مجله علمی پژوهشی «علوم و فناوری‌های پدافند نوین»
سال پنجم، شماره ۲، تابستان ۱۳۹۳؛ ص ۱۱۹-۱۰۷

شناسایی حملات برنامه‌های کاربردی تحت وب با استفاده از ترکیب دسته‌بندهای تک‌کلاسی

حسین شیرازی[1*]، امینه جمالی فرد[2]، سیدمحمدرضا فرشچی[3]

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)

چکیده

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 آسیب‌پذیرترین اهداف دشمن تبدیل شده‌اند. کشف 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 با استفاده از ترکیب دسته‌بندهای تک‌کلاسی پیشنهاد شده است. در مرحله آموزش بردارهای ویژگی استخراج شده مرتبط با هر درخواست HTTP، وارد سامانه شده و نمونه شبیه‌سازی شده درخواست عادی توسط هر دسته‌بند یادگیری می‌شود. سپس با استفاده از روش‌های مختلف ترکیب دسته‌بندهای تک‌کلاسی، بار دیگر نمونه شبیه‌سازی شده درخواست عادی HTTP به سامانه یادگیری منتقل می‌شود. برای ترکیب دسته‌بندها از استراتژی‌های مختلف ترکیب، جهت تصمیم‌گیری گروهی استفاده شده است. نتایج ارزیابی‌های کمی و کیفی روش پیشنهادی بر روی پایگاه‌داده CSIC2012، بیانگر نرخ تشخیص حدود ۹۹ درصد در مدل‌سازی با روش‌های ترکیبی و حداکثر نرخ هشدار نادرست ۰/۲ می‌باشد. رویکرد سامانه پیشنهادی در استفاده از تصمیم‌گیری گروهی، معیارهای کارآیی سامانه تشخیص ناهنجاری را به خوبی بهبود بخشیده است.

کلید واژه‌ها: امنیت سایبری، سامانه‌های تحت وب، دسته‌بندهای تک‌کلاسی، تصمیم‌گیری گروهی، عملگر S-OWA.

Detection of Attacks against Web Applications Using Combination of One-Class Classifiers

H. Shirazi[*], A. Jamalyfard, S. M. R. Farshchi
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 می‌توان جدیدترین سلاح معاصر و به صورت بالقوه تأثیرگذارترین و ویرانگرترین آن‌ها دانست. امروزه کشورهای جهان تلاش می‌کنند تا افسار این شبکه سرکش را جهت رام کردن و تحت نظر گرفتن آن به دست گیرند. شاید یکی از بارزترین و مهم‌ترین تهدیدات اینترنتی صهیونیستی، تولید ویروس استاکس نت[1] به منظور اختلال در مراکز هسته‌ای ایران بود که بر اساس گزارش‌های مطبوعاتی این ویروس‌ها پس از تولید توسط شرکت‌های مهندسی، ماه‌ها در نیروگاه‌های هسته‌ای دیمونا در صحرای نقب[2] مورد آزمایش قرار گرفتند. این سلاح سایبری نمونه‌ای قابل توجه از اقدامات و گام‌های دشمن در تغییر مفاهیم جنگ‌های سنتی قلمداد می‌شود، جنگ‌هایی که از جنگ با توپ و مواد منفجره به جنگ سایبری تبدیل شده‌اند [۱]. زمانی که درباره تهدیدات امنیتی برنامه‌های وب سخن به میان می‌آید، تهاجم علیه سایت‌ها، سرقت اطلاعات کارت‌های اعتباری، حملات منع سرویس به وب سایت‌ها در جهت مستأصل ارائه خدمات و سرویس‌های تعریف شده آنان، ویروس‌ها، تروجان‌ها، کرم‌ها و...، در ذهن تداعی می‌شود. می‌بایست بپذیریم 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 کاهش می‌دهد و امکان بازسازی مناطق آسیب‌دیده را با کمترین هزینه فراهم می‌سازد. در حقیقت طرح‌های پدافند غیرعامل قبل از انجام مراحل تهاجم و در زمان صلح تهیه و اجرا می‌شوند. با توجه به فرصتی که در زمان صلح جهت تهیه چنین طرح‌هایی فراهم می‌شود 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 به‌کار می‌برند و از روش ساده تشخیص مبتنی بر قانون استفاده می‌کنند. این سامانه‌ها به سادگی پیاده‌سازی می‌شوند ولی نیازمند دانش اولیه انواع حملات هستند و نمی‌توانند حملات جدیدی که قبلاً مشاهده نکرده‌اند را شناسایی کنند، می‌توان 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 می‌کنند. نمونه شبیه‌سازی شده از رفتارهای عادی سامانه تحت شرایط کنترل شده خاصی ایجاد می‌شود. ساختن نمونه شبیه‌سازی شده رفتار سامانه در مرحله آموزش و در بخش برون خط تشخیص ناهنجاری است. هنگامی که این نمونه شبیه‌سازی شده ساخته شد، وضعیت شبکه متناوباً با این نمونه شبیه‌سازی شده مقایسه می‌شود تا انحراف‌ها از شرایط طبیعی کشف شوند. این بخش قسمت برخط تشخیص ناهنجاری را تشکیل می‌دهد.

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 به‌کارگیری 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 می‌توان 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 می‌باشد.

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 می‌شود که از ترکیبی از دسته‌بندهای تک‌کلاسه استفاده می‌نماید. نرخ تشخیص[3] و نرخ هشدار نادرست[4] این روش در مقایسه با سایر روش‌های ارائه شده، شناسایی مناسب با خطای اندکی را نشان می‌دهند. ادامه مقاله بدین ترتیب است: در بخش دوم پیشینه تحقیق را مطالعه می‌شود. در بخش سوم به بررسی برخی تعاریف اولیه مورد ارجاع در مقاله، پرداخته می‌شود. در بخش چهارم روش پیشنهادی مقاله در تشخیص ناهنجاری در برنامه‌های کاربردی تحت وب را ارائه داده و در روش تشخیص ناهنجاری پیشنهادی به منظور بهبود کارآیی از سامانه‌های مبتنی بر ترکیب چند دسته‌بند تک‌کلاسی بهره گرفته می‌شود. بدین منظور ابتدا چهار دسته‌بند تک‌کلاسه رایج را به صورت مستقل به‌کار گرفته و سپس روش‌های ترکیبی متداول و ترکیب با استفاده از

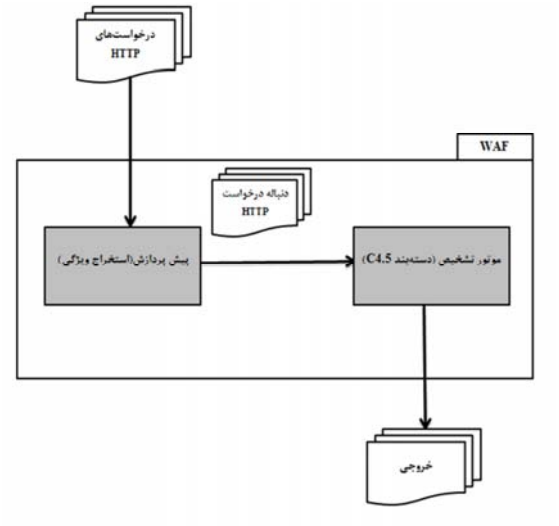
[1] Stuxnet
[2] Negev Dessert
[3] Detection Rate
[4] False Alarm Rate

عملگر S-OWA را برای ترکیب دسته‌بندها به کار برده می‌شود. سپس در بخش پنجم روش پیشنهادی را مورد ارزیابی قرار داده و نتایج حاصل را بیان می‌شود. در بخش ششم به نتیجه‌گیری و بیان پژوهش‌های آینده پرداخته شده است.

۲. پیشینه تحقیق

نخستین سامانه تشخیص نفوذ مبتنی بر تشخیص ناهنجاری، از تخمین پارامتر بیزین درخواست HTTP برای تشخیص ناهنجاری برنامه‌های کاربردی تحت وب استفاده کرد.کروگل و ویگنا، روش‌هایی ارائه دادند که روی تحلیل پارامترهای درخواست‌های HTTP تمرکز دارند و اساساً شامل ترکیبی از نمونه شبیه‌سازی شده‌های تشخیص مختلف می‌باشند. این نمونه‌های شبیه‌سازی شده روی طول ویژگی‌ها، توزیع کاراکتری ویژگی‌ها، استنتاج ساختاری، حضور یا 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 مقیاس‌دهی هر متغیر بر مبنای انحراف معیار استاندارد و کوواریانس، استفاده می‌کند [۴]. وانگ و همکارانش تشخیص دهنده ناهنجاری محتوا بر اساس تحلیل n-gram پیشنهاد دادند که از فیلترهای bloom استفاده می‌کرد و مقاومت در برابر حملات مشابه و چندریخت را فراهم می‌کرد [۵]. اینگهام و همکارانش نشان دادند که چگونه می‌توان سامانه‌ای ارائه داد که با استفاده از الگوریتم استنتاج DFA به همراه ابتکاری‌های[1] کاهش اتوماتا، خطر مثبت نادرست را کمینه کند. روش آنها الگوریتم آموزش دارای قابلیت کار با داده‌های غیرایستا با طول دلخواه را فراهم می‌کند [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 کردن درخواست‌های HTTP و استخراج الگوریتم آموزشی مرتبط به کار برده‌اند [۷].

روشی برای امنیت برنامه‌های کاربردی تحت وب پیشنهاد شده که دیواره آتش برنامه کاربردی تحت وب[2] نام گرفته است. این روش پس از پیش پردازش داده‌ها، 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 DARPA بیان شده است. WAF پیشنهادی برای دسته‌بندی درخواست‌های HTTP با نمونه‌های عادی و حمله آموزش داده می‌شود. در مرحله آموزش، مجموعه‌دادگان اولیه برای آموزش دسته‌بندی درخت تصمیم به سامانه ارائه می‌شود. ساختار روش پیشنهادی در شکل (۱) آمده است. نخست، پیش‌پردازش برای استخراج ویژگی‌های و برچسب هر بسته HTTP انجام می‌شود. سپس، یک سوم مجموعه‌دادگان برای آموزش تشخیص دهنده با استفاده از الگوریتم C4.5 در نرم‌افزار WEKA؛ به‌کار گرفته می‌شود. نرم‌افزار درخت تصمیمی که موتور تشخیص WAF را بازنمایی می‌کند، به عنوان خروجی، بازمی‌گرداند [۳].



شکل ۱. WAF ارائه شده در مرجع [۳]

۳. تعاریف اولیه

آسیب‌پذیری‌های تحت وب به عنوان بخش عمده‌ای در زمینه امنیت سامانه‌های رایانه‌ای مطرح می‌باشند. به منظور شناسایی حملات شناخته شده تحت وب، سامانه‌های تشخیص نفوذ به تعداد زیادی از امضای حملات مجهز می‌شوند. متأسفانه، همگام‌سازی با استفاده از به‌روزرسانی تغییرات رخ داده در حملات اینترنتی امر بسیار سختی است. همچنین ممکن است آسیب‌پذیری با نصب برنامه‌های کاربردی تحت وب مشخصی رخ دهد. به همین دلیل سامانه‌های تشخیص نفوذ بهتر است با رویکرد تشخیص ناهنجاری پیاده‌سازی می‌شوند.

تشخیص ناهنجاری در حیطه مسائلی است که تلاش می‌شود در میان داده‌ها الگوهایی که با رفتار از پیش مورد انتظار مغایرت دارند، کشف شود. این الگوهای ناهمگون بیشتر مربوط به داده‌های پرت، مشاهدات ناسازگار، موارد استثنا، موارد انحرافی، ویژگی‌های گمراه کننده و فعالیت‌های مختل کننده در زمینه‌های کاربردی گوناگون می‌باشد. واژه‌های داده پرت[3] و داده ناهنجار[4] به صورت متناوب در متون تخصصی این حیطه به جای یکدیگر به کار می‌روند که معادل هم می‌باشند [۸].

برنامه کاربردی تحت وب هر نوع برنامه کاربردی است که از مرورگر وب به عنوان سرویس‌گیرنده استفاده می‌کند. تمامی پایگاه‌های موجود بر روی اینترنت از پروتکل HTTP استفاده 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 این پروتکل HTTP است که به عنوان زبان مشترک ارتباطی بین سرویس گیرنده و سرویس دهنده وب به رسمیت شناخته شده و از آن استفاده می‌شود. در واقع مرورگر وب صدای خود را با استفاده از پروتکل HTTP به گوش سرویس دهنده وب می‌رساند و تقاضای سرویس می‌کند [۹].

درخواست HTTP مجموعه‌ای از خطوط متنی است (با CRLF از یکدیگر جدا شده‌اند) که به سرویس دهنده وب ارسال می‌شود و شامل خط درخواست[1]، قسمت‌های سرپیام درخواست[2] و بدنه درخواست[3] می‌باشد. خط درخواست از سه بخش تشکیل شده که با فاصله از یکدیگر جدا شده‌اند. این سه بخش نام روشی که می‌بایست اعمال شود، مسیر محلی منبع درخواست و نسخه پروتکلی که مورد استفاده قرار می‌گیرد را مشخص می‌کند. نخستین کلمه‌ای که در درخواست HTTP ظاهر می‌شود کلمه method است. بیشتر درخواست‌های HTTP از نوع روش GET هستند ولی انواع روش‌های دیگری همانند POST و HEAD نیز وجود دارند. بعد از method، مسیر منبع (URI) ذکر می‌شود که عموماً یک فایل، یک فهرست در سامانه فایل یا ترکیبی از هر دو است. آخرین بخش، نسخه پروتکل استفاده شده توسط سرویس گیرنده را مشخص می‌کند (عموماً HTTP/1.0 یا HTTP/1. 1).

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"[4] که با CRLF خاتمه می‌یابد. در HTTP 1.0 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 هر دو ارائه می‌دهند (جستجوگر، سیستم عامل و غیره). هر یک از این خطوط از نامی تشکیل شده که نوع سرپیام را مشخص می‌کند و با (:) و مقدار سرپیام دنبال می‌شود [۱۰].

۴. الگوی شبیه‌سازی شده پیشنهادی

۴-۱. ساختار

در حوزه تشخیص ناهنجاری در برنامه‌های کاربردی تحت وب که مورد بحث ما در این پژوهش می‌باشد، تنها مجموعه‌دادگان درخواست‌های HTTP عادی در دسترس است و در مرحله آموزش ما می‌خواهیم رفتار عادی درخواست‌های HTTP را شبیه‌سازی 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 می‌شود. در واقع هر کدام از دسته‌بندهای تک‌کلاسی 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.

۴-۲. استخراج ویژگی

یکی از مهم‌ترین ملزومات برای ارائه یک سامانه تشخیص ناهنجاری برنامه‌های کاربردی تحت وب بر مبنای پروتکل HTTP، شناخت رفتار عادی این پروتکل است تا بتوان ناهنجاری‌ها و حمله‌ها را که به عنوان انحراف از حالت عادی تعریف می‌شود، 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 می‌گیرد و به تبع آن رفتار عادی به عنوان قیدی روی مقدار ویژگی‌های تعریف شده یا رابطه‌ای مابین ویژگی‌ها تعریف می‌شود. برای تعریف ویژگی‌هایی از درخواست‌های HTTP که در مسئله تشخیص ناهنجاری تعیین‌کننده باشند، نیازمند شناخت حملات و نحوه تأثیر 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)).

جدول ۱. ویژگی‌های درخواست HTTP [۳].

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

هر کدام از این ویژگی‌ها به نوعی در حملات متداول تحت وب، تحت تأثیر فعالیت‌های مخرب مهاجمان قرار گرفته‌اند. بعضی 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. دسته‌بندهای تک‌کلاسی

در فرآیند آموزش در روش‌های دسته‌بندی دو یا چند کلاسی، داده‌های مربوط به همه کلاس‌ها موجود است. در صورتی که در تشخیص ناهنجاری تک‌کلاسی، 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 [۱۲]. در این گونه مسائل مجبور به استفاده از دسته‌بندی‌های تک‌کلاسی بوده تا بتوان مشخصات یک کلاس موجود را یادگیری نماییم.

در روش‌های دسته‌بندی تک‌کلاسی دو مؤلفه اصلی باید مشخص شود. مؤلفه اول عبارت است از اندازه‌گیری مقدار فاصله (d(x)) یا شباهت (p(x)) یک شیء x در فضای ویژگی 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 (d(x)<θ) یا شباهت آن به کلاس عادی بزرگ‌تر از حد آستانه باشد (p(x)>θ). دسته‌بندی‌های تک‌کلاسی را با توجه به روشی که در حل مسئله دسته‌بندی تک‌کلاسی به‌کار می‌گیرند و نمونه شبیه‌سازی شده که از کلاس هدف ارائه می‌کنند در سه گروه قرار می‌دهد، روش‌های مبتنی بر مرز (مانند SVM و SVDD)، روش‌های مبتنی بر چگالی (PDE و MOG) و روش‌های مبتنی بر دوباره‌سازی[1] (مانند SOM، K-means و PCA).

در این مقاله برای یادگیری رفتار عادی درخواست HTTP از دسته‌بندی‌های تک‌کلاسی زیر استفاده می‌شود:

۱- دسته‌بندی تک‌کلاسی SVDD : یک ابرکره را بر داده‌های کلاس موجود (به عنوان کلاس هدف) محاط می‌کند. محدوده ابرکره توسط اشیائی از کلاس هدف تعیین می‌شود. این اشیاء بردارهای پشتیبان نامیده می‌شوند. در SVDD فاصله شیء x از کلاس هدف طبق رابطه زیر محاسبه می‌شود:

$$D_{SVDD}(x) = k(x,x) - 2\sum_i \alpha_i * k(x,x_i) + \sum_{i,j} \alpha_i\alpha_j * k(x_i,x_j)\, D_{SVDD}(x) \quad (۱)$$

که در آن، k نشان دهنده تابع هسته، x_i و α_i ضریب لاگرانژ منتسب به بردار پشتیبان x_i است.

۲- اختلاط نمونه‌های شبیه‌سازی شده‌های گوسی (MOG)[2]: اختلاط نمونه شبیه‌سازی شده‌های گوسی یک ترکیب خطی از توزیع نرمال است که تابع چگالی آن طبق رابطه زیر به دست می‌آید:

$$P_{MOG}(x) = \frac{1}{N_{MOG}} \sum_i \alpha_i P_N(x;\mu_i,\sum_i) \quad (۲)$$

که در آن، α_i 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 شبیه‌سازی شده‌های گوسی، N_{MOG}، مشخص باشد، میانگین و کوواریانس هر کدام از نمونه شبیه‌سازی شده‌های گوسی با روش بیشینه‌سازی امید ریاضی[3] تعیین می‌شود.

۳- تصمیم چگالی پارزن[4] (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:

$$P_{PDE}(x) = \frac{1}{N}\sum_i K_h(x - x_i) \quad (۳)$$

که تابع هسته مورد استفاده (اغلب گوسی)، N تعداد اشیاء موجود در مجموعه‌داده آموزش، x_i، i امین شیء موجود در مجموعه‌داده آموزش و h عرض هسته است که با آموزش و با استفاده از روش بیشینه مقدار احتمال[5] تعیین می‌شود [۱۲].

۴- ماشین بردار پشتیبان (SVM): این روش ابرصفحه‌هایی با حداکثر حاشیه را به دست می‌آورد که دسته‌های داده‌ها 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:

$$w_1X_1 + w_2X_2 + b = 0 \quad (۴)$$

در حالت n بعدی خواهیم داشت:

$$\sum_{i=0}^{n} w_i x_i + b = 0 \quad (۵)$$

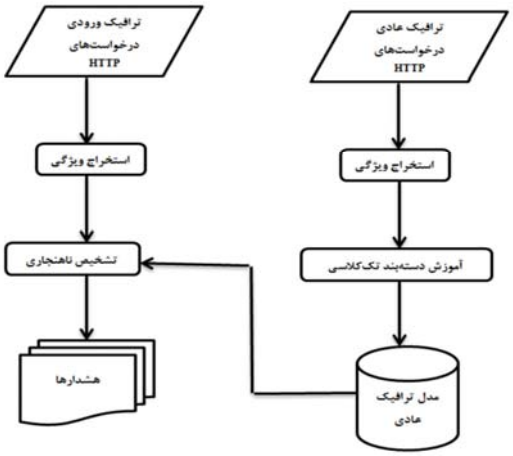
نمونه شبیه‌سازی شده سامانه تشخیص ناهنجاری پیشنهادی برای هر درخواست HTTP در شکل (۲) نشان داده شده است.

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

[1] Reconstruction

[2] Mixture of Gaussian Model
[3] Expectation-Maximization
[4] Parzen Density Estimator
[5] Maximum Likelihood

می‌شود. در روش پیشنهادی، مرحله آموزش دسته‌بندها به صورت برون‌خطی انجام می‌شود. به عبارت دیگر نمونه شبیه‌سازی شده رفتار عادی درخواست‌ها مبتنی بر پروتکل HTTP، توسط دسته‌بندهای تک‌کلاسی قبل از شروع به استفاده از برنامه کاربردی تحت وب یادگیری می‌شود و سپس از نمونه شبیه‌سازی شده‌های یادگیری شده در هنگام کارکرد برنامه کاربردی برای تشخیص ناهنجاری استفاده می‌شود. روش پیشنهادی، در مرحله یادگیری رفتار عادی برنامه کاربردی تحت وب، نیازی به مجموعه-داده حمله ندارد و صرفاً از مجموعه‌دادگان رفتار عادی برنامه کاربردی برای ساختن مرزهای تصمیم بهره می‌گیرد.



شکل ۲. آموزش و تشخیص سامانه تشخیص ناهنجاری.

ما در روش پیشنهادی خود، مسئله تشخیص ناهنجاری را به صورت مسئله تصمیم‌گیری گروهی دنبال می‌کنیم و روش‌های متداول ترکیب و نوعی روش میانگین مرتب شده وزن‌دار[1] (OWA) موسوم به S-OWA، را در آن‌ها به‌کار می‌گیریم. در حالت کلی، فرآیند تصمیم‌گیری گروهی عبارت است از حالتی که دو یا چند متخصص، هر کدام با عقاید و ویژگی‌های منحصر بفرد خود سعی می‌کنند تا یک تصمیم مشترک بگیرند. مهم‌ترین مسئله‌ای که در تصمیم‌گیری گروهی مطرح می‌شود این است که چگونه نظرات متخصصین با هم ترکیب شود طوری که تصمیم گرفته شده در جهت ارضای معیار مشخصی باشد [۱۱].

دسته‌بندهای تک‌کلاسی به سختی می‌توانند تمامی مشخصات[2] 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 حمله‌های موجود انتخاب شود و در آینده حمله‌ای 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 می‌شود.

روش‌های مختلفی مانند میانگین‌گیری، 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

مسئله بازشناسی الگویی را در نظر می‌گیریم که الگوی Z به یکی از m کلاس ممکن $(w_1,\ldots,w_m)$ تعلق گیرد. فرض می‌شود R دسته‌بند داریم که هر یک الگوی Z را با بردار اندازه‌گیری[3] مشخصی بازنمایی می‌کنند. بردار اندازه‌گیری مرتبط با دسته‌بند i ام با x_i نشان داده می‌شود. در فضای اندازه‌گیری هر کلاس w_k با تابع چگالی احتمال شبیه‌سازی می‌شود و احتمال اولیه رخداد[4] آن با $P(w_k)$ نمایش داده می‌شود. فرض می‌شود نمونه‌های شبیه‌سازی شده متقابلاً منحصر به‌فرد[5] هستند که این بدان معنی است که در نهایت تنها یک نمونه شبیه‌سازی شده با هر الگو مرتبط است.

حال، با توجه به تئوری بیز، اندازه‌گیری‌های داده شده x_i، $i=1\ldots R$، الگوی Z می‌بایست به کلاس w_i تعلق بگیرد که احتمال ثانویه آن تعبیر به صورت زیر بیشینه شود:

Assign $z \rightarrow w_j$ if

$$P\left(w_j \mid x_1,\ldots,x_R\right) = argmax_k P(w_k \mid x_1,\ldots,x_R) \qquad (۶)$$

قانون تصمیم بیز فوق بیان می‌دارد برای استفاده صحیح از تمامی اطلاعات در دسترس به منظور تصمیم‌گیری، ضروری است احتمالات فرضیات متعددی، با در نظر گرفتن تمامی اندازه‌گیری‌ها در یک زمان، محاسبه شود. این مطلب، عبارتی صحیح در مسئله کلاس‌بندی است ولی ممکن است راهکاری عملی نباشد. محاسبه توابع احتمال ثانویه به دانشی از آمار اندازه‌گیری در سطح بالاتری بستگی دارد که به صورت توابع چگالی احتمال پیوسته[6] $P\left(x_1 \ldots x_R \mid w_k\right)$ هستند و به سختی قابل استنتاج می‌باشند. به همین دلیل می‌بایست قانون فوق را ساده‌سازی کرد و آن را به صورت محاسبات تصمیم‌گیری انجام شده توسط دسته‌بندهای تکی، که تنها شامل اطلاعات موجود در بردار x_i بیان کرد. مشاهده می‌شود 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 می‌کند، بلکه ترکیب دسته‌بندها که در عمل از آن‌ها استفاده می‌شود را نیز ممکن می‌سازد. به‌علاوه این رویکرد حیطه‌ای را برای گسترش استراتژی‌های ترکیب دسته‌بندها، فراهم می‌آورد.

می‌توان از قانون تصمیم بیز شروع کرد و آن را با در نظر گرفتن فرضیات مشخصی، روشن‌تر کرد. ابتدا احتمال ثانویه $P(w_k|\, x_1, \dots, x_R)$ را با استفاده از تئوری بیز بازنویسی می‌کنیم. خواهیم داشت:

$$P(w_k|\, x_1, \dots, x_R) = \frac{p\,(x_1 \dots x_R|w_k)P(w_k)}{p(x_1, \dots, x_R)} \qquad (۷)$$

که در آن، $p(x_1, \dots, x_R)$ اندازه‌گیری غیر شرطی چگالی احتمال پیوسته است. این احتمال را می‌توان به صورت توزیع‌های اندازه‌گیری شرطی به صورت زیر نوشت:

$$p(x_1, \dots, x_R) = \sum_{j=1}^{m} P\,(x_1, \dots, x_R|w_j)P(w_j) \qquad (۸)$$

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

قانون حاصل ضرب: همان طور که ملاحظه شد، $p\,(x_1 \dots x_R|w_k)$ توزیع احتمال پیوسته اندازه‌گیری‌های استخراج شده توسط دسته‌بندها را بازنمایی می‌کند. فرض می‌شود بازنمایی‌های استفاده شده همگی از لحاظ آماری استقلال شرطی[1] دارند. در حالات خاص، استفاده از بازنمایی‌های مختلف را می‌توان دلیل چنین استقلالی دانست. با استفاده از چنین فرضی می‌توان نوشت:

$$p\,(x_1, \dots, x_R|w_k) = \prod_{i=1}^{R} p(x_i|\, w_k) \qquad (۹)$$

که در آن، $p(x_i|\, w_k)$، نمایه فرآیند اندازه‌گیری i امین بازنمایی است. با جایگذاری (۸ و ۹) در (۷) خواهیم داشت:

$$P(w_k|\, x_1, \dots, x_R) = \frac{P(w_k)\prod_{i=1}^{R} p(x_i|\, w_k)}{\sum_{j=1}^{m} P(w_j)\ \prod_{i=1}^{R} p(x_i\,|\, w_j)} \qquad (۱۰)$$

$$assign\ Z \rightarrow\ w_j\ \ if$$

$$P^{-(R-1)}(w_j)\prod_{i=1}^{R}\left(P(w_j|\, x_i)\right) = \max_{k=1,..,m} P^{-(R-1)}(w_k)\prod_{i=1}^{R} P(w_k|x_i) \qquad (۱۱)$$

با بیان احتمالاتی ثانویه حاصل شده از دسته‌بندهای مورد بحث خواهیم داشت:

$$assign\ Z \rightarrow\ w_j\ \ if$$

$$P^{-(R-1)}(w_j)\prod_{i=1}^{R} P(w_j|\, x_i) = \max_{k=1,..,m} P^{-(R-1)}(w_k)\ \prod_{i=1}^{R} P(w_k|\, x_i) \qquad (۱۲)$$

قانون تصمیم‌گیری (۱۲) درستنمایی[2] فرضیه را با ترکیب احتمالات ثانویه تولید شده توسط دسته‌بندهای تکی را با استفاده از قانون حاصل ضرب بیان می‌کند. این قانون برای ترکیب خروجی دسته‌بندها بسیار کارآمد است. به این صورت که در موتور تشخیص ترکیبی با نزدیک به صفر کردن خروجی احتمال یک بازنمایی، از آن جلوگیری می‌کند. همان‌طور که در ادامه این بخش خواهیم دید، این امر در ترکیب قانون‌های تصمیم‌گیری نامطلوب است. زیرا تمامی دسته‌بندها می‌بایست برای هر شناسه کلاس مفروض گزینه رد یا قبول را تولید کنند.

قانون حاصل جمع[3]: 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 نمی‌آید. 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:

$$P(w_k|x_i) = P(w_k(\,1 + \delta_{ki})) \qquad (۱۲)$$

که در آن، $\delta_{ki} \ll 1$ است. با جایگذاری (۱۲) در (۱۱) به عنوان احتمال ثانویه خواهیم داشت:

$$P^{-(R-1)}(w_k)\prod_{i=1}^{R} P(w_k|x_i) = P(w_k)\prod_{i=1}^{R}(\,1 + \delta_{ki}) \qquad (۱۳)$$

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:

$$P(w_k)\prod_{i=1}^{R}(\,1 + \delta_{ki}) = P(w_k) + P(w_k)\sum_{i=1}^{R}\delta_{ki} \qquad (۱۴)$$

با جایگذاری (۱۲ و ۱۴) در (۱۱) خواهیم داشت:

$$assign\ Z \rightarrow\ w_j\ \ if$$

$$(1-R)P(w_j) + \sum_{i=1}^{R} P(w_j|\, x_i) = \max_{k=1,..,m}\left[(1-R)P(w_k) + \sum_{i=1}^{R} P(w_k|x_i)\right] \qquad (۱۵)$$

قوانین تصمیم‌گیری (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:

$$\prod_{i=1}^{R} P(w_k|x_i) \le \min_{i=1,..,R} P(w_k|x_i) \le \frac{1}{R}\sum_{i=1}^{R} P(w_k|x_i) \le \max_{i=1,..,R} P(w_k|x_i) \qquad (۱۶)$$

رابطه (۱۶) پیشنهاد می‌کند که قوانین ترکیب حاصل جمع و حاصل ضرب را به وسیله باندهای پایینی و بالایی تقریب زده شود. همچنین می‌توان با سخت‌سازی[4] 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

$$\Delta_{ki}=\begin{cases}1, \text{if } P(w_k|x_i)=\max\limits_{j=1,..,m} P(w_j|x_i)\\ 0, otherwise\end{cases} \quad (۱۷)$$

از این نتیجه می‌توان در خروجی ترکیب دسته‌بندها به جای ترکیب احتمالات ثانویه استفاده کرد. با این تقریب بقیه قوانین ترکیب را می‌توان به دست آورد.

قانون میانه[1]: 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:

$$assign\ Z \rightarrow w_j \ \ if$$

$$\frac{1}{R}\sum_{i=1}^{R} P(w_j|\,x_i)=\max_{k=1,..,m}\frac{1}{R}\sum_{i=1}^{R} P(w_k|\,x_i) \quad (۱۸)$$

بنابراین، قانون الگو را به کلاسی متعلق می‌داند که میانگین احتمال ثانویه آن بیشینه شود. چنانچه خروجی یکی از دسته‌بندها احتمال ثانویه‌ای باشد که متعلق به کلاس ناهنجار است، روی میانگین تأثیر می‌گذارد و ممکن است سبب تصمیم‌گیری 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:

$$assign\ Z \rightarrow w_j \ \ if$$

$$\underset{i=1,..,R}{\text{med}}\ P(w_j|\,x_i)=\max_{k=1,..,m}\ \underset{i=1,..,R}{\text{med}}\ P(w_k|\,x_i) \quad (۱۹)$$

قانون رأی اکثریت[2]: با شروع از ۱۵ با فرض مساوی بودن احتمال اولیه و با سخت‌سازی احتمالات با توجه به (۱۷) خواهیم داشت:

$$assign\ Z \rightarrow w_j \ \ if$$

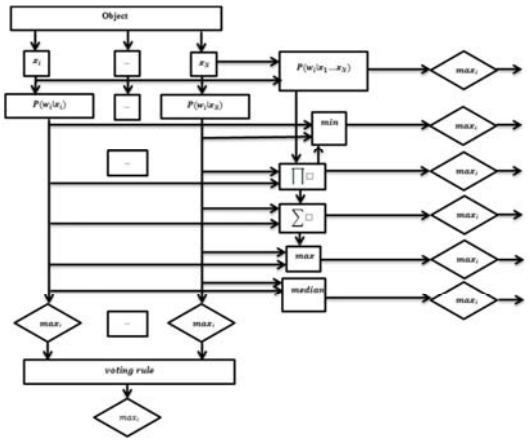
$$\sum_{i=1}^{R}\Delta_{ji}=\max_{k=1,..,m}\sum_{i=1}^{R}\Delta_{ki} \quad (۲۰)$$

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 رأی اکثریت انتخاب می‌شود [۱۳].

تمامی استراتژی‌های فوق و روابط آن‌ها در شکل (۳) آمده است.

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 به‌کار بسته‌ایم.



شکل ۳. استراتژی‌های متداول ترکیب [۱۳]

درجه orness را به صورت زیر تعریف می‌کنیم [۱۴]:

$$\frac{1}{2}\sum_{i=1}^{n} w_i=\frac{1}{2} \quad (۲۱)$$

برای به دست آوردن تعریف جدید، این فرمول را تغییر می‌دهیم و از تساوی زیر استفاده می‌کنیم:

$$\frac{1}{2}\sum_{i=1}^{n} w_i=\frac{1}{2} \quad (۲۲)$$

فرمول درجه orness را به صورت زیر بازنویسی می‌کنیم:

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

$$=\frac{1}{2}+\sum_{i=1}^{n}\frac{(n-2i+1)}{2(n-1)}\,w_i \quad (۲۳)$$

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

حال در نظر می‌گیریم وضعیتی را که n زوج باشد، n = 2m ؛ همچنین i = k ، k <= m و i = n+1-k . خواهیم داشت:

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

$$q_{n+1-k}=\frac{n-2n-2+2k+1}{n-1} \quad (۲۴)$$

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

همچنین

$$orness(W)=\frac{1}{2}\sum_{i=1}^{n} q_k\,(w_k-w_{n+1-k}) \quad (۲۵)$$

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

$$orness(W)=\frac{1}{2}\sum_{i=1}^{n} q_k\,(w_k-w_{n+1-k}) \quad (۲۶)$$

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

[1] Median Rule
[2] Majority Vote Rule

$$q_{m+1} = \frac{2m + 1 - 2(m-1) + 1}{2(n-1)} \quad (۲۷)$$

پس برای orness خواهیم داشت:

$$orness(W) = \frac{1}{2}\sum_{i=1}^{n} q_k\,(w_k - w_{n+1-k}) \quad (۲۸)$$

با استفاده از این عبارت مستقیماً روشی را برای ساختن عملگرهای S-OWA با وزن‌هایی با درجه orness از پیش تعیین شده بیان می‌کنیم.

فرض می‌کنیم درجه orness با نام Ω از پیش داده شده باشد. می‌توانیم فرض کنیم که تمامی وزن‌های به جز w_1 و w_n مساوی باشند. با این فرض تابع orness به سادگی به صورت زیر درمی‌آید:

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

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

$$w_1 - w_n = 2(\Omega - 0.05) \quad (۳۰)$$

می توان w_1 و w_n را هر عددی در بازه بین صفر و یک انتخاب کرد به طوری که شرط فوق را برآورده نمایند. سپس مجموع وزن‌های باقی‌مانده می‌بایست بین صفر و یک باقی بماند. بنابراین داریم:

$$\mathrm{wi} = \frac{1}{n}\,[\,1 - (\mathrm{w1} - \mathrm{wn})\,],\ \mathrm{i} = 2,3,\ldots,\mathrm{n}-1 \quad (۳۱)$$

سپس فرآیند فوق را با الگوریتم زیر تغییر اندکی می‌دهیم.

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

با چنین فرآیند وزن‌های S-OWA تولید می‌شود. در واقع ما پس از تولید وزن‌ها به خروجی هر دسته‌بند، یکی از وزن‌های تولید شده توسط عملگر S-OWA را اختصاص می‌دهیم. برای ما در این پژوهش چنانچه $\Delta > 0$ باشد، $w_1 = L + \Delta$ و $w_i = L$ برای $i = 1,..n$ خواهد بود. در این حالت برای مقدار ترکیب یافته که در واقع خروجی حاصل از ترکیب خروجی دسته‌بندهاست خواهیم داشت:

$$F(a_1, \ldots a_n) = \Delta\, Max_i[\alpha_i] + L\sum_{i=1}^{n}\alpha_i \quad (۳۲)$$

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

بنابراین خواهیم داشت:

$$F(a_1, \ldots a_n) = \Delta\, Max_i[\alpha_i] + (1-\Delta)\, Ave(a_1, \ldots, a_n) \quad (۳۳)$$

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

$$F(a_1, \ldots a_n) = \Delta\, Min_i[\alpha_i] + (1-\Delta)\, Ave(a_1, \ldots, a_n) \quad (۳۴)$$

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 از رفتار برنامه‌های کاربردی تحت وب از 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 HTTP، از ترکیب نظرات این دسته‌بندها استفاده می‌کنیم و با این سامانه ترکیبی 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 اندازه‌گیری می‌شود. در اینجا نیز، مرحله آموزش دسته‌بند به صورت برون‌خطی انجام می‌شود. به عبارت دیگر نمونه شبیه‌سازی شده رفتار عادی درخواست‌ها مبتنی بر پروتکل HTTP، توسط دسته بند تک‌کلاسی ترکیبی قبل از شروع به استفاده از برنامه کاربردی تحت وب یادگیری می‌شود و سپس از نمونه یادگیری شده در هنگام کارکرد برنامه کاربردی برای تشخیص ناهنجاری استفاده می‌شود. 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 ساختن مرزهای تصمیم بهره‌گیری می‌شود.

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 هم ترکیب می‌شوند تا وضعیت عادی یا غیرعادی بودن هر درخواست بر مبنای آن شکل بگیرد.

شکل۴. نمودار سامانه ترکیبی تشخیص ناهنجاری.

در مورد مسئله مورد پژوهش ما نیز از روش‌های معمول ترکیب به همان صورت استفاده می‌شود و فرآیند تصمیم‌گیری گروهی انجام می‌پذیرد. در واقع هر دسته‌بند تک‌کلاسی به عنوان عنصری شرکت‌کننده در تصمیم‌گیری گروهی نظر خود را پیرامون عادی یا ناهنجار بودن درخواست HTTP ورودی اعلام می‌کند و در نهایت با یکی از روش‌های ترکیب تجمیع نظرات دسته‌بندها پیرامون آن درخواست صورت می‌پذیرد و تصمیم نهایی اعلام می‌شود. علاوه بر روش‌های معمول، ما در پژوهش خود از روش‌های رأی اکثریت و ترکیب با استفاده از عملگر S-OWA به عنوان استراتژی ترکیب، استفاده می‌نماییم و تصمیم‌گیری گروهی را انجام می‌دهیم.

برای تشخیص حمله، با استفاده از یادگیری ماشین و با رویکرد تشخیص ناهنجاری ابتدا نمونه شبیه‌سازی شده رفتار عادی برنامه کاربردی تحت وب بر مبنای پروتکل HTTP یادگیری شده و سپس با اعمال درخواست‌های HTTP، انحراف از حالت عادی اندازه‌گیری می‌شود. در روش پیشنهادی، مرحله آموزش دسته‌بندها و ترکیب آنها به صورت برون‌خطی انجام می‌شود [۶ و ۱۳]. به عبارت دیگر نمونه شبیه‌سازی شده رفتار عادی درخواست‌ها مبتنی بر پروتکل HTTP، توسط دسته‌بندهای تک‌کلاسی یا ترکیب آنها قبل از شروع به کار برنامه کاربردی تحت وب یادگیری می‌شود و سپس از آن نمونه‌های یادگیری شده در هنگام کارکرد برنامه کاربردی برای تشخیص ناهنجاری استفاده می‌شود. روش پیشنهادی، در مرحله یادگیری رفتار عادی برنامه کاربردی تحت وب، نیازی به مجموعه‌داده حمله ندارد و صرفاً از مجموعه‌داده رفتار عادی برنامه کاربردی برای ساختن مرزهای تصمیم بهره می‌گیرد.

برای یادگیری رفتار عادی پروتکل ما چهار دسته‌بند تک‌کلاسی SVDD، MOG، PDE و SVM را روی بردار خصیصه‌های استخراج شده از مجموعه‌داده CSIC2012 به‌کار می‌بریم. سپس 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 می‌باشد [۱۶]. این مجموعه‌داده شامل درخواست‌های عادی یا ناهنجار متعلق به تمامی صفحات وب مرتبط با یک برنامه کاربردی تحت وب تجاری است و پارامترهای مرتبط با درخواست‌های HTTP آن شامل مقادیر مختلفی است. مجموعه‌داده CSIC2012 شامل حملات وب نوینی نظیر تزریق injection، سرریز بافر، تزریق CRLF و XSS می‌باشد. در این مجموعه‌داده، نمونه درخواست‌های HTTP به عنوان عادی یا حمله برچسب خورده و در فایل‌های مجزا از یکدیگر جدا شده‌اند. پس از شناسایی ویژگی‌های مورد نیاز برای بررسی که در جدول (۱) به آنها اشاره شد، عملیات استخراج آن‌ها از مجموعه‌دادگان CSIC2012 انجام پذیرفت. با توجه به اینکه درخواست‌های HTTP در این مجموعه‌دادگان به صورت خام بوده و همگی در یک فایل xml ذخیره شده بودند. برنامه‌ای به زبان Microsoft Visual Studio 2010 تهیه شد که درخواست‌های HTTP را به قالب استانداردی تبدیل نماید که قابل پردازش توسط نرم‌افزار متلب و به صورت ماتریس ویژگی نمونه‌ها 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 برنامه‌های کاربردی تحت وب می‌توان استفاده کرد. برای این دو معیار داریم:

$$DR = \frac{TP}{TP + FN} \quad (۳۵)$$

$$FAR = \frac{FP}{FP + TN} \quad (۳۶)$$

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 می‌بایست نرخ تشخیص ۱۰۰٪ و نرخ هشدار نادرست ۰٪ داشته باشد. با این حال در عمل این امر به سختی محقق می‌شود.

منحنی ROC[1]: ایده اساسی سامانه‌های 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 سامانه‌های تشخیصی منحنی‌های ROC از اهمیت ویژه‌ای برخوردند [۱۷]. تحلیل‌های ROC رویکردی استاندارد است که برای مشخص‌کردن حساسیت و ویژگی تشخیص‌ها به‌کار می‌روند. برای این منظور، منحنی ROC برای تعریف کردن رابطه حساسیت و ویژگی سامانه تشخیصی به‌کار 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[2]) به عنوان یک معیار معمول و شناخته شده برای مقایسه روش‌های دسته‌بندی و داده‌کاوی به‌کار می‌رود. شش الگوریتم مختلف دسته‌بندی را روی شش مجموعه‌دادگان پزشکی واقعی مورد آزمایش قرار داده شده و مشخص شده که AUC خواص دقت بهتری نسبت به ROC از خود نشان می‌دهد و معیار بهتری برای مقایسه الگوریتم‌های دسته‌بندی می‌باشد [۱۸].

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

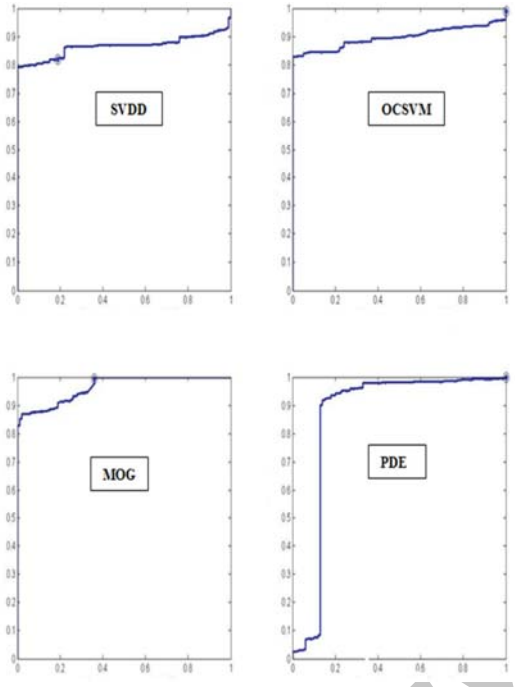
۵-۳. ارزیابی نتایج

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 Intel Core i5، 2.53 GHz 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 ۰/۰۱ و تعداد تکرار الگوریتم ۲۵ بوده است. برای دسته‌بند PDE، مقدار تخمین شباهت بیشینه[3] برای هموارسازی تخمین چگالی Parzen، ۰/۰۵بوده است. پارامتر کرنل گاوسی برای دسته‌بند SVDD، ۰/۰۰۵ و پارامتر کرنل RBF برای دسته‌بند OCSVM، ۰/۰۲ 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 می‌باشد.

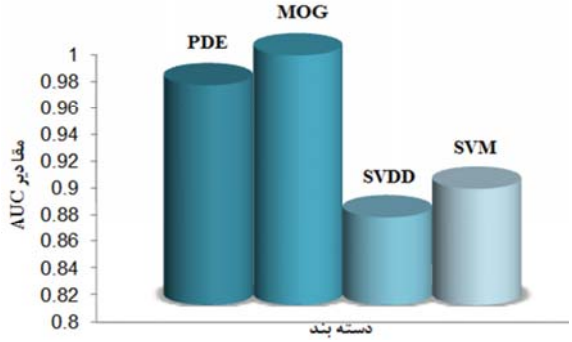
مقدار AUC نیز با استفاده از این دسته‌بندها در شکل (۶) نشان داده شده است.



شکل ۵. منحنی ROC دسته‌بندهای تک‌کلاسی

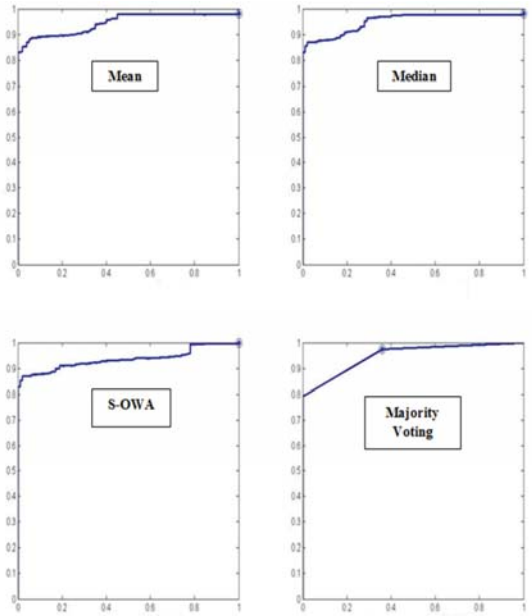
شکل (۷) نمودار ROC مربوط به داده‌های حمله‌های مختلف موجود در مجموعه‌دادگان را با استفاده از استراتژی‌های مختلف ترکیب دسته‌بندهای مذکور را نشان می‌دهد.

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

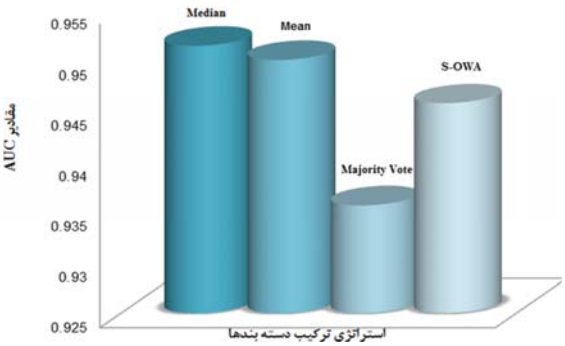


شکل ۶. مقادیر AUC دسته‌بندهای تک‌کلاسی

[1] Receiver Operating Characteristics Curve
[2] Area Under the Curve
[3] Maximum Likelihood Estimation



شکل ۷. منحنی‌های ROC روش‌های ترکیب دسته‌بندهای تک‌کلاسی



شکل۸. مقادیر AUC روش‌های ترکیب دسته‌بندهای تک‌کلاسی

با توجه به نمودارهای ROC به دست آمده، مشاهده می‌شود که مساحت زیر نمودار ROC به طور قابل ملاحظه‌ای از مساحت بالای منحنی بیشتر است که این امر نشان‌دهنده تستی موفقیت‌آمیز است.

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.

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.

جدول ۲. نتایج حاصل از روش پیشنهادی

عنوان روش	نرخ تشخیص	نرخ هشدار نادرست
SVDD	۹۸/۹	۲/۲۸
MOG	۹۵/۸	۳/۸
PDE	۹۸/۶	۴
SVM	۹۷/۷۷	۳/۰۱
استراتژی میانه	۹۸/۹	۱/۸
استراتژی رای اکثریت	۹۹/۴	۲/۱۳
استراتژی میانگین	۹۹/۱	۲/۰۸
استراتژی S-OWA	۹۹	۰/۲

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.

جدول ۳. ارزیابی WAF ارائه شده در [۳]

نرخ تشخیص	۹۵/۷
نرخ هشدار نادرست	۴/۷

با مقایسه نتایج ارزیابی WAF با روش‌هایی که ما در این پژوهش از آن‌ها بهره گرفتیم، مشاهده می‌شود نرخ تشخیص آن از بسیاری 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 تک‌کلاسی است، هرچند از آن‌ها کمتر است. نکته قابل توجه نرخ هشدار نادرست بسیار پایین‌تر 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. نتیجه‌گیری

در این مقاله از ترکیب دسته‌بندهای تک‌کلاسی 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 دسته‌بندها از استراتژی‌های رایج ترکیب دسته‌بندها برای تصمیم‌گیری گروهی استفاده شده است؛ همچنین از عملگر S-OWA، به منظور ترکیب دسته‌بندهای تک‌کلاسی استفاده شده است. استفاده از تصمیم‌گیری گروهی به ویژه با روش S-OWA، معیارهای کارایی سامانه تشخیص ناهنجاری را به خوبی بهبود بخشیده است. فرآیند تشخیص ناهنجاری با تصمیم‌گیری گروهی بر مبنای عملگر S-OWA سبب

بهبود نرخ هشدار نادرست به طور چشمگیری می‌شود و نرخ تشخیص مناسبی نیز دارد، به طوری که نرخ تشخیص به ۹۹ درصد و نرخ 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 روش‌های مختلف تأمین امنیت برنامه‌های کاربردی تحت وب و ابزارهای گوناگون آن، عدم استفاده این ابزارها و روش‌ها از سامانه ترکیبی پیشنهادی ما، صحت 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

[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.csic.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.