Journal of Information and Communication Technology

Volume 16, Issue 59-60, Spring and Summer 2024, pp. 84-92

Anomaly and Intrusion Detection Through Data Mining and Feature Selection using PSO Algorithm

Fereidoon Rezaei[1], Mohammad Ali Afshar Kazemi[2*], Mohammad Ali Keramati[2]

[1] Department of Information Technology Management, Central Tehran Branch, Islamic Azad University, Tehran, Iran
[2] Department of Industrial Management, Central Tehran Branch, Islamic Azad University, Tehran, Iran

Received: 17 January 2023, Revised: 07 March 2023, Accepted: 04 April 2023
Paper type: Research

Abstract

Today, considering technology development, increased use of Internet in businesses, and movement of business types from physical to virtual and internet, attacks and anomalies have also changed from physical to virtual. That is, instead of thieving a store or market, the individuals intrude the websites and virtual markets through cyberattacks and disrupt them. Detection of attacks and anomalies is one of the new challenges in promoting e-commerce technologies. Detecting anomalies of a network and the process of detecting destructive activities in e-commerce can be executed by analyzing the behavior of network traffic. Data mining systems/techniques are used extensively in intrusion detection systems (IDS) in order to detect anomalies. Reducing the size/dimensions of features plays an important role in intrusion detection since detecting anomalies, which are features of network traffic with high dimensions, is a time-consuming process. Choosing suitable and accurate features influences the speed of the proposed task/work analysis, resulting in an improved speed of detection. In this article, by using data mining algorithms such as Bayesian, Multilayer Perceptron, CFS, Best First, J48 and PSO, we were able to increase the accuracy of detecting anomalies and attacks to 0.996 and the error rate to 0.004.

Keywords: PSO, J48, data mining, cyberattack, NLC-KDD.

[*] Corresponding Author's email: m_afsharkazemi@iauec.ac.ir

دوفصلنامه
فناوری اطلاعات و ارتباطات ایران

سال شانزدهم، شماره های ۵۹ و ۶۰، بهار و تابستان ۱۴۰۳، صفحه ۸۴ الی ۹۲

تشخیص نفوذ و ناهنجاری‌ها با استفاده از داده‌کاوی و انتخاب ویژگی‌ها به‌وسیله الگوریتم PSO

فریدون رضائی[۱]، محمدعلی افشار کاظمی[۲]*، محمدعلی کرامتی[۲]

[۱] گروه مدیریت فناوری اطلاعات، واحد تهران مرکزی، دانشگاه آزاد اسلامی، تهران، ایران

[۲] گروه مدیریت صنعتی، واحد تهران مرکزی، دانشگاه آزاد اسلامی، تهران، ایران

تاریخ دریافت: ۱۴۰۱/۱۰/۲۷ تاریخ بازبینی: ۱۴۰۱/۱۲/۱۶ تاریخ پذیرش: ۱۴۰۲/۰۱/۱۵

نوع مقاله: پژوهشی

چکیده

امروزه با توجه به پیشرفت فناوری و توسعه استفاده از اینترنت در کسب و کارها و تغییر نوع کسب و کارها از حالت فیزیکی به مجازی و اینترنت، باعث شده است که نوع حملات و ناهنجاری‌های مرتبط نیز از حالت فیزیکی به حالت مجازی تغییر کند. یعنی بجای دستبرد به یک فروشگاه یا مغازه، افراد با استفاده از حملات سایبری به سایت‌ها و فروشگاه‌های مجازی نفوذ کرده و در آنها اخلال ایجاد می‌کنند. آشکارسازی حملات و ناهنجاری‌ها یکی از چالش‌های جدید در مسیر پیشبرد تکنولوژی تجارت الکترونیک می‌باشد. تشخیص ناهنجاری‌های یک شبکه و فرآیند شناسایی فعالیت‌های مخرب در کسب و کارهای تجارت الکترونیک با تجزیه و تحلیل رفتار ترافیک شبکه امکان‌پذیر است. سیستم‌های داده‌کاوی بطور گسترده‌ای در سیستمهای تشخیص نفوذ (IDS) برای تشخیص ناهنجاری‌ها استفاده می‌شوند. کاهش ابعاد ویژگی‌ها نقش بسیار مهمی در تشخیص نفوذ ایفا می‌کند، زیرا تشخیص ناهنجاری‌ها از ویژگی‌های ترافیک شبکه با ابعاد بالا فرآیندی زمان‌بری است. انتخاب ویژگی‌های درست و مناسب بر سرعت تجزیه و تحلیل و کار پیشنهادی تاثیر می‌گذارد و می‌تواند سرعت تشخیص را بهبود بخشد. در این مقاله با استفاده از الگوریتم‌های داده‌کاوی مانند بیزین، پرسپترون چندلایه، CFS، Best First، J48 و PSO، میزان دقت تشخیص ناهنجاری‌ها و حملات به ۰,۹۹۶ و میزان خطای آن ۰,۰۰۴ رسانده شده است.

کلیدواژگان: PSO، J48، داده‌کاوی، حملات سایبری، NLC-KDD.

* رایانامهٔ نویسنده مسؤول: m_afsharkazemi@iauec.ac.ir

۱- مقدمه

امروزه اینترنت به یک جزء ضروری از زیرساخت‌های اجتماعی و اقتصادی روزمره مردم تبدیل شده است. حجم بالای اطلاعات محرمانه و امنیتی اینترنت باعث می‌شود انواع تهدیدات و حملات مختلف در آن به وجود آید که ممکن است باعث خسارت مالی، سرقت هویت، از دست دادن اطلاعات خصوصی، آسیب شهرت نام تجاری و از دست دادن اعتماد مشتریان در تجارت الکترونیک شود [۱]. افزایش دامنه استفاده از تجارت الکترونیک باعث افزایش مصرف انرژی، پیچیده تر شدن مدیریت آن، افزایش حجم داده‌ها، نیاز به پهنای باند بالا به جهت ارسال داده‌ها و سیستم‌های پردازش با سرعت بالا می‌باشد. یکی از مهمترین این چالش‌ها حفظ حریم خصوصی و امنیت اطلاعات می‌باشد [۵، ۴، ۳]. یقینا رضایت مصرف‌کنندگان این تکنولوژی به این چالش‌ها گره خورده است. تهدیدات و حملاتی که در زمینه تجارت الکترونیک رخ می‌دهد را می‌توان به چهاردسته کلی تقسیم‌بندی نمود: ۱) حملات DOS، ۲) حملات R2L، ۳) حملات U2R، و ۴) حملات PROB.

عواملی که تهدید و حمله را در یک شبکه اینترنتی به وجود می‌آورند عبارتند از: دسترسی بدون محدودیت به اینترنت، گمنامی افراد، سرعت بالای انتشار، عدم ارتباط چهره به چهره، دسترسی آزاد به خدمات و محتویات ارزشمند و همچنین عدم وجود قوانین مناسب [۶]؛ بنابراین، مناسب بودن اینترنت به‌عنوان یک کانال برای انجام معاملات تجاری مطرح می‌شود.

سیستم‌های تشخیص نفوذ مبتنی بر داده‌کاوی می‌تواند بسیار موثر باشد و می‌تواند با استفاده از آموزش در شبکه‌های واقعی و مجموعه داده‌های واقعی و ممیزی نشده، روش‌های جدید و ترکیبی را شناسایی کند. در واقع سیستم تشخیص با گرفتن بسته‌ای ورودی به عنوان یک دروازه عمل می‌کند و شروع به جمع آوری اطلاعات می‌کند و با پیش‌پردازش داده‌ها ویژگی‌های بی‌استفاده آنها را حذف کرده و سپس داده‌ها را برای پردازش ارسال می‌کند و با تجزیه و تحلیل داده‌ها آنها را طبقه‌بندی می‌کند و سپس رکوردهای طبیعی به شبکه اصلی ارسال می‌شود و رکوردهای غیر طبیعی به مراحل پردازش بعدی ارسال می‌گردد [۷].

در این مقاله سعی شده است با استفاده از الگوریتم‌های داده‌کاوی مانند بیزین، پرسپترون چندلایه، CFS، Best First، J48 و PSO، ابعاد ویژگی‌های داده‌ها کاهش داده شده تا هم سرعت و هم دقت الگوریتم را برای تشخیص ناهنجاری‌ها بهبود یابد. در واقع با کاهش ویژگی‌ها در داده‌ها میزان زمان جستجوی داده‌ها کاهش یافته که این کار دقت تشخیص ناهنجاری‌ها را افزایش داده است، که نتایج حاصله نشان دهنده این موضوع می‌باشد.

در این مقاله از نرم‌افزار وکا نسخه ۳،۹،۲ (Weka 3.9.2) استفاده شده است[۸].

۲- مروری بر کارهای انجام شده

در سال‌های گذشته افراد بسیاری بر روی داده‌کاوی و استفاده از آن در تشخیص نفوذ و ناهنجاری‌ها استفاده کرده‌اند. که در ادامه به تعدادی از آنها پرداخته شده است:

در مقاله [۲،۲۵] از شبکه عصبی به عنوان آموزش عمیق[1] و به جهت تشخیص حملات و ناهنجاری‌های سیستم تجارت الکترونیک استفاده می‌شود. درشبکه‌های عصبی چندلایه بروز خطای بیش‌برازش[2] بسیار رایج است. در این مقاله برای جلوگیری از این اثر، ویژگی‌ها را با استفاده از الگوریتم کرم شب‌تاب کاهش داده‌اند. نتایج شبیه‌سازی نشان می‌دهد سیستم شبکه عصبی با استفاده از کاهش ویژگی دقت بسیار بالاتری دارد.

آقای اندرسون در سال ۱۹۸۰ با استفاده از داده‌های ماهانه و هفتگی در سیستم‌های حسابرسی به بررسی رکوردها پرداخته و با استفاده از پردازش آن‌ها به بررسی دسترسی‌های غیر مجاز رسیده و مجموعه قوانینی را ایجاد کرده و سیستمی را ارائه می‌دهد [۱].

یکی از راه‌های مقابله با حملات و تهدیدات در شبکه گسترده تجارت الکترونیک استفاده از مدل‌ها و سازوکارهای یادگیری ماشین می‌باشد [۹-۱۰] یادگیری ماشین زیرمجموعه‌ای از فناوری هوش مصنوعی[3] است که عمدتا روی یادگیری ماشین‌ها بر اساس تجربیات خود ماشین و پیش‌بینی‌های مبتنی بر این تجربیات استوار است. الگوریتم‌های یادگیری ماشین، با استفاده از مجموعه داده‌هایی با عنوان داده‌های آموزشی یادگیری کرده و مدل‌های موردنیاز را ایجاد می‌کنند. زمانی که داده‌های جدیدی به الگوریتم یادگیری ماشین معرفی می‌شوند، سیستم می‌تواند بر اساس مدل ایجاد شده، فرایند پیش‌بینی را انجام دهد. یکی از معروفترین، پرکاربردترین و قوی‌ترین الگوریتم‌ها و مدل‌های یادگیری ماشین شبکه عصبی مصنوعی می‌باشد [۱۱-۱۲].

سینک و همکاران در سال ۲۰۱۵ از چهار الگوریتم پیش‌پردازش

[1] Deep Learning
[2] Overfitting
[3] Artificial Intelligence

CFS، IG، زیر مجموعه مبتنی بر سازگاری[1]، PCA برای کاهش ابعاد و پنج الگوریتم دسته‌بندی؛ J48، بیزین ساده، SVM، جنگل تصادفی و آدابوست برای ارزیابی عملکرد الگوریتم‌های کاهش ویژگی از نظر دقت و AUC استفاده کردند [۱۳].

علیزاده بهرامی و همکاران در سال ۱۳۹۶ از روش‌های انتخاب ویژگی‌ها و درخت تصمیم‌گیری J48 برای کاهش ویژگی‌ها و تشخیص نفوذ استفاده کردند[۱۴]. بهارلو و همکاران در سال ۱۳۹۹ از روش دو مرحله‌ای Wrapper+PCA و الگوریتم جنگل تصادفی برای تشخیص و شناسایی وب سایت فیشینگ استفاده کرده است [۱۵].

در مقاله [۱۶] نیز یک روش کاهش ویژگی ترکیبی هایبرید[2] ارائه شده است که در کنار روش طبقه‌بندی جنگل تصادفی بهترین عملکرد را داشته است. در مقاله [۱۷] با استفاده از الگوریتم‌های یادگیری ماشین علاوه بر استفاده از یک مجموعه داده جدید مربوط به تشخیص فیشینگ که شامل ۵۰۰۰ صفحه وب قانونی و ۵۰۰۰ فیشینگ است، به این مسئله پرداخته که به منظور دستیابی به بهترین نتایج، الگوریتم‌های مختلف یادگیری ماشین مورد آزمایش قرار گرفتند و نهایتا J48، جنگل تصادفی و پرسپترون چند لایه انتخاب شدند. در مقاله [۱۸] نیز روش‌های یادگیری ماشین بطور کلی در کنار کاهش ویژگی‌ها برای تشخیص صفحات وب و ایمیل فیشینگ مورد بررسی و ارزیابی قرار گرفته‌اند.

۳- مروری بر ساختار و طرح پیشنهادی

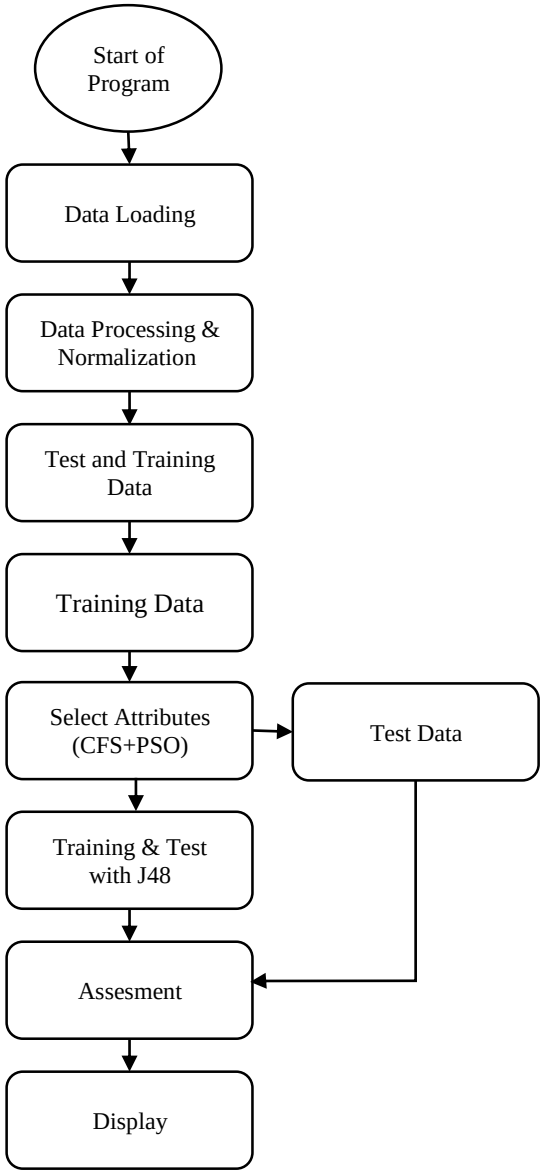
ساختار و فلوچارت کلی طرح پیشنهادی در شکل ۱ ترسیم شده است. همانطور که در شکل نشان داده شده است در گام اول می‌بایست داده‌های سیستم پیشنهادی جمع آوری شود. در این مقاله از مجموعه داده‌های NSL-KDD استفاده شده است. در مرحله دوم عملیات پیش پردازش داده‌ها شامل پاکسازی داده‌های مشابه، نرمالیزه کردن داده‌ها و مهندسی ویژگی داده‌ها[3] و بردارسازی آنها انجام می‌شود. در این مقاله ابتدا با استفاده از الگوریتم PSO به جهت کاهش و استخراج ویژگی‌ها استفاده می‌شود. در ادامه داده‌هایی که به صورت جمع و جور تبدیل شده‌اند به دو دسته داده‌های آموزش (۸۰ درصد داده‌ها) داده‌های تست (۲۰ درصد داده‌ها) تبدیل می‌شوند. در انتها مدل پیشنهادی برروی داده‌های نهائی اعمال شده است.

۳-۱- مجموعه داده‌های NSL-KDD

دیتاست مجموعه‌ای از داده‌های گردآوری شده در رابطه با یک موضوع واحد بوده و بیشترین کاربرد آن در داده‌کاوی است اما یکی از ابزارهای بسیار مناسب و کارآمد برای آزمون و ارزیابی الگوریتم‌های طراحی شده در یک حوزه خاص نیز به شمار می‌رود برای مثال دیتاست KDD CUP 99 با هدف آزمون الگوریتم‌های تشخیص نفوذ (Intrusion Detection) گردآوری و طراحی شده است این مجموعه داده با استفاده از حجم عظیم داده‌های گردآوری شده در پروژه DIDE یا Darpa Intrusion Detection Evalution که با همکاری سازمان پروژه‌های تحقیقاتی پیشرفته دفاعی، وزارت دفاع ایالات متحده آمریکا و آزمایشگاه لینکلن دانشگاه MIT انجام شد، تهیه گردیده است هدف از تهیه این دادگان، ایجاد یک مجموعه داده استاندارد برای ارزیابی سیستم‌های تشخیص نفوذ (Intrusion Detection System) است [۲۰].

از این رو کلیه رکوردهای موجود در این مجموعه داده، توسط افراد خبره در حوزه امنیت اطلاعات برچسب گذاری شده است بگونه‌ای که تعلق هر رکورد به کلاس خاصی از حمله و یا عادی بودن رکورد به آسانی قابل تشخیص است. این دادگان از دو مجموعه داده جداگانه تشکیل می‌شود که عبارتند از : مجموعه داده‌های آموزشی (Training) که مجموعه یادگیری نیز نامیده می‌شود و مجموعه آزمون (Test) که از مجموعه یادگیری برای تحلیل دقیق رفتار حمله و تدوین قوانین موثر و کارآمد استفاده می‌شود و برای آزمون و ارزیابی الگوریتم پیشنهادی نیز از هر دو مجموعه یادگیری و آزمون استفاده می‌شود. یکی از دیتاست‌های مطرح برگرفته شده از KDD CUP 99 دیتاست NLS-KDD است [۱۴] که با انجام تحلیل‌های آماری دقیق در خصوص دیتاست KDD CUP 99 و برای حل برخی از مشکلات ذاتی دیتاست KDD CUP 99 تهیه گردیده است که نسبت به KDD CUP 99 دارای برتری‌های زیر است [۱۹]:

[1] Consistency Subset
[2] HEFS
[3] Feature engineering



شکل ۱. فلوچارت استخراج مدل تشخیص نفوذ از روی مجموعه دیتاست NSL-KDD

هر دو مجموعه داده‌های یادگیری و آزمون فاقد رکورد تکراری هستند که این ویژگی موجب بالاتر رفتن دقت و کارآیی الگوریتم‌های داده‌کاوی و یادگیری ماشینی شده و مانع از تاثیر منفی رکوردهای تکراری بر خروجی الگوریتم خواهد شد. تعداد رکوردها در مجموعه یادگیری و آزمون مناسب و خردمندانه انتخاب شده است که این ویژگی سرعت الگوریتم‌های یادگیری ماشینی و داده‌کاوی را افزایش می‌دهد.

جدول ۱. مقایسه تعداد رکوردها در دو مجموعه داده[۱۹]

S/N	Name	Type	
1	Duration	Basic Features	
2	protocol_type		
3	Service		
4	flag		
5	src_bytes		
6	dst_bytes		
7	Land		
8	wrong_fragment		
9	Urgent		
10	Hot		
11	num_failed_logins	Content features	
12	logged_in		
13	num_compromised		
14	root_shell		
15	su_attempted		
16	num_root		
17	num_file_creations		
18	num_shells		
19	num_access_files		
20	num_outbound_cmds		
21	Is_hot_login		
22	Is_guest_login		
23	Count	Same service features	Traffic features
24	Srv_count		
25	serror_rate		
26	srv_serror_rate		
27	rerror_rate		
28	srv_rerror_rate		
29	same_srv_rate		
30	diff_srv_rate		
31	srv_diff_host_rate		
32	dst_host_count	Same host features	
33	dst_host_srv_count		
34	dst_host_same_srv_rate		
35	dst_host_diff_srv_rate		
36	dst_host_same_src_port_rate		
37	dst_host_srv_diff_host_rate		
38	dst_host_serror_rate		
39	dst_host_srv_serror_rate		
40	dst_host_rerror_rate		
41	dst_host_srv_rerror_rate		

جدول ۲. ویژگی‌های دیتاست NSL-KDD قبل کاهش ویژگی‌ها[۱۹]

تعداد رکوردها در دیتاست NLS-KDD	تعدا رکوردها در دادگان KDD CUP 99
۱۲۵۹۷۳	۴۹۴۰۲۱

مجموعه داده NLS-KDD شامل ۴۲ ویژگی یا فیلد است که عبارتند از: ۴۱ ویژگی عادی مربوط به اتصالات شبکه و یک ویژگی کلاس که در آن ۵ کلاس مختلف شامل یک کلاس عادی (Normal) و ۴ کلاس حمله تعریف شده است. کلاس‌های حمله عبارتند از : R2L، U2R، DoS و Prob [۱۹].

۴- کاهش ویژگی‌ها

یکی از روش‌های کاهش ابعاد داده، روش پوششی می‌باشد. در این روش انتخاب زیر مجموعه ویژگی با استفاده از الگوریتم یادگیری انجام می‌شود. الگوریتم دسته‌ای از ویژگی‌ها را برای یادگیری انتخاب می‌نماید و نهایتا آن دسته از ویژگی‌ها که دقت بالاتری دارند، انتخاب می‌شود. الگوریتمی که کار ارزیابی زیر مجموعه ویژگی‌ها و انتخاب بهترین زیر مجموعه را انجام می‌دهد، خود به‌عنوان بخشی از تابع ارزیابی، کار جستجو برای انتخاب بهترین مدل را انجام می‌دهد[۱۵].

۴-۱- روش CFS

CFS مقدار همبستگی بین ویژگی‌ها و کلاس‌هایشان و همچنین همبستگی بین خود ویژگی‌ها را اندازه‌گیری می‌کند. ایده کلی این است که زیر مجموعه ویژگی‌های خوب، همبستگی زیادی با کلاس‌ها دارند، اما نباید با یکدیگر همبستگی داشته باشند [۱۵]. در الگوریتم CFS، هیوریستیکی برای ارزیابی ارزش یا شایستگی یک زیر مجموعه ویژگی وجود دارد [۸].

۴-۲- روش Best First

جستجوی ابتدا – بهترین (Best-First Search) یک الگوریتم جستجو است که یک گراف را با بسط دادن محتمل‌ترین راس، که بنابر قوانین خاص انتخاب می‌شود، پیمایش می‌کند. این نوع جستجو را به‌عنوان تخمین تعهد نود n به وسیله Heuristic Evaluation Function توصیف می‌کند، که به‌صورت کلی، ممکن است بر پایه توصیف n، توصیف هدف، اطلاعات جمع آوری شده به وسیله جستجو تا آن نقطه و هر گونه اطلاعات اضافی در زمینه مسئله باشد [۲۱].

۴-۳- الگوریتم بیزین (Bayes Net)

الگوریتم بیزین یکی از الگوریتم‌های مدل‌سازی است. در این روش، دسته‌بندی برپایه احتمالات و با فرض استقلال متغیرهای تصادفی ساخته می‌شود. این روش از ساده‌ترین الگوریتم‌های دسته‌بندی است که دقت قابل قبولی داشته و بر پایه احتمال وقوع یا عدم وقوع یک پدیده شکل می‌گیرد [۲۰].

۴-۴- الگوریتم پرسپترون چند لایه

پرسپترون چند لایه[1]، دسته‌ای از شبکه‌های عصبی مصنوعی پیشخور است. یک MLP شامل حداقل سه لایه گره است: یک لایه ورودی، یک لایه پنهان و یک لایه خروجی. به جز گره‌های ورودی، هر گره یک نورون است که از یک تابع فعالسازی غیر خطی استفاده می‌کند. MLP از تکنیک یادگیری نظارت شده به نام بازپرداخت برای آموزش استفاده می‌کند. لایه‌های متعدد آن و فعال‌سازی غیر خطی آن MLP را از یک پرسپترون خطی متمایز می‌کند. در واقع می‌تواند داده‌هایی را متمایز کند که به‌صورت خطی قابل تفکیک نیستند[۲۲].

۴-۵- الگوریتم J48

این روش از معیار شاخص جینی[2]جهت انتخاب ویژگی استفاده می‌کند [۲۳]. از میان ویژگی‌ها، هرکدام که مقدار شاخص جینی آن کوچک‌تر است، برای گروه جاری درخت تصمیم در نظر گرفته می‌شود.

۴-۶- الگوریتم PSO

روش بهینه‌سازی ازدحام ذرات[3] یا به اختصارPSO ، یک روش سراسری بهینه‌سازی است که با استفاده از آن می‌توان با مسائلی که جواب آن‌ها یک نقطه یا سطح در فضای n بعدی می‌باشد، برخورد نمود. در اینچنین فضایی، فرضیاتی مطرح می‌شود و یک سرعت ابتدایی به آن‌ها اختصاص داده می‌شود، همچنین کانال‌های ارتباطی بین ذرات در نظر گرفته می‌شود. سپس این ذرات در فضای پاسخ حرکت می‌کنند، و نتایج حاصله بر مبنای یک «ملاک شایستگی» پس از هر بازه زمانی محاسبه می‌شود. با گذشت زمان، ذرات به سمت ذراتی که دارای ملاک شایستگی بالاتری هستند و در گروه ارتباطی یکسانی قرار دارند، شتاب می‌گیرند [۲،۲۵].

۵- روش پیشنهادی

همانطوریکه در شکل شماره ۱ مشاهده می‌کنید، روش پیشنهادی شامل مراحل زیر است؛ یک جمع‌آوری داده و نرمال‌سازی آنها، دوم تقسیم داده‌ها به دو قسمت آموزش و تست. سوم اعمال روش‌های انتخاب ویژگی‌ها و در نهایت مدل‌سازی و تست.

در اینجا با استفاده از نرم‌افزار وکا [۸]، داده‌ها را بارگذاری کرده و

[1] Multilayer Perceptron

[2] Gini coefficient

[3] Particle swarm optimization

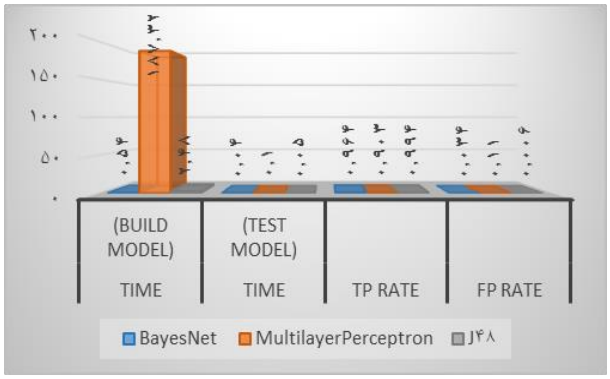
الگوریتم‌ها و روش‌های مختلف را بر روی آنها اعمال می‌کنیم.

در ابتدا الگوریتم CFS را به‌عنوان ارزیاب و الگوریتم BestFirst را به‌عنوان جستجوگر جهت کاهش ویژگی‌ها بر روی داده‌ها اعمال می‌کنیم. تعداد ویژگی‌ها از ۴۱ مورد (جدول ۲) به ۶ ویژگی (جدول ۳) کاهش می‌یابد.

جدول ۳. انتخاب ویژگی‌ها با استفاده از CFS+BestFirst

S/N	Name	Type
4	flag	Symbolic
5	src_bytes	Continuous
6	dst_bytes	Continuous
12	logged_in	Symbolic
26	srv_serror_rate	Continuous
30	diff_srv_rate	Continuous

سپس الگوریتم‌های بیزین، پرسپترون چندلایه و درخت تصمیم گیری J48 را با روش تست Cross-validation و Folds برابر ۱۰ بر روی آنها اعمال می‌کنیم و نتایج زیر حاصل می‌شود:



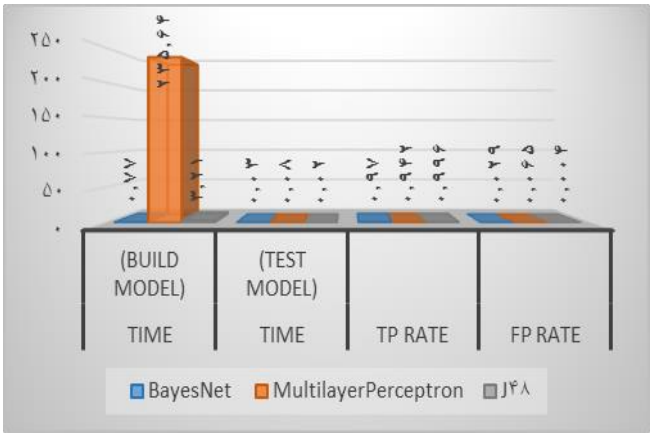
شکل ۲. نمودار ۱. نتایج حاصل از پیاده‌سازی الگوریتم‌های بیزین، پرسپترون چندلایه و J48 بر روی ۶ ویژگی

همانطوریکه در نمودار ۱؛ مشاهده می‌کنید، الگوریتم J48 بهترین نتایج را دارد. حال برای انتخاب ویژگی‌ها به‌جای BestFirst از الگوریتم PSO استفاده می‌کنیم. در این حالت تعداد ویژگی‌ها به ۹ تا به شرح جدول ۴ کاهش یافت.

جدول ۴. انتخاب ویژگی‌ها با استفاده از CFS+PSO

S/N	Name	Type
4	flag	Symbolic
5	src_bytes	Continuous
6	dst_bytes	Continuous
12	logged_in	Symbolic
26	srv_serror_rate	Continuous
29	same_srv_rate	Continuous
30	diff_srv_rate	Continuous
37	dst_host_srv_diff_host_rate	Continuous
39	dst_host_srv_serror_rate	Continuous

حال الگوریتم‌های بیزین، پرسپترون چندلایه و درخت تصمیم گیری J48 را بر روی ۹ ویژگی انتخاب شده در جدول شماره ۴ با روش تست Cross-validation و Folds برابر ۱۰ بر روی آنها اعمال می‌کنیم. نتایج حاصل در نمودار ۲ ارائه شده است.



شکل ۳. نمودار ۲. نتایج حاصل از پیاده‌سازی الگوریتم‌های بیزین، پرسپترون چندلایه و J48 بر روی ۹ ویژگی

۶- مقایسه نتایج

با مقایسه نتایج حاصل از دو روش انتخاب ویژگی‌ها و سپس اعمال الگوریتم‌های مشابه به دو نتایج حاصل از دو حالت، همانطوریکه در جدول ۵ مشاهده می‌کنید، درخت تصمیم‌گیری J48 هم دارای زمان مناسب و نتایج مناسب نسبت به روش‌های بیزین و پرسپترون چندلایه می‌باشد. البته در روش دوم نتایج حاصل با اعمال الگوریتم جستجوی PSO بهبود یافت و نتایج آن بهتر شده است.

جدول ۵. نتایج حاصل از مقایسه دو روش پیاده‌سازی شده

Method (Select Attributes)	Attribute	Model	Time (Build Model)	Time (Test Model)	TP Rate	FP Rate
CFS+BestFirst	6	J48	2.48	0.05	0.994	0.006
CFS+PSO	9	J48	3.21	0.02	0.996	0.004
The rate of improvement			-0.73	0.03	-0.002	0.002

همانطوریکه در جدول شماره ۷ مشاهده می‌کنید، با استفاده از روش انتخاب ویژگی‌های CFS+PSO سیستم عملکرد بهتری داشته است. تشخیص درست (TP) به میزان ۰،۰۰۲ بهبود داشته و در تشخیص اشتباهات نیز به میزان ۰،۰۰۲ نیز بهبود داشته است. توجه داشته باشید که TP هر چه بیشتر باشد، عملکرد سیستم بهتر بوده و FP هر چه کمتر باشد سیستم بهتر عمل کرده است.

۷- بحث

همانطوریکه در جدول ۵ مشاهده می‌کنید، نتایج حاصل از دو روش پیاده‌سازی شده یعنی CFS+BestFirst و CFS+PSO نشان میدهد که در حالت استفاده از الگوریتم PSO میزان زمان ساخت مدل افزایش داشته ولی زمان تست مدل کاهش نشان میدهد. در ضمن میزان TP بهتر شده و میزان FP هم کاهش پیدا می‌کند که این

خروجی‌ها نشانه بهبود روش تشخیص می‌باشد.

در [۲۴] از الگوریتم PSO برای آموزش شبکه عصبی برای تشخیص حملات و ناهنجاری‌های سیستم اینترنت اشیاء بر روی داده‌های NLC-KDD استفاده شده است. اگر چه الگوریتم PSO دارای مزایای زیادی است، اما در برخی موارد ممکن است تنوع جمعیت را کاهش داده و منجر به همگرایی زودرس شود. بنابراین برای حل این مشکل از الگوریتم TLBO استفاده شده است، که دقت تشخیص حمله تا ۹۰ درصد رسیده است. این در حالیست که در این مقاله، الگوریتم PSO با الگوریتم CFS ترکیب شده و ویژگیها کاهش داده شده است و سپس با استفاده از درخت تصمیم گیری J48 عملیات تشخص حملات انجام شده است که باعث بهبود عملکرد شده و میزان دقت تشخیص بطور قابل توجهی افزایش یافته و به میزان ۹۹,۶ درصد رسیده است.

در [۲،۲۵] با استفاده از الگوریتم کرم شب تاب (FA) بر روی داده‌های NLC-KDD، ابتدا ویژگیهای کاهش داده شده، که همین موضوع باعث افزایش دقت تشخیص شده است و سپس با استفاده از ترکیب الگوریتم فوق با الگوریتم PSO میزان دقت[1] به ۹۴,۳ رسیده است، این در حالیست که در این مقاله ابتدا با استفاده از ترکیب دو الگوریتم PSO و CFS باعث کاهش ویژگی‌ها شده و سپس با استفاده از الگوریتم درخت تصمیم J48 میزان دقت تشخیص را بطور چشم‌گیری افزایش داده و در نتیجه این روش از روش استفاده شده در مقاله فوق بهتر عمل کرده است.

یکی از محدودیتهای مطالعه نبود داده‌های واقعی برای تشخیص حملات تجارت الکترونیک بود. برای کم کردن اثر این مشکل از داده‌های KDD CUP که از لحاظ ویژگی‌ها شباهت بسیاری به داده‌های تجارت الکترونیک دارند استفاده شده است.

۸- نتیجه‌گیری

همانطوریکه در مقایسه نتایج مشاهده شد، استفاده از داده‌کاوی و ترکیب آن با الگوریتم‌هایی مانند بهینه‌سازی ازدحام ذرات و یا همان الگوریتم حرکت پرندگان، می‌توان میزان تشخیص درست را نسبت به روش‌های دیگر بهبود داده و میزان تشخیص‌های نادرست را نیز کاهش داد. در ضمن کاهش ویژگی‌ها، هم می‌تواند در زمان جستجو بسیار موثر بوده و هم در بهینه‌سازی عملکرد الگوریتم موثر باشد.

از روش فوق میتوان در تشخیص سایتهای فیشینگ و همچنین تشخیص ویروس‌های کامپیوتری بخصوص متامورفیک‌ها استفاده نمود. همچنین استفاده از مدل کاهش ویژگی‌ها که در این مقاله استفاده شده است، می‌تواند در افزایش سرعت تشخیص حملات در روشهای مختلف مورد استفاده قرار گیرد.

مراجع

[1] Abdelhamid, N., Ayesh, A., Thabtah, F., “Phishing detection based Associative Classification data mining”, Expert Systems with Applications 41 5948–5959, 2014.

[2] Rezaei F, Afshar Kazemi M A, Keramati M A. Detection of E-commerce Attacks and Anomalies using Adaptive Neuro-Fuzzy Inference System and Firefly Optimization Algorithm . itrc 2021; 13 (1) :32-39 URL: http://ijict.itrc.ac.ir/article-1-477-en.html

[3] Hasan, Mahmudul, et al. "Attack and anomaly detection in IoT sensors in IoT sites using machine learning approaches." Internet of Things 7 (2019): 100059.

[4] Kotenko, Igor, et al. "Attack detection in IoT critical infrastructures: a machine learning and big data processing approach." 2019 27th Euromicro International Conference on Parallel, Distributed and NetworkBased Processing (PDP). IEEE, 2019.

[5] Foley, John, Naghmeh Moradpoor, and Henry Ochen. "Employing a Machine Learning Approach to Detect Combined Internet of Things Attacks against Two Objective Functions Using a Novel Dataset." Security and Communication Networks 2020 (2020).

[6] Assistant, Masoud, "Detection of attacks in electronic banking using fuzzy-rough combined system" computer department of Imam Reza University (AS), 2014.

[7] Al-jarrah, O., Arafat, A., “Network Intrusion Detection System using attack behavior classification.”, Paper presented at the Information and Communication Systems (ICICS), 2014 5th International Conference on.

[8] Kohavi, R., John, G. H., “Wrappers for feature subset selection”, Artificial Intelligence,Vol. 97, pp. 273-324, 1997.

[9] Doshi, Rohan, Noah Apthorpe, and Nick Feamster. "Machine learning ddos detection for onsumer internet of things devices." 2018 IEEE Security and Privacy Workshops (SPW). IEEE, 2018.

[10] Syed, Naeem Firdous, et al. "Denial of service attack detection through machine learning for the IoT." Journal of Information and Telecommunication (2020): 1-22.

[11] Manimurugan, S., et al. "Effective Attack Detection in Internet of Medical Things Smart Environment Using a Deep Belief Neural Network." IEEE Access 8 (2020): 77396-77404.

[12] Latif, Shahid, et al. "A Novel Attack Detection Scheme for the Industrial Internet of Things Using a Lightweight Random Neural Network." IEEE Access 8 (2020): 89337-89350.

[13] Singh, P., Jain, N., Maini, A., “Investigating the Effect Of Feature Selection and Dimensionality Reduction On Phishing Website Classification Problem”, 1st International Conference on Next Generation Computing Technologies (NGCT) Dehradun, India, IEEE, pp. 388-393, 2015.

[14] Alizadeh Bahrami, Karimi, Abdullahi Fard, "J48 Decision Tree in Intelligent Intrusion Detection Systems", National Conference on New Researches in Electrical, Computer and Medical Engineering, Islamic Azad University, Kazeroon Branch, July 27, 2016

[15] Baharlo, Yari, "Improving the method of identifying phishing websites using data mining on web pages", two scientific quarterly magazines of Iran Information and Communication Technology,

[1] Accuracy

[21] Fatemeh Mirjalili & Jafar Razmara, "An intelligent behavior-based intrusion detection method for virtual machines ", Signal and data processing journal, 2021, number 2, serial 48

[22] Rana, A., Singh Rawat, A., Bijalwan, A., & Bahuguna, H. (2018). "Application of Multi Layer (Perceptron) Artificial Neural Network in the Diagnosis System": A Systematic Review. 2018 International Conference on Research in Intelligent and Computing in Engineering (RICE). 978-1-5386-2599-6/18/$31.00 ©2018 IEEE

[23] Ghafari Gosheh, Zainab, 2019, intrusion detection using decision tree-based data mining technique, international conference on modern researches in electrical, computer, mechanical and mechatronic engineering in Iran and the Islamic world, Karaj, https://civilica.com/ doc/1118442

[24] M. Nazarpour, N. Nezafati, S. Shokouhyar, "Detection of Attacks and Anomalies in The Internet of Things System Using Neural Networks Based on Training with PSO and TLBO Algorithms", Signal Processing and Renewable Energy, ISSN: 2588-7327, eISSN: 2588-7335, December 2020, (pp. 81-94).

[25] Rezaei F, Afshar Kazemi M A, Keramati M A. (2023). Presenting a Model for Recognizing Phishing Sites and Privacy Violations in the Tourism Industry .Iranian Journal of Educational Sociology. 6(1), 222-232. Doi:10.61186/ijes.6.1.222 URL :http://iase-idje.ir/article-1-1261-en.html

Iran Information and Communication Technology Association, 12th year, numbers 43 and 44, Spring and summer 2019, pages 27-38

[16] K. L. Chiew, C. L. Tan, K. Wong, K. S. Yong, and W. K. Tiong, "A new hybrid ensemble feature selection framework for machine learningbased phishing detection system," Information Sciences, vol. 484, pp. 153–166, 2019.

[17] M. Almseidin, A. A. Zuraiq, M. Alkasassbeh, and N. Alnidami, "Phishing detection based on machine learning and feature selection methods," International Journal of Interactive Mobile Technologies (iJIM), vol. 13, no. 12, pp. 171–183, 2019.

[18] Meenu , Sunila godara, "Phishing Detection using Machine Learning Techniques", International Journal of Engineering and Advanced Technology (IJEAT) , Volume-9 Issue-2, December, 2019.

[19] S. Revathi, Dr. A. Malathi, "A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 2 ISSue 12, December-2013

[20] Rouhaninejad, Tayyaba, 2014, Combining Decision Tree and Bayesian Data Mining Algorithms in Intrusion Detection, Second National Conference on Computer Engineering and Information Technology Management, Tehran, https://civilica.com/doc/422878