر نتایج حاصله به گونهای باشد که 0 => ∆ باشد خواهیم داشت:

$$F(a_1, \dots a_n) = \Delta Min_i[\alpha_i] \tag{(3.4)}$$

$$+(1-\Delta) Ave(a_1,\ldots,a_n)$$

۴-۵. الگوریتم پیشنهادی

همانطور که بیان شد، در تشخیص ناهنجاری برنامههای کاربردی تحت وب، تنها مجموعهدادگان درخواستهای HTTP عادی در دسترس است و در مرحله آموزش ما میخواهیم رفتار عادی درخواستهای HTTP را شبیهسازی شده کنیم تا در مرحله تشخیص، درخواستهای ورودی به سامانه پیشنهادی ما با این نمونه شبیهسازی شده عادی مقایسه شوند. دیدیم برای تهیه نمونه شبیهسازی شده عادی مقایسه شوند. در این بخش به جای استفاده دستهبندهای تککلاسی استفاده شد. در این بخش به جای استفاده از هر کدام از دستهبندهای تککلاسی به صورت مستقل، به عنوان یک سامانه تشخیص ناهنجاری برای درخواستهای باین سامانه ترکیب نظرات این دستهبندها استفاده میکنیم و با این سامانه ترکیب مراحل آموزش و تشخیص را انجام میدهیم.

در معماری سامانه ترکیبی پیشنهادی نیز برای تشخیص ناهنجاری، ابتدا نمونه شبیهسازی شده رفتار عادی برنامه کاربردی تحت وب بر مبنای پروتکل HTTP با استفاده از دستهبند ترکیب یافته از دستهبندهای تککلاسی، یادگیری شده و سپس با اعمال درخواستهای HTTP، انحراف از حالت عادی اندازه گیری می شود. در اینجا نیز، مرحله آموزش دستهبند به صورت برون خطی انجام می شود. به عبارت دیگر آموزش دستهبند به صورت برون خطی انجام می شود. به عبارت دیگر HTTP، توسط دسته بند تککلاسی ترکیبی قبل از شروع به استفاده از شده در هنگام کارکرد برنامه کاربردی برای تشخیص ناهنجاری استفاده می شود. همچنین، در مرحله یادگیری رفتار عادی برنامه کاربردی تحت وب، به مجموعه داده حمله دسترسی وجود ندارد و صرفاً از مجموعه داده رفتار عادی برنامه کاربردی برای سخیص ناه بهره گیری می شود.

شکل (۴) نحوه عملکرد هر درخواست HTTP در روشهای ترکیبی را نشان میدهد. درخواستهای تولید شده مربوط به هر برنامه کاربردی، به بردار ویژگی تبدیل میشوند و بردار ویژگی تولید شده متناسب را به عنوان ورودی به دستهبندهای تککلاسی میدهند. سپس خروجی دستهبندها (به عنوان معیار شباهت بردار ویژگی به

کلاس رفتار عادی) با هـم ترکیـب مـیشـوند تـا وضـعیت عـادی یـا غیرعادی بودن هر درخواست بر مبنای آن شکل بگیرد.



در مورد مسئله مورد پژوهش ما نیز از روشهای معمول ترکیب به همان صورت استفاده می شود و فرآیند تصمیم گیری گروهی انجام میپذیرد. در واقع هر دسته بند تککلاسی به عنوان عنصری شرکت کننده در تصمیم گیری گروهی نظر خود را پیرامون عادی یا ناهنجار بودن درخواست HTTP ورودی اعلام می کند و در نهایت با یکی از روشهای ترکیب تجمیع نظرات دسته بندها پیرامون آن درخواست صورت می پذیرد و تصمیم نهایی اعلام می شود. علاوه بر روشهای معمول، ما در پژوهش خود از روش های رأی اکثریت و ترکیب با استفاده از عملگر SOW به عنوان استراتژی ترکیب، استفاده می نماییم و تصمیم گیری گروهی را انجام می دهیم.

برای تشخیص حمله، با استفاده از یادگیری ماشین و با رویکرد تشخیص ناهنجاری ابتدا نمونه شبیه سازی شده رفتار عادی برنامه کاربردی تحت وب بر مبنای پروتکل HTTP یادگیری شده و سپس با اعمال درخواستهای HTTP، انحراف از حالت عادی اندازه گیری می شود. در روش پیشنهادی، مرحله آموزش دسته بندها و ترکیب آنها به صورت برون خطی انجام می شود [۶ و ۱۳]. به عبارت دیگر نمونه شبیه سازی شده رفتار عادی درخواستها مبتنی بر پروتکل HTTP توسط دسته بندهای تککلاسی یا ترکیب آنها قبل از شروع به کار برنامه کاربردی تحت وب یادگیری می شود و سپس از آن نمونه های ناهنجاری استفاده می شود. روش پیشنهادی، در مرحله یادگیری رفتار عادی برنامه کاربردی برای ساختن درفتار عادی برنامه کاربردی برای ساختن ندارد و صرفاً از مجموعه داده رفتار عادی برنامه کاربردی برای ساختن مرزهای تصمیم بهره می گیرد.

برای یادگیری رفتار عادی پروتکل ما چهار دستهبند تـککلاسـی PDE ،MOG ،SVDD و SVM را روی بردار خصیصههای اسـتخراج شده از مجموعهداده CSIC2012 بهکار میبریم. سـپس از روشهای ترکیب با استفاده از عملگر S-OWA برای ترکیب نتایج حاصله از آن

دستهبندها میپردازیم. معماری کلی روش پیشنهادی در شکل (۱) نمایش داده شده است.

۵. نتایج و بحث ۵-۱. مجموعهداده

مجموعهداده مورد استفاده در آزمایشهای انجام شده در این مقاله، مجموعهداده CSIC2010 میاشد [۱۶]. این مجموعهداده شامل درخواست.های عادی یا ناهنجار متعلق به تمامی صفحات وب مرتبط با یک برنامه کاربردی تحت وب تجاری است و پارامترهای مرتبط با درخواست.های HTTP آن شامل مقادیر مختلفی است. مجموعهداده CSIC2012 شامل حملات وب نوينى نظير تزريـق injection، سـرريز بافر، تزريق CRLF و XSS مى باشد. در اين مجموعهداده، نمونه درخواست.های HTTP به عنوان عادی یا حمله برچسب خورده و در فایلهای مجزا از یکدیگر جدا شدهاند. پس از شناسـایی ویژگـیهـای مورد نیاز برای بررسی که در جدول (۱) به آنها اشاره شد، عملیات استخراج آنها از مجموعهدادگان CSIC2012 انجام پذیرفت. با توجـه به اینکه درخواستهای HTTP در این مجموعهدادگان به صورت خام بوده و همگی در یک فایل xml ذخیره شده بودند. برنامهای به زبان HTTP تهیه شد که درخواستهای Microsoft Visual Studio 2010 را به قالب استانداردی تبدیل نماید که قابل پردازش توسط نرمافزار متلب و به صورت ماتریس ویژگی نمونه ها باشد. یعنی برای هر درخواست به عنوان یک نمونه، تمامی ویژگیهای آن را به صورت جداگانه بتوان مورد تحلیل قرار داد. خروجی برنامه به صورت یک فایل در قالب داده بود که به ازای هر درخواست مقادیر ویژگیهای مرتبط با آن به صورت جداگانه و مشخص ذکر شده است.

۵–۲. معیارهای ارزیابی

نرخ تشخیص و نرخ هشدار نادرست: از دو معیار نرخ تشخیص (DR) و نرخ هشدار نادرست (FAR) برای ارزیابی کارایی سامانه تشخیص ناهنجاری پیشنهادی برای برنامه های کاربردی تحت وب می توان استفاده کرد. برای این دو معیار داریم:

$$DR = \frac{TP}{TP + FN} \tag{(\%)}$$

$$FAR = \frac{FP}{FP + TN} \tag{(\%)}$$

که در آن، TP تعداد درخواستهای HTTP ناهنجاری هستند که به درستی تشخیص داده شدهاند و FN تعداد درخواستهای ناهنجاری هستند که به عنوان عادی تشخیص داده شدهاند. FP تعداد درخواستهای عادی هستند که به اشتباه ناهنجار تشخیص داده شدهاند و TN تعداد درخواستهایی است که به درستی عادی تشخیص داده شدهاند.

به صورت ایدهآل سامانه تشخیص ناهنجاری می،ایست نرخ تشخیص ۱۰۰٪ و نرخ هشدار نادرست ۰٪ داشته باشد. با این حال در عمل این امر به سختی محقق میشود.

منحنی "ROC ایده اساسی سامانه ای تشخیص ناهنجاری محاسبه احتمال ناهنجار بودن درخواستهای HTTP، بر اساس نتایج آزمون تشخیص ناهنجاری میباشد. تحلیلهای ROC برای مشخص کردن دقت واقعی نتایج تشخیص است. برای بررسی عملکرد سامانه ای تشخیصی منحنی های ROC از اهمیت ویژهای برخواردند [۱۷]. تحلیلهای ROC رویکردی استاندارد است که برای مشخص کردن حساسیت و ویژگی تشخیص ها به کار میروند. برای این منظور، منحنی ROC برای تعریف کردن رابطه حساسیت و ویژگی سامانه تشخیصی به کار میرود.

منحنیها بین صفر و یک قرار میگیرند. منحنیهای که در همسایگی نیمساز ۴۵ درجه هستند معرف سامانههای تشخیصی نامناسب هستند و همچنین نمودارهای که مساحت زیر منحنی ROC مساوی یا کمتر از مساحت بالای منحنی باشد نشان دهنده تستی غیر موفقیتآمیز هستند.

مقادیر AUC: سطح زیر نمودار ROC (AUC^{*}) به عنوان یک معیار معمول و شناخته شده برای مقایسه روش های دسته بندی و داده کاوی به کار میرود. شش الگوریتم مختلف دسته بندی را روی شش مجموعه دادگان پزشکی واقعی مورد آزمایش قرار داده شده و مشخص شده که AUC خواص دقت بهتری نسبت به ROC از خود نشان می دهد و معیار بهتری برای مقایسه الگوریتم های دسته بندی می باشد [1۸].

معیارهایی که برای ارزیابی دستهبندهای تککلاسی به عنوان تشخیص دهنده ناهنجاری در این پژوهش به کار بردهایم، علاوه بر نرخ تشخیص و نرخ هشدار نادرست که در فصل دوم مورد مطالعه قرار گرفت، سطح زیر نمودار AUC نیز بوده است.

۵–۳. ارزیابی نتایج

به منظور مقایسه کارایی روش ترکیب دستهبندها در مقایسه با استفاده از دستهبندها به صورت مستقل، نتایج مقایسهای در جدول (۲) روی بردارهای ویژگی استخراج شده از مجموعهداده CSIC2012 حاصل شده است. همه نتایج با استفاده از پردازنده GHz Intel Core i5 به دست آمده است.

برای هر دستهبند پارامترها طوری تنظیم شدهاند که عملکرد آن دستهبند بهینه شود. در دستهبند MOG پارامتر regularization برای ماتریس کوواریانس ۲۰۱۰ و تعداد تکرار الگوریتم ۲۵ بوده است. برای دستهبند PDE، مقدار تخمین شباهت بیشینه^۲ برای هموارسازی تخمین چگالی Parzen، ۵۰/۰بوده است. پارامتر کرنل گاوسی برای دستهبند SVDD، ۵٬۰۲۵ و پارامتر کرنل RBF برای دستهبند OCSVM، ۲۰/۰ در نظر گرفته شده است.

شکل (۵) نمودار ROC مربوط به دادههای حملههای مختلف موجود در مجموعهدادگان را با استفاده از چهار دستهبند مذکور را نشان میدهد. در نمودارهای ROC، محور عمودی نشاندهنده نرخ تشخیص حمله و محور افقی نشاندهنده نرخ هشدار نادرست می باشد.

مقدار AUC نیز با استفاده از این دستهبندها در شکل (۶) نشان داده شده است.



شکل (۷) نمودار ROC مربـوط بـه دادههـای حملـههـای مختلـف موجود در مجموعهدادگان را بـا اسـتفاده از اسـتراتژیهـای مختلـف ترکیب دستهبندهای مذکور را نشان میدهد.

مقدار AUC نیز با استفاده از روش های مختلف ترکیب دستهبندها در شکل (۸) نشان داده شده است.



شکل ۶. مقادیر AUC دستهبندهای تککلاسی

¹ Receiver Operating Characteristics Curve
² Area Under the Curve

³ Maximum Likelihood Estimation





عنوان روش نرخ تشخیص نرخ هشدار نادرست ۲/۲۸ ۹۸/۹ SVDD ۳/۸ ۹۵/۸ MOG ۴ ۹۸/۶ PDE ۳/۰۱ ۹۷/۷۷ SVM

جدول ۲. نتایج حاصل از روش پیشنهادی

استراتژی میانه

استراتژی رای اکثریت

استراتژی میانگین

استراتژی S-OWA

برای ارزیابی کارایی روش پیشنهادی خود آن را با روش دیواره آتش ارائه شده در [۳] مقایسه مینماییم. این رساله برای ارزیابی کارایی روش خود از معیارهای نرخ تشخیص و نرخ هشدار نادرست استفاده کرده است که در جدول (۳) آمده است.

٩ ٨/٩

99/4

۹٩/۱

٩٩

۱/٨

۲/۱۳

۲/• ۸

٠/٢

جدول ۳. ارزیابی WAF ارائه شده در [۳]

| ۹۵/۷ | نرخ تشخيص |
|------|------------------|
| ۴/۷ | نرخ هشدار نادرست |

با مقایسه نتایج ارزیابی WAF با روشهایی که ما در این پـژوهش از آنها بهره گرفتیم، مشاهده میشود نـرخ تشـخیص آن از بسـیاری از روشهای پیشنهادی به مراتب کمتر است و صرفاً در حد روشهایی نظیر MOG و روش ترکیبی کمینه است. نرخ هشدار نادرست WAF در حد دستهبندهای تک کلاسی است، هرچند از آنها کمتـر است. نکته قابل توجه نرخ هشدار نادرست بسیار پایین تر روشهای ترکیبی در مقایسه با این روش است که تفاوت بسیار مشهود است و کارایی بسیار بالاتر این روشها نسبت به WAF را نشان میدهد.

۶. نتیجه گیری

در این مقاله از ترکیب دست بندهای تک کلاسی رایج به منظور تشخیص درخواستهای HTTP ناهنجار در برنامههای کاربردی تحت وب استفاده شد و پردازش روش پیشنهادی روی درخواستهای مجموعهداده CSIC2012 انجام گرفت. برای ترکیب دسته بندها از استراتژیهای رایج ترکیب دسته بندها برای تصمیم گیری گروهی استفاده شده است؛ همچنین از عملگر OWA-۵، به منظور ترکیب دسته بندهای تک کلاسی استفاده شده است. استفاده از تصمیم گیری گروهی به ویژه با روش OWA-۵، معیارهای کارایی سامانه تشخیص ناهنجاری را به خوبی به بود بخشیده است. فرآیند تشخیص ناهنجاری با تصمیم گیری گروهی بر مبنای عملگر OWA-۵ سبب با توجه به نمودارهای ROC به دست آمده، مشاهده مـیشـود کـه مساحت زیر نمودار ROC به طور قابل ملاحظهای از مسـاحت بـالای

نتایج حاصل از محاسبه AUC نیز مقادیر نزدیک به یک را نشان میدهد و حاکی از این است که دستهبندهای مورد پژوهش ما کارکرد خوب و قابل قبولی دارند. نرخ تشخیص و نرخ هشدار نادرست هر دستهبند یا ترکیب دستهبندها پس از مرحله یادگیری و در مرحله آزمایش، به عنوان مقیاس کارایی ذکر شده است.

منحنى بيشتر است كه اين امر نشاندهنده تستى موفقيت آميز است.

همانطوری که در جدول (۲) مشاهده میشود، در ترکیب دستهبندهای تککلاسی با استفاده از عملگر S-OWA نرخ تشخیص ترکیب دستهبندها نرخ تشخیص بهبود یافته و نرخ هشدار نادرست نیز در مقایسه با بهکارگیری مستقل دستهبندها کاهش مییابد و کارایی روش پیشنهادی و ایده ترکیب دستهبندها در تشخیص ناهنجاری برنامههای کاربردی تحت وب به خوبی اثبات میشود.

- [5] Ingham, K. L. "Anomaly Detection for HTTP Intrusion Detection: Algorithm Comparisons and the Effect of Generalization on Accuracy"; Ph.D. Thesis, University of New Mexico, USA, 2007.
- [6] Kruegel, C.; Vigna, G.; Robertson, W. "A Multi-Model Approach to the Detection of Web-Based Attacks"; Computer Networks 2005, 48, 717-738.
- [7] Khandelwal, S. Shah, P. Bhavsar, M. K.; Gandhi, S. "Frontline Techniques to Prevent Web Application Vulnerability"; Int. J. Adv. Res. Comput. Sci. Elec. Eng. 2013, 2, 208-217.
- [8] Chandola, V.; Banerjee, A.; Kumar, V. "Anomaly detection: A Survey"; ACM Computing Surveys 2009, 41, 3-75.
- [9] Nascimento, G. M. "Anomaly Detection of Web-Based Attacks"; M. S. Thesis, University of Lisbon, Portugal, 2010.
- [10] Berners-Lee, T.; Fielding, R.; Frystyk, H. "Hypertext Transfer Protocol-HTTP/1.0"; 1996.
- [11] Torrano-Gimenez, C.; Nguyen, H. T.; Alvarez, G.; Franke, K. "Combining Expert Knowledge with Automatic Feature Extraction for Reliable Web Attack Detection"; Security Comm. Networks 2012, 119-132.
- [12] Tax, D. M. J. "One-Class Classification"; Ph.D. Thesis, Delft University, Netherland, 2001.
- [13] Kittler, J.; Hatef, M.; Duin, R.; Matas, J. "On Combining Classifiers"; IEEE Transactions on Pattern Analysis and Machine Intelligence 1998, 20, 226-239.
- [14] Reformat, M.; Yager, R. "Building Ensemble Classifiers using Belief Functions and OWA Operators"; Soft Computing 2008, 12, 543-558.
- [15] Filev, D.; Yager, R. "On the Issue of Obtaining OWA Operator Weights"; Fuzzy Sets and Systems 1998, 94, 157-169.
- [16] The HTTP Dataset CSIC2012, http://iec.esic.es/dataset/, Department of Information Processing and Codification (T.I.C.), of the Institute of Applied Physics (I.F.A.), Spanish Scientific Research Council (C.S.I.C.), 2012.
- [17] Bradley, A. "The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms"; Pattern Recognition 1997, 30, 1145-1159.
- [18] Ling, X.; Huang, J.; Zhang, H. "Advances in Artificial Intelligence: AUC: a Better Measure than Accuracy in Comparing Learning Algorithms"; Springer: Berlin-Heidelberg, 2003.
- [19] Tax, D. M. J. "Ddtools 2012, the Data Description Toolbox for MATLAB"; Version 1.9.1, 2012.

بهبود نرخ هشدار نادرست به طور چشمگیری می شود و نرخ تشخیص مناسبی نیز دارد، به طوری که نـرخ تشـخیص بـه ۹۹ درصـد و نـرخ هشدار نادرست نیز به ۰/۲ درصد رسیده است.

در روش ذکر شده با ترکیب دستهبندهای تککلاسی متداول با روش ترکیب با استفاده از عملگر S-OWA، نرخ تشخیص افزایش یافته و نرخ هشدار نادرست نیز به خوبی کاهش یافت.

با توجه به مطالعات انجام شده بر روی روش های مختلف تـأمین امنیت برنامههای کاربردی تحت وب و ابزارهای گوناگون آن، عـدم استفاده این ابزارها و روش ها از سامانه ترکیبی پیشنهادی ما، صـحت قول دیدگاه نوآورانه پژوهش این مقاله را تأیید میکند.

پژوهشهای آینده میتوانند روی میزان تأثیر استفاده از روشهای دیگر دستهبندی تککلاسی به صورت مستقل و روشهای دیگر ترکیب این دستهبندها، در بهبود کارایی سامانههای تشخیص نفوذ مبتنی بر تشخیص ناهنجاری متمرکز شوند. با توجه به گستردگی و نوآوری ویژه این پژوهش بحث بر روی مقایسه زمان اجرای روش پیشنهادی و سایر روشهای ارائه شده همچنان در حال بررسی است. میتوان با مطالعات بیشتر ویژگیهای دیگری برای توصیف درخواستهای HTTP تعریف کرد تا به صورت جامعتری عملکرد درخواستها را توصیف نماید. دستهبندهای مختلف دیگری را میتوان با روشهای متنوع ترکیب نمود و نتایج را مورد ارزیابی قرار داد.

۷. مراجع

- [1] "Internet, the Newest and Most Effective Weapon"; http://paydarymelli.ir/fa/news/2499, 2013.
- [2] "Now Cyber War"; http://paydarymelli.ir/fa/news/970 (In Persian).
- [3] Nguyen, H. T. "Reliable Machine Learning Algorithms for Intrusion Detection Systems"; Ph.D. Thesis, Gjøvik University College, Norway, 2012.
- [4] Kruegel, C.; Vigna, G. "Anomaly Detection of Web-Based Attacks"; In Proc. of the 10th ACM Conf. on Computer and Communications Security 2003, 251-261