مجله علمی- پژوهشی "پدافند الکترونیکی و سایبری"

سال ششم، شماره ۲، تابستان ۱۳۹۷، ص ۶۳-۴۹

انتخاب خصایص سامانه تشخیص نفوذ با استفاده از الگوریتم کلونی مورچگان به شیوه حرکت روبه‌جلو

مهدی عباسی[1*]، صادق بجانی[2]

۱- دانشجوی کارشناسی ارشد ۲-استادیار، دانشگاه جامع امام حسین(ع)

(دریافت: ۹۶/۰۳/۱۷، پذیرش: ۹۶/۰۹/۲۵)

چکیده

سامانه تشخیص نفوذ یکی از مهم‌ترین ابزارهای امنیتی در تشخیص حملات رایانه‌ای است که بر پایه یکی از دو روش تشخیص مبتنی بر سوءاستفاده و مبتنی بر ناهنجاری عمل می‌کند. مهم‌ترین چالش ارتقای آی.دی.اس، محدودیت زمانی پاسخ و استفاده از الگوریتم با کارایی پایین جهت شناسایی نفوذ است. انتخاب دقیق خصایصی از سامانه تشخیص نفوذ که بر پایه آن‌ها بتوان قدرت تشخیص را در این سامانه‌ها بالا برد، یکی از مراحل مهم در فرآیند تشخیص نفوذ است. در این مقاله شیوه‌ای جدید جهت تعیین مؤثرترین خصایص در سامانه تشخیص نفوذ مبتنی بر تشخیص سوءاستفاده، ارائه‌شده است. در این شیوه، خصایص مجموعه داده *NSL-KDD* با استفاده از الگوریتم بهینه‌سازی کلونی مورچگان، درحرکت روبه‌جلو با بهره‌گیری از الگوریتم دسته‌بندی *PART*، کاهش داده‌شده است. برای ارزیابی میزان موفقیت این شیوه، نرم‌افزاری به زبان جاوا پیاده‌سازی شده که در آن از توابع کتابخانه نرم‌افزار *WEKA* استفاده شده است. نتایج ارزیابی در مقایسه با سایر کارهای موفق، نشان می‌دهد که این طرح، نرخ صحت تشخیص نفوذ را با تعیین هم‌زمان دسته حمله، از متوسط ۸۴/۱٪ به ۸۵/۳۵٪ ارتقا داده است. همچنین زمان تشخیص نفوذ برای یک مجموعه داده حدوداً بیست هزار عضوی از متوسط ۰/۳۱ ثانیه به کم‌تر از ۰/۲۵ ثانیه کاهش یافته است.

واژه‌های کلیدی: تشخیص نفوذ، انتخاب خصایص، داده‌کاوی، الگوریتم کلونی مورچگان، الگوریتم PART

۱- مقدمه

بر اساس نظر آماروسو[1] [۱]، تشخیص نفوذ، فرآیند نظارت بر وقایع رخ‌داده در یک شبکه و یا سامانه رایانه‌ای جهت کشف موارد انحراف از سیاست‌های امنیتی و پاسخ به فعالیت‌های مشکوک علیه منابع پردازشی و شبکه‌ای است. سامانه‌های تشخیص نفوذ از حیث شیوه تحلیل و تشخیص، غالباً به دو دسته سامانه‌های تشخیص سوءاستفاده[2] (مبتنی بر امضا یا سناریو) و سامانه‌های تشخیص ناهنجاری تقسیم می‌شوند [۲]. سامانه‌های تشخیص ناهنجاری ابتدا نمایه‌هایی از رفتارهای هنجار(یا نرمال) از سامانه‌ای که در آن مستقر است را تشکیل داده، سپس هرگونه تخطی یا انحراف از نمایه هنجار که بالاتر از یک حد آستانه باشد را به‌عنوان رفتار ناهنجار و مهاجمانه تلقی می‌کند [۳]. در سامانه‌های تشخیص نفوذ مبتنی بر امضاء، سامانه با در اختیار داشتن مجموعه‌ای از الگوهای حمله، در بین هشدارهای موجود به دنبال هشدار یا زنجیره‌ای از هشدارها می‌گردد که با یکی از الگوهای حمله مطابقت داشته باشد [۲] لذا در سامانه‌های تشخیص نفوذ مبتنی بر امضا نیازمند مجموعه داده‌ای خواهد بود که در آن الگوهای حمله در دو بخش آموزش و آزمایش، تعریف‌شده باشند. روش‌های مبتنی بر امضاء نیازمند بهنگام‌سازی مکرر پایگاه داده حملات در فواصل زمانی کوتاه هستند. به‌علاوه حملاتی که از روش‌های رمزنگاری داده‌ها استفاده می‌نمایند توسط این روش به‌سختی قابل‌تشخیص هستند [۴]. در سامانه‌های تشخیص نفوذ مبتنی بر امضا، جهت تحلیل الگوهای نفوذ و ساخت مدل تشخیص، از فنون داده‌کاوی یا تحلیل آماری استفاده می‌شود. قبل از به‌کارگیری فنون داده‌کاوی، لازم است آماده‌سازی داده‌ها انجام گیرد.

یکی از مهم‌ترین مراحل آماده‌سازی داده‌های سامانه تشخیص نفوذ، تعیین خصایص یا ویژگی‌هایی از داده‌ها است که بر اساس مقادیر آن خصایص، بتوان مراحل داده‌کاوی را به‌خوبی دنبال کرد. در همین راستا فرایندی موسوم به انتخاب یا کاهش خصیصه نیز وجود دارد که منجر به انتخاب مطلوب خصایص، متناسب با هدف مورد نظر می‌گردد. هرچه انتخاب خصایص دقیق‌تر باشد؛ تشخیص نفوذ نیز دقیق‌تر خواهد بود.

* رایانامه نویسنده مسئول: m.abbasi@sndu.ac.ir

1- Edward Amoroso
2- Misuse Detection

سؤال اصلی تحقیق این است که چگونه می‌توان فهرست دقیقی از خصایص یا ویژگی‌ها را از مجموعه خصایص عمومی یک سامانه تشخیص نفوذ انتخاب کرد؛ به‌گونه‌ای که دقت و سرعت سامانه تشخیص نفوذ ارتقا یابد.

بهره‌گیری از الگوریتم کلونی مورچگان می‌تواند انتخاب مناسبی برای تحقق این هدف باشد. در این مقاله با محوریت الگوریتم کلونی مورچگان، الگویی جهت گزینش خصایص سامانه تشخیص نفوذ، با هدف دسته‌بندی دقیق‌تر هشدارهای این سامانه، ارائه شده است.

در ادامه پس از مقدمه، الگوریتم کلونی مورچگان تبیین شده و انواع معیارهای دسته‌بندی داده‌ها توضیح داده می‌شود. بررسی کارهای مرتبط و نتایج آن در ادامه کار آورده شده، سپس ایده اصلی تحقیق مطرح گردیده و در پایان به ارزیابی طرح پیشنهادی و نتایج طرح، پرداخته می‌شود.

۱-۱- الگوریتم کلونی مورچگان

بر اساس تحقیقات آقای دوریگو[1] [۵] الگوریتم کلونی مورچگان الهام گرفته‌شده از مطالعات و مشاهدات روی کلونی مورچه‌ها جهت یافتن غذا است. مورچه‌ها در طول مسیر حرکت ردّی از فرومون[2] به‌جا می‌گذارند و چنانچه مورچه‌ای به غذا برسد در مسیر بازگشت رَد فرومون خود را تقویت می‌کند. مورچه‌های دیگر وقتی به این مسیر برخورد می‌کنند، پرسه زدن را رها کرده و مسیر جدیدی که دارای فرومون بیش‌تری است، دنبال می‌کنند و با تقویت مداوم آن مسیر و تبخیر ردهای دیگر، به‌مرور همه‌ی مورچه‌ها، هم مسیر می‌شوند.

برای پیاده‌سازی الگوریتم کلونی مورچگان ابتدا مسئله را در قالب یک گراف مدل می‌کنند؛ گره‌های این گراف موقعیت‌های مختلف و وزن یال‌ها نشان‌دهنده هزینه عبور از موقعیت‌های دو طرف یال است. گراف یادشده عموماً دارای گره‌های زیاد و ساختار پیچیده است به‌گونه‌ای که مرتبه زمانی جستجوی کامل این گراف جهت یافتن کوتاه‌ترین مسیر، با توجه به محدودیت‌های زمانی یا پردازشی، معمولاً غیرقابل‌پذیرش است. لذا با استفاده از روش‌های ابتکاری سعی می‌شود با پذیرش محدودیت‌های موجود، بهترین و کوتاه‌ترین مسیر ممکن، شناسایی شود. اولین الگوریتم کلونی مورچگان معروف به ACO در شکل (۱) ارائه شده است. در این الگوریتم احتمال این‌که مورچه k در زمان t از گره i به گره j حرکت کند با رابطه (۱) محاسبه می‌گردد.

$$P_{ij}^{k}(t)=\begin{cases}\dfrac{[\tau_{ij}]^{\alpha}.[\eta_{ij}]^{\beta}}{\sum_{l\epsilon N_i^k}[\tau_{il}]^{\alpha}*[\eta_{il}]^{\beta}} & if\ \ j\epsilon N_i^k \\ 0 & otherwise\end{cases} \qquad (۱)$$

در رابطه (۱)، τ_{ij} مقدار فرومون و η_{ij} مقدار مؤلفه ابتکاری یال ij است که گویای جذابیت انتخاب این مسیر هست. هم‌چنین α مؤلفه میزان تأثیر اثر فرومون و β مؤلفه میزان تأثیر مقدار ابتکاری است. N_i^k به ازای مورچه k، مجموعه گره‌های همسایه گره i است که قبلاً ملاقات نشده‌اند.

در زمان t از میان گره‌هایی که مورچه k می‌تواند انتخاب کند، گره‌ای که احتمال پذیرش آن بر اساس رابطه (۱) بیش‌تر است، انتخاب شده و در آرایه S_k قرار گرفته و پس از هر دور کامل(یا یک گشت) که مورچه‌ها طی می‌کنند، مقدار فرومون در تکرار t+1 طبق رابطه (۲) بهنگام می‌گردد.

$$\tau_{ij}(t+1)=(1-\rho).\tau_{ij}(t)+\sum_{k=1}^{m}\Delta\tau_{ij}^{k}(t) \qquad (۲)$$

در رابطه (۲) ρ نرخ تبخیر اثر فرومون است که $0 \le \rho \le 1$. هم‌چنین m تعداد مورچه‌ها و $\Delta\tau_{ij}^{k}$ مقدار فرومونی است که مورچه k روی یال‌ij به‌جا می‌گذارد و معمولاً با رابطه (۳) تعریف می‌شود.

$$\Delta\tau_{ij}^{k}(t)=\begin{cases}\dfrac{Q}{|S_k(t)|} & if\ ij\ visited\ by\ ant_k \\ 0 & otherwise\end{cases} \qquad (۳)$$

در رابطه (۳)، $|S_k(t)|$ طول مسیری است که k امین مورچه در آخرین دور طی کرده است. به این معنا که اگر مورچه‌ای در آخرین دور موفق شده باشد مسیر کوتاه‌تری را طی نماید فرومون بیش‌تری را روی این یال به‌جا می‌گذارد. هم‌چنین Q تابع کیفیت است که در یک حالت ساده می‌تواند برابر یک باشد.

۱. شروع
۲. مقداردهی اولیه $n,m,T,\alpha,\beta,\rho,\tau 0$
۳. مقداردهی $\tau_{ij}=\tau_0$
۴. قرار دادن m مورچه بر روی گره‌های گراف
۵. به ازای مورچه $k=1..m, S_k=\{\}$
۶. $h=1$
۷. مراحل زیر را تا زمانی که $h<T$ تکرار کن
۸. $t=1$
۹. تا زمانی که شرط توقف ارضاء نشده و $t<n$ مراحل زیر را به ازای مورچه $k=1$ تا $k=m$ تکرار کن
۱۰. η_{ij} را برای یال‌هایی که از گره i شروع می‌شوند محاسبه کن
۱۱. به ازای مورچه kام احتمال انتخاب گره بعدی را مطابق تابع احتمال(۱-۱) بررسی کرده و گره انتخابی را به مجموعه S_k اضافه کن.
۱۲. $t=t+1$, $h=h+1$
۱۳. از میان آرایهS هر مورچه، آرایه‌ای که در این دور طول مسیر کمتری دارد را به‌عنوان کوتاه‌ترین مسیر ذخیره کن.
۱۴. τ_{ij} را برای تمام یال‌ها با استفاده از رابطه(۲-۲) بهنگام کن
۱۵. کوتاه‌ترین مسیری که در h دور حاصل‌شده را به‌عنوان جواب نهایی اعلام کن
۱۶.پایان

شکل (۱): الگوریتم کلونی مورچگان [۶]

1- Marco Dorigo
2- Pheromone

۱-۲- مجموعه داده

مهم‌ترین مجموعه داده‌ای که محققان این حوزه جهت ساخت الگوی خود استفاده می‌کنند مجموعه داده KDD99 است که توسط آزمایشگاه MIT Lincoln ارائه‌شده و شامل ۴۸۹۸۴۳۱ اتصال آموزش و ۳۱۱۰۲۹ اتصال آزمایش هست. در این مجموعه داده، هر ردیف نماینده یک ثبت اتصال نرمال و یا یکی از حملاتی است که در جدول (۱) آمده است [۷].

در سال ۲۰۰۹ مجموعه داده NSL-KDD توسط آقای محمود تولایی و همکارانش [۸] برای رفع مشکلات KDD ارائه شد. در این مجموعه داده که به لحاظ تعداد ردیف تقریباً ۱۰ درصد مجموعه داده KDD99 است، توزیع حملات اصلاح‌شده و افزونگی داده‌های مجموعه داده اولیه نیز برطرف شده است.

جدول (۱): انواع حملات بر اساس دسته‌بندی KDD99 [۹]

حملات منع سرویس(DoS)	Apache2,Back,Land,Mail-Bomb,Neptune,Pod,Process-Table,Smurf,Teardrop,UDP-Storm,Worm
تفحص (Probe)	IP-Sweep,Mscan,Nmap,Port-Sweep,Saint,Satan
دسترسی راه دور به محلی (R2L)	FTP-Write,Guess-Password,HTTP-Tunnel,Imap,Multihop,Named,PHF,Sendmail,Snmp-Getattack,Snmp-Guess,Spy,Warez-Client,Warez-Master,Xlock,Xsnoop
دسترسی کاربر به ریشه(U2R)	Buffer-Overflow,Load-Module,Perl,Ps,Rootkit,SQL-Attack,Xterm

در جدول (۲) توزیع انواع حملات در ۲ مجموعه داده KDD99, NSL-KDD مقایسه شده است.

جدول (۲): مقایسه مجموعه داده KDD99,NSL-KDD [۹]

مجموعه داده	تعداد نمونه	Normal	DoS	U2R	R2L	Probe
KDD99	Train۴۹۴۰۲۱۰	۹۷۲۷۸۰	۳۹۱۴۵۸۰	۵۴۰	۱۱۲۴۰	۴۱۷۰
	Test۳۱۱۰۲۹	۶۵۹۳	۲۲۹۸۵۳	۲۶۳۶	۱۳۷۸۱	۴۱۶۶
NSL-KDD	Train۱۲۵۹۷۳	۶۷۳۴۳	۴۵۹۲۷	۵۲	۹۹۵	۱۱۶۵۶
		۵۳,۴۶%	۳۶,۴۶%	۰,۰۴%	۰,۷۹%	۹,۲۵%
	Test۲۲۵۴۴	۹۷۱۱	۷۴۵۸	۲۰۰	۲۷۵۴	۲۴۲۱
		۴۳,۰۸%	۳۳,۰۸%	۰,۸۹%	۱۲,۲۲%	۱۰,۷۴%

در مجموعه داده NSL-KDD ۴ جدول داده در دو قالب csv,arff ارائه‌شده که فهرست آن‌ها در جدول (۳) آمده است.

جدول (۳): جداول ارائه‌شده در مجموعه داده NSL-KDD [۹]

نام جدول	تعداد نمونه	محتوا
KDDTrain+	۱۲۵۹۷۳	شامل تمام نمونه‌های آموزشی
KDDTrain+_20Percent	۲۵۱۹۲	شامل ۲۰ درصد اول نمونه‌های آموزشی
KDDTest+	۲۲۵۴۴	شامل تمام نمونه‌های آزمایشی
KDDTest-21	۱۱۸۵۰	شامل ۵۲ درصد تصادفی از نمونه-های آزمایشی

۱-۳- معیارهای دسته‌بندی داده

چهار معیار دسته‌بندی پایه به نام‌های TP,FP,TN,FN طبق جدول (۴) تعریف می‌گردند که در آن True به معنای تشخیص صحیح و False به معنای تشخیص غلط است. هم‌چنین P به معنای پیش‌بینی مثبت شدن و N به معنای پیش‌بینی منفی شدن است.

مهم‌ترین معیارهای دسته‌بندی داده در جدول (۵) آمده است که در آن N=TN+FN و P=TP+FP لذا P+N شامل همه نمونه‌ها است.

جدول (۴): معیارهای دسته‌بندی پایه [۱۰]

	پیش‌بینی مثبت بودن	پیش‌بینی منفی بودن
مثبت محاسبه شدن	P*P=>TP	P*N=>FN
منفی محاسبه شدن	N*P=>FP	N*N=>TN

جدول (۵): معیارهای دسته‌بندی داده [۱۰]

نام معیار	فرمول محاسبه
TP Rate, Sensitivity, Recall, Hit Rate	TP/P=TP/(TP+FN)
TN Rate , Specificity	TN/N=TN/(FP+TN)
Precision, PPV(Positive Predictive Value)	TP/(TP+FP)
NPV (Negative Predictive Value)	TN/(TN+FN)
Fall-Out , FPR(False Positive Rate)	FP/N = FP/(FP+TN) = 1-TNR
FDR(False Discovery Rate)	FP/(FP+TP) = 1-PPV
Miss Rate , FNR(False Negative Rate)	FN/P = FN/(FN+TP)=1-TRP
ACC(Accuracy)	(TP+TN)/(P+N)

۲- کارهای مرتبط

بررسی تحقیقات انجام‌شده نشان می‌دهد که دو رویکرد عمده در گزینش خصایص وجود دارد. در رویکرد اول به موضوع انتخاب خصایص به‌عنوان یکی از سرفصل‌های داده‌کاوی توجه شده و با رویکردی مبتنی بر مبانی ریاضی، روشی جهت انتخاب خصایص مؤثرتر معرفی شده است. در رویکرد دوم، موضوع انتخاب خصایص، منحصراً بر مبنای کاهش یا گزینش خصایص هشدارهای سامانه تشخیص نفوذ صورت پذیرفته است. رویکرد دوم معمولاً به پاسخ‌های دقیق‌تری می‌رسد. در ادامه مقالات رویکرد دوم بررسی می‌گردد.

گائو[1] [۱۱] با استفاده از الگوریتم بهینه‌سازی کلونی مورچگان خصایص مجموعه داده KDD99 را کاهش داده و برای دسته‌بندی داده‌ها از الگوریتم SVM و تابع تخمین LS استفاده کرده است. وی برای مؤلفه ابتکاری الگوریتم کلونی مورچگان طبق رابطه (۴) از ضریب FDR[2] استفاده کرده که در آن X_{ij} وزن یال ij است.

$$\eta_{ij} = \frac{\sum_{n=1}^{N} x_{ni} x_{nj}}{\sqrt{\sum_{n=1}^{N} x_{ni}^2 \sum_{n=1}^{N} x_{nj}^2}} \quad (۴)$$

زرگری [۱۲] ابتدا اصلاحاتی بر مجموعه داده KDD99 انجام داده و تعداد نمونه‌های آن را به ۳۱۱۰۲۹ نمونه کاهش داده است. سپس ۵ مجموعه داده کوچک‌تر به تعداد ۱۵۶۰۴ و ۱۱۷۰۷ و ۷۸۰۹ و ۳۹۰۵ و ۱۳۷۲ از آن استخراج نموده و خصایص مجموعه داده‌های جدید را با استفاده از دو روش CfsSubset, InfoGainVal کاهش داده و درنهایت با استفاده از درخت تصمیم Random Forest نمونه‌های مرتبط با خصایص تعیین‌شده را دسته‌بندی کرده است. اَذَب[3] [۱۳] با استفاده از آنتروپی ابتدا بهره اطلاعات هر خصیصه را تعیین کرده است و بر این اساس ۱۲ خصیصه‌ای که دارای بالاترین بهره اطلاعات هستند را به‌عنوان خصایص منتخب، برگزیده است. سپس با استفاده از درخت تصمیم J48 اقدام به دسته‌بندی مجموعه داده NSL-KDD نموده است. ژانگ[4] [۱۴] از الگوریتم BayesNet برای انتخاب خصایص مجموعه داده NSL-KDD استفاده کرده است. در این روش ابتدا تمام خصایص با الگوریتم انتخابی آموزش داده می‌شود. سپس در یک حلقه تکرار هر بار، یکی از خصایص حذف‌شده و چنانچه حذف این خصیصه موجب کاهش نرخ صحت[5] گردد، این خصیصه به خصایص انتخابی اضافه می‌گردد. پارک [۱۵] در مجموعه داده NSL-KDD از متوسط مقادیر هر ستون و مقایسه آن با متوسط مقادیر ستون رده، استفاده کرده است. درنهایت ستون‌هایی که به متوسط مقادیر رده نزدیک‌تر هستند را به‌عنوان خصایص منتخب برگزیده است.

پنج محقق مصری [۱۶] بر مبنای مجموعه داده NSL-KDD ابتدا با شش روش مختلف به نام‌های PCA, SFFS, SFBS, CFS, IG,RS خصایص هشدارها را گزینش کرده و سپس با استفاده از الگوریتم ژنتیک هشدارها را دسته‌بندی کرده‌اند. تِسفهون[6] [۱۷] در مجموعه داده NSL-KDD از روش IG برای انتخاب خصایص و از درخت تصمیم Random Forest برای دسته‌بندی هشدارها استفاده کرده است اما قبل از دسته‌بندی، در بخش پیش‌پردازش با روش SMOTE[7] توزیع هشدارها را بر اساس رده، یکنواخت کرده و یا به عبارتی نمونه هشدارها را بازنمونه‌گذاری[8] کرده است. طباخی [۱۸] با استفاده از الگوریتم کلونی مورچگان یک روش غیر نظارتی جهت انتخاب خصیصه معرفی کرده است. وی در این روش هر خصیصه را یک گره گراف فرض کرده و وزن یال‌های این گراف را بر اساس شباهت بین گره‌های دو سر هر یال محاسبه کرده است. تابع محاسبه شباهت مطابق رابطه (۵) محاسبه شده که در آن p تعداد نمونه‌هاست.

$$Sim(A,B) = \left| \frac{\sum_{i=1}^{p}(a_i b_i)}{(\sqrt{\sum_{i=1}^{p} a_i^2})(\sqrt{\sum_{i=1}^{p} b_i^2})} \right| \quad (۵)$$

در روش غیرنظارتی، ستون رده وجود ندارد. تابع ابتکاری در الگوریتم کلونی مورچگان همان تابع شباهت رابطه (۵) است و گره‌ای که در هر مرحله انتخاب شود فرومون بیشتری به آن اختصاص می‌یابد و درنهایت گره‌هایی که دارای فرومون بیش‌تری باشند به‌عنوان خصایص برتر انتخاب می‌شوند.

امبوسعیدی[9] [۱۹] یک روش خصیصه گزینی مبتنی بر روش فیلتر پیشنهاد کرده و نام آن را FMIFS[10] گذاشته است. در این روش تابعی بر مبنای آنتروپی برای سنجش اهمیت هر خصیصه به‌صورت رابطه (۶) در نظر گرفته‌شده است.

$$G_{MI} = argmax_{f_i \epsilon F}(I(C;f_i) - \frac{1}{|S|}\sum_{f_i \epsilon F} MR) \quad (۶)$$

در رابطه (۶)، C برچسب ستون رده، S مجموعه خصایص و f_i خصیصه جاری است. سایر مؤلفه‌های آن طبق روابط (۷)، (۸)، (۹)، (۱۰) و (۱۱) محاسبه می‌گردد.

$$MR = \frac{I(f_i;f_s)}{I(C,f_i)} \quad (۷)$$

$$I(U;V)=H(U)+H(V)-H(U,V) \quad (۸)$$

$$H(U) = -\int_u p(u) \log p(u) d(u) \quad (۹)$$

$$H(V) = -\int_v p(v) \log p(v) d(v) \quad (۱۰)$$

$$I(U;V) = \int_u \int_v p(u,v) \log \frac{p(u,v)}{p(u)p(v)} dudv \quad (۱۱)$$

در روابط (۷) تا (۱۱)، توابع H(U) و H(V) درواقع، توابع آنتروپی هستند. آقدم [۹] با استفاده از الگوریتم کلونی مورچگان خصیصه گزینی هشدارهای تشخیص نفوذ را انجام داده است. وی نیز همچون تحقیقات مشابه، فهرست خصایص را با گره‌های گراف، متناظر کرده است. در این روش هر مورچه به‌صورت تصادفی از یک گره حرکت خود را شروع می‌کند و شرط پایان حرکت او در هر گشت این است که نرخ شناسایی[11] در سیر افزایشی خود، شروع به کاهش کند. وی همچنین این روش را به

1- Hai-Hua Gao
2- Fisher Discrimination Rate
3- Ammar Alazab
4- Fengli Zhang
5- Accuracy
6- Abebe Tesfahun
7- Synthetic Minority Oversampling Technique
8- Resampling
9- Mohammed Ambusaidi
10- Flexible Mutual Information Based Feature Selection
11- Detection Rate (DR)

ازای هر یک از مقادیر ستون رده (یعنی حالت نرمال و انواع حملات) جداگانه تکرار کرده و لذا ۵ زنجیره خصایص را به ازای هر مقدار رده به دست آورده است.

۳- معرفی طرح پیشنهادی

در طرح پیشنهادی، با استفاده از فنون داده‌کاوی، الگویی از یک سامانه تشخیص نفوذ مبتنی بر تشخیص سوءاستفاده ارائه‌شده که در آن با بهره‌گیری از الگوریتم کلونی مورچگان، نرخ صحت تشخیص نوع حمله ارتقا یافته است. ازآنجاکه در طرح پیشنهادی، الگوی ابتکاری، نوعی انتخاب خصیصه با استفاده از الگوریتم کلونی مورچگان است، لذا نام این طرح ACFSM[1] انتخاب گردید.

۳-۱- مراحل کلی طرح

مراحل کلی آماده‌سازی سامانه تشخیص نفوذ در طرح ACFSM مطابق شکل (۲) است.

شکل (۲): مراحل کلی آماده‌سازی سامانه تشخیص نفوذ در طرح ACFSM

در ابتدا مجموعه داده اتصالات، انتخاب‌شده و پس از پیش‌پردازش، فهرست عناوین خصایص قابل بهره‌برداری از آن تعیین می‌گردد. در ادامه، بردار خصایص که درواقع جدول مقادیر خصایص است، تکمیل‌شده و پس از پیش‌پردازش، بخشی از آن به‌عنوان داده آموزشی و بخش دیگر به‌عنوان داده آزمایشی مورد بهره‌برداری قرار می‌گیرد. در ادامه هر یک از مراحل فوق به همراه جزئیات بیش‌تری از طرح ACFSM ارائه می‌گردد.

انتخاب مجموعه داده: در این طرح، مجموعه داده NSL-KDD به‌عنوان مجموعه داده مبنا انتخاب‌شده و از داده‌های آن جهت آموزش و آزمایش سامانه تشخیص نفوذ استفاده می‌شود.

استخراج اتصالات از پشته: عموماً مجموعه داده‌هایی که محققان از آن استفاده می‌کنند در قالب پشته تی.سی.پی در فضای اینترنت ارائه می‌گردد. پشته تی.سی.پی درواقع انباشته‌ای از ترافیک اتصالات است که در یک فایل با فرمت دودویی نگهداری می‌گردد. در این مرحله فایل پشته تی.سی.پی به‌عنوان ورودی دریافت شده و محتوای اتصالات در قالب فایل متنی و به‌صورت جداول قابل‌فهم ذخیره می‌گردد.

پیش‌پردازش اتصالات: در مرحله پیش‌پردازش داده‌ها، به مسائلی همچون حذف داده‌های اضافی، جبران داده‌های مفقودی و همسان‌سازی فرمت‌ها، پرداخته می‌شود.

تعیین فهرست خصایص: جهت داده‌کاویِ یک مجموعه داده، باید خصایصی از این مجموعه شناسایی شود که تحلیل‌ها یا اعمال الگوریتم‌های متنوع داده‌کاوی بر داده‌های برخاسته از این خصایص انجام شود. این خصایص باید کاملاً شفاف و قابل استخراج از مجموعه داده باشند.

در مجموعه داده NSL-KDD برای هر اتصال ۴۱ خصیصه تعیین شده است که در جدول (۱۷) این خصایص فهرست شده‌اند. این خصایص به ۴ گروه خصایص پایه، خصایص مربوط به محتوای بسته، خصایص ترافیکی وابسته به زمان و خصایص ترافیکی وابسته به میزبان تقسیم شده‌اند. از این خصایص، خصیصه شماره ۲۰ (یعنی Num-outbound-cmds) به ازای تمام ردیف‌های آموزشی و آزمایشی صفر است، لذا تأثیری در دسته‌بندی نداشته و حذف می‌گردد. از طرفی برچسب رده به‌عنوان خصیصه نهایی به این خصایص اضافه می‌شود که دارای یکی از ۵ مقدار normal,dos,probe,r2l,u2r است. این برچسب نشان‌دهنده نتیجه اتصال بوده و گویای نوع حمله یا وضعیت نرمال است.

وجود یا عدم وجود خصیصه رده تأثیر زیادی در انتخاب فن داده‌کاوی دارد. اگر رده یک اتصال در مرحله آموزش نامشخص باشد در این صورت، کار بسیار مشکل شده و باید از فنون خوشه‌یابی یا رگرسیون استفاده کرد تا اتصالات بر اساس سایر مؤلفه‌های قابل پیش‌بینی، از هم تفکیک گردند. اما اگر نتیجه اتصالات در مرحله آموزش مشخص باشد در این صورت کار ساده‌تر است و معمولاً از فنون دسته‌بندی استفاده می‌شود تا اتصالات بر اساس شاخص رده در دسته‌های جداگانه توزیع گردند. ازآنجاکه در مجموعه داده NSL-KDD نتیجه هر ثبت اتصال مشخص شده است؛ لذا در طرح ACFSM از دسته‌بندی جهت تفکیک اتصالات بر مبنای شاخص رده استفاده می‌شود. با افزودن خصیصه رده به خصایص قبل در نهایت تعداد خصایصی که بر اساس آن دسته‌بندی انجام می‌شود ۴۱ خصیصه است.

1- Ant Colony Featur Selection Method

ساخت بردار خصایص: پس از مشخص شدن خصایصی که به ازای هر اتصال قابل استخراج است، در این مرحله عملاً مقادیر مربوط به خصایص هر یک از اتصالات، استخراج شده و فایل داده جدیدی که آن را بردار خصایص یا فایل نمونه داده‌ها می‌نامند، تشکیل می‌شود. در مجموعه داده NSL-KDD درمجموع ۱۴۸۵۱۷ نمونه یا ردیف وجود دارد که ۱۲۵۹۷۳ نمونه جهت آموزش و ۲۲۵۴۴ نمونه جهت آزمایش در فایل‌هایی بنام Train و Test ذخیره‌شده است.

پیش‌پردازش بردار خصایص: بردار خصایص یا همان نمونه‌ها در فایل‌های متنی با فرمت csv,arff ارائه شده‌اند. جهت سهولت انجام محاسبات عددی اطلاعات فایل‌های Train.txt,Test.txt به جداول معادل آن در بانک اطلاعاتی SQL-Server منتقل گردید. در ستون آخر مجموعه داده‌های آموزشی و آزمایشی Train,Test به‌جای نام دسته حمله که درواقع باید خصیصه رده باشد، نام حمله آمده است. لذا طبق جدول (۱)، نام حمله را در جدول‌های پایگاه داده، با نام دسته حمله جایگزین می‌شود.

ازآنجاکه برنامه‌های محاسباتی، محاسبات مقادیر عددی را سریع‌تر از محاسبات مقادیر متنی انجام می‌دهند و با توجه به اهمیت استفاده بهینه از زمان در مرحله آموزش و آزمایش، مقادیر متنی به مقادیر عددی تبدیل شد. لذا در این مرحله خصایص ستون ۲، ۳ و۴ مطابق جدول (۶) به خصایص عددی تبدیل شد.

در خصیصه Protocol-type ۳ مقدار متنی و در خصیصه Service-Network ۷۰ مقدار متنی و در خصیصه Flag ۱۱ مقدار متنی وجود دارد که طبق جدول (۶)، در پایگاه داده SQL-Server مقادیر متنی با معادل عددی آن، جایگزین شد.

جدول (۶): تبدیل مقادیر غیر عددی NSL-KDD به مقادیر عددی

۲	Protocol-type	TCP=۱ , UDP=۲ , ICMP=۳
۳	Service Network	نوع سرویس شامل یکی از موارد زیر: Aol=۱, auth=۲, bgp=۳, courier=۴, csnet_ns=۵, ctf=۶, daytime=۷, discard=۸, domain=۹, domain_u=۱۰, echo=۱۱, eco_i=۱۲, ecr_i=۱۳, efs=۱۴, exec=۱۵, finger=۱۶, ftp=۱۷, ftp_data=۱۸, gopher=۱۹, harvest=۲۰, hostnames=۲۱, http=۲۲, http_2784=۲۳, http_443=۲۴, http_8001=۲۵, imap4=۲۶, IRC=۲۷, iso_tsap=۲۸, klogin=۲۹, kshell=۳۰, ldap=۳۱, link=۳۲, login=۳۳, mtp=۳۴, name=۳۵, netbios_dgm=۳۶, netbios_ns=۳۷, netbios_ssn=۳۸, netstat=۳۹, nnsp=۴۰, nntp=۴۱, ntp_u=۴۲, pm_dump=۴۳, pop_2=۴۴, pop_3=۴۵, printer=۴۶, private=۴۷, red_i=۴۸, remote_job=۴۹, rje=۵۰, shell=۵۱, smtp=۵۲, sql_net=۵۳, ssh=۵۴, sunrpc=۵۵, supdup=۵۶, systat=۵۷, telnet=۵۸, tftp_u=۵۹, tim_i=۶۰, time=۶۱, urh_i=۶۲, urp_i=۶۳, uucp=۶۴, uucp_path=۶۵, vmnet=۶۶, whois=۶۷, X11=۶۸, Z39_50=۶۹, other=۷۰
۴	Flag	SH=۱, SF=۲, S3=۳, S2=۴, S1=۵, S0=۶, RSTR=۷, RSTOS0=۸, RSTO=۹, REJ=۱۰, OTH=۱۱

1- Attack Category

گزینش خصایص: در مرحله گزینش خصایص، در صورت امکان بخشی از خصایص، که نقش منفی یا غیرمفید در فرآیند داده‌کاوی دارند حذف می‌شوند. برای این منظور دو روش کلی وجود دارد. در روش اول بدون توجه به الگوریتم داده‌کاوی، خصایص انتخاب می‌شوند در روش دوم، ابتدا الگوریتم داده‌کاوی تعیین می‌گردد تا متناسب با نوع عملکرد آن الگوریتم، فهرست خصایص اصلاح گردند. خصایص انتخاب‌شده در روش دوم معمولاً خصایص مؤثرتری هستند و در طرح ACFSM نیز از این روش استفاده می‌گردد.

در طرح ACFSM، قبل از تصمیم‌گیری در خصوص تغییر خصایص، از انواع روش‌های دسته‌بندی، یک الگوریتم نمونه انتخاب‌شده و توان آن را در دسته‌بندی مجموعه داده آزموده شد. در این مرحله الگوریتم‌های زیر را انتخاب شدند:

۱) از روش‌های دسته‌بندی مبتنی بر ساخت درخت تصمیم، الگوریتم J48 انتخاب شد که تعمیم‌یافته درخت تصمیم C4.5 است.
۲) از روش‌های دسته‌بندی مبتنی بر تحلیل آماری، الگوریتم BayesNet انتخاب شد.
۳) از روش‌های دسته‌بندی مبتنی بر شبکه‌های عصبی مصنوعی، الگوریتم MLP انتخاب شد.
۴) از روش‌های دسته‌بندی مبتنی بر رگرسیون الگوریتم LogitBoost انتخاب شد.
۵) از روش‌های دسته‌بندی مبتنی بر تحلیل تلازمی، الگوریتم PART انتخاب شد.

در مرحله آموزش و آزمایش، مقادیر ستون رده که در این تحقیق نوع حمله است باید مشخص باشد. از مقادیر ستون رده در مرحله آموزش جهت ساخت مدل استفاده می‌شود و در مرحله آزمایش جهت بررسی نرخ صحت دسته‌بندی استفاده می‌گردد.

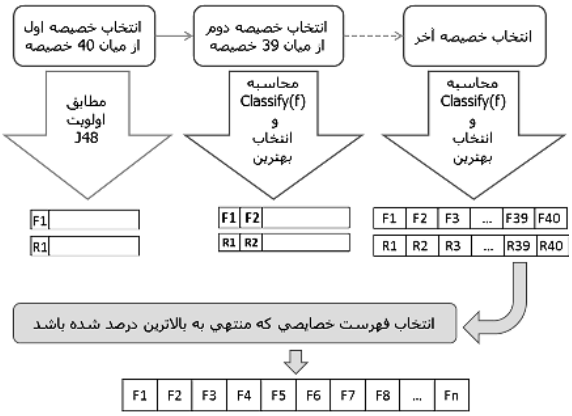
نتایج حاصل از دسته‌بندی مجموعه داده NSL-KDD در مرحله آموزش و آزمایش با نرم‌افزار وکا و با مقادیر اولیه پیش‌فرض، بر اساس نرخ صحت در جدول (۷) آمده است. زمان‌های ثبت‌شده در این جدول در رایانه‌ای با پردازنده Intel-Core-i7 و حافظه DDR3-8GB حاصل‌شده است.

جدول (۷): نتایج دسته‌بندی در مرحله آموزش و آزمایش

الگوریتم	زمان ساخت مدل(s)	دسته‌بندی در آموزش	زمان آزمایش (s)	پیش‌بینی در آزمایش
BayesNet	**۵/۹۲**	۹۵/۷۳۸۰%	۰/۴۵	۷۳/۳۱۳۲%
J48	۳۲/۵۷	۹۹/۹۱۴۳%	۰/۳۴	۷۵/۱۵۴۱%
MLP	۱۱۶۹/۱۳	۹۸/۶۱۸۰%	۰/۴۲	۷۲/۷۱۸۸%
PART	۳۷/۹۱	**۹۹/۹۴۸۴%**	**۰/۳۱**	**۷۶/۲۲۷۷%**
LogitBoost	۶۲/۹۳	۹۸/۴۳۷۸%	۰/۵۵	۷۴/۰۴۵۲%

این نتایج نشان می‌دهد که الگوریتم PART بهتر از سایر الگوریتم‌ها، مجموعه داده را در مرحله آموزش دسته‌بندی کرده است. هم‌چنین مدت‌زمان آزمایش برای ۲۲۵۴۳ رکورد، ۰٫۳۱s بوده که بهتر از سایر الگوریتم‌ها است. مهم‌تر از همه قدرت بیش‌تر این الگوریتم در پیش‌بینی نوع حمله است که حدود ٪۷۶/۲۳ است.

در این مرحله، الگوریتم PART به‌عنوان الگوریتم منتخب انتخاب شد اما هدف ارائه روشی است که نرخ صحت دسته‌بندی را از ٪۷۶/۲۳ کنونی تا حد امکان افزایش دهد.



شکل (۳): مراحل طی شده جهت انتخاب خصایص

طرح پیشنهادی، انتخاب زیرمجموعه‌ای از خصایص موجود است، به‌گونه‌ای که نرخ صحت دسته‌بندی الگوریتم منتخب با استفاده از این زیرمجموعه خصایص تا حد امکان افزایش یابد. برای این منظور همان‌گونه که در شکل (۳) نشان داده‌شده، ابتدا یکی از خصایص ۴۰ گانه انتخاب می‌شود و سپس در ۳۹ تکرار، هر بار به ازای مجموعه این خصیصه و یکی از ۳۹ خصیصه باقی‌مانده، مدل دسته‌بندی ساخته‌شده و نرخ صحت پیش‌بینی برای این مجموعه ۲ عضوی محاسبه و در آرایه‌ای ذخیره می‌شود. در ادامه مجموعه‌ای که منجر به بهترین نرخ صحت دسته‌بندی شود برگزیده شده و سپس از بین ۳۸ خصیصه باقی‌مانده در ۳۸ تکرار، هر بار به ازای مجموعه برگزیده ۲ عضوی قبل و یکی از ۳۸ خصیصه باقی‌مانده مدل دسته‌بندی ساخته‌شده و مجدداً نرخ صحت پیش‌بینی محاسبه می‌شود. خصیصه سوم به‌گونه‌ای انتخاب می‌شود که در کنار دو خصیصه قبلی بهترین نتیجه حاصل شود. این مراحل تا انتخاب خصیصه چهلم ادامه می‌یابد. پس از پایان دور اول، دور دوم شروع می‌شود و این بار یکی دیگر از خصایص به‌عنوان خصیصه اول انتخاب‌شده و فرآیند فوق تا تعیین خصیصه چهلم ادامه می‌یابد. در هر بار اجرای فرآیند فوق، زنجیره‌ای از خصایص به همراه نرخ صحت دسته‌بندی در هر خصیصه حاصل می‌شود. در انتها، زنجیره‌ای که در آن بالاترین نرخ صحت دسته‌بندی حاصل شده باشد، انتخاب می‌شود. زنجیره خصایص بهینه این طرح شامل اولین خصیصه مجموعه منتخب تا خصیصه‌ای است که در آن خصیصه به بالاترین نرخ صحت دسته‌بندی رسیده باشد. مثلاً اگر در مجموعه {۱۷،۵،۲۱،۲،۱۹،۳۳،۱،۸} در خصیصه ۱۹ بالاترین نرخ صحت به‌دست آید در این صورت مجموعه {۱۷،۵،۲۱،۲،۱۹} نشان‌دهنده خصایص منتخب است.

در این بخش به محاسبه مرتبه زمانی انجام طرح پرداخته می‌شود. اگر تعداد خصایص n فرض شود، در یک جستجوی کامل، در گام اول انتخاب یک خصیصه از n خصیصه خواهد داشت و در گام دوم انتخاب دو خصیصه از n خصیصه به‌گونه‌ای که تکرار خصایص نداشته و ترتیب انتخاب نیز بی‌تأثیر باشد (یعنی اگر یکبار خصیصه ۱ و ۲ انتخاب شده، در این صورت انتخاب ترتیب خصایص ۲ و ۱ نخواهد بود) و در گام r انتخاب r خصیصه از n خصیصه را خواهد داشت به‌گونه‌ای که در r خصیصه، خصیصه تکراری نباشد و ترتیب انتخاب خصایص نیز بی‌تأثیر باشد و نهایتاً در گام nاُم تنها یک خصیصه جهت انتخاب باقی می‌ماند. لذا مرتبه زمانی اجرای این الگوریتم، حاصل جمع جایگشت‌های یک تا n از n خصیصه است که طبق رابطه (۱۲) محاسبه می‌شود.

$$O(n) = C_1^n + C_2^n + \cdots + C_r^n + \cdots + C_{n-1}^n + C_n^n \qquad (۱۲)$$

که در آن : $C_r^n = \frac{n!}{r!(n-r)!}$

مقدار تابع به ازای n=40 برابر است با: $O(40) = 1.1*10^{12}$

برای محاسبه زمان انجام طرح باید مقدار فوق در متوسط زمان صرف‌شده، به ازای هر بار اجرای الگوریتم محاسبه گردد. زمان اجرای الگوریتم برابر مجموع زمان آموزش و آزمایش است که این زمان برای الگوریتم منتخب PART مطابق جدول (۷) برابر ۳۷/۹۱+۰/۳۱ ثانیه یعنی حدود ۴۰s است. لذا زمان انجام طرح برابر است با:

$Total = O(40)*40 = 4.4*10^{13}(Second)$

با توجه به اینکه یک قرن ۳،۱۵۳،۶۰۰،۰۰۰ ثانیه است، درواقع پاسخ نهایی این طرح پس از چهارده هزار قرن مشخص می‌شود! این زمان در صورت انتخاب الگوریتم BayesNet نیز کمتر از ۲ هزار قرن نخواهد بود.

در چنین شرایطی که مرتبه زمانی جستجوی کامل بسیار بالا و غیرقابل‌پذیرش است، یک‌راه بیشتر نمی‌ماند و آن، به‌کارگیری الگوریتم‌های جستجوی ابتکاری است. این الگوریتم‌ها برخلاف الگوریتم‌های حریصانه به دنبال پیمایش کامل فضای جستجو نیستند، بلکه با پذیرش محدودیت زمان، سعی می‌کنند در روشی ابتکاری نزدیک‌ترین پاسخ ممکن که شاید متفاوت با جواب بهینه اصلی باشد را بیابند. الگوریتم‌های جستجوی ابتکاری متنوعی در حال حاضر مطرح هستند و در مسائل پیچیده و زمان‌بر از آن‌ها استفاده می‌شود که از آن جمله می‌توان به الگوریتم ژنتیک و الگوریتم کلونی مورچگان اشاره کرد. استفاده از نمونه‌های مختلف

الگوریتم‌های ابتکاری، در حل این مسئله می‌تواند موضوع یک پژوهش باشد اما ازآنجاکه به‌کارگیری الگوریتم‌های ژنتیک برای حل این مسئله بیشتر متداول بوده است، در این طرح از الگوریتم کلونی مورچگان جهت حل این مسئله استفاده شده است. البته همان‌گونه که در بخش پژوهش‌های مرتبط بررسی شد برخی محققین از الگوریتم کلونی مورچگان نیز جهت حل این مسئله استفاده کرده‌اند اما نویسندگان این مقاله بر آنند که با ابتکار عمل نتیجه بهتری کسب نمایند.

ساخت مدل دسته‌بندی خصایص(آموزش): پس از نهایی شدن زنجیره خصایص منتخب، در این مرحله، بردار خصایص مجموعه داده آموزش که شامل۱۲۵۹۷۳ردیف داده است با استفاده از الگوریتم PART و با نرم‌افزار وکا، تنها به ازای خصایصی که در مرحله قبل تعیین‌شده است، دسته‌بندی می‌گردد تا با توجه به نوع حمله، مدلی به‌دست آید که در مرحله آزمایش قادر به پیش‌بینی نوع حمله باشد.

آزمایش مدل دسته‌بندی خصایص: در این مرحله با به‌کارگیری الگوریتم منتخب، مجموعه داده آزمایشی که مشتمل بر ۲۲۵۴۳ ردیف است؛ تنها به ازای خصایص منتخبی که در مرحله کاهش خصایص تعیین شده‌اند، آزمایش می‌شود. نرخ صحت پیش‌بینی نوع حمله در این مرحله گویای میزان موفقیت طرح ACFSM در بهبود دسته‌بندی هشدارهای سامانه تشخیص نفوذ است.

۳-۲- جزئیات طرح ACFSM

در طرح کلی شناسایی نوع حمله، از دسته‌بندی ۴۱ خصیصه استفاده گردید. تعداد بالای خصایص موجب بروز دو چالش خواهد شد:

۱) پیچیده شدن مدل پیش‌بینی و در نتیجه کاهش نرخ صحت تشخیص،

۲) افزایش زمان تشخیص نوع حمله.

جهت کاهش تعداد خصایص، در بخش قبل ایده‌ای مطرح شد که بنا بر تحلیل زمانی صورت گرفته، مشخص شد که پیاده‌سازی این ایده با روش‌های معمول سال‌ها طول می‌کشد. در این بخش طرحی ارائه می‌شود که با استفاده از الگوریتم ابتکاری کلونی مورچگان، فضای جستجو جهت یافتن خصایص موردنظر با در نظر گرفتن محدودیت‌های زمانی پیمایش شود.

در طرح ACFSM که فرآیند آن در شکل (۵) آمده است، برای این منظور ۴۰ خصیصه اصلی (بدون ستون رده) گره‌های یک گراف کامل فرض می‌شود. سپس تعداد m مورچه (که m<n) در گره‌های مختلف این گراف قرار می‌گیرد. در روش‌های معمول این گره‌ها به‌صورت تصادفی انتخاب می‌شوند. اما در اینجا از نتایج دسته‌بندی درخت تصمیم J48 استفاده می‌شود زیرا در درخت تصمیم، گره‌هایی که در سرشاخه بالاتری قرار می‌گیرند، قدرت تفکیک بالاتری دارند [۲۰]. لذا نقطه شروعِ بهتری برای پیمایش یک مورچه هستند. در شکل(۴) بخشی از مدل دسته‌بندی درخت تصمیم J48 آمده است. در این مثال خصیصه‌های src_bytes,counts در سرشاخه بالاتری قرار داشته و لذا گزینه بهتری برای شروع پیمایش هستند.

```
17.  src_bytes <= 28
18.  |  counts <= 3
19.  |  |  dst_host_same_src_port_rate <= 0.5
20.  |  |  |  dst_host_serror_rate <= 0.89
21.  |  |  |  |  dst_host_srv_count <= 2
22.  |  |  |  |  |  dst_host_rerror_rate <= 0.02
23.  |  |  |  |  |  |  wrong_fragment <= 0
24.  |  |  |  |  |  |  |  flag_cost <= 1: probe (6.0)
25.  |  |  |  |  |  |  |  flag_cost > 1
26.  |  |  |  |  |  |  |  |  dst_host_srv_serror_rate <= 0.75
27.  |  |  |  |  |  |  |  |  |  num_compromised <= 0
28.  |  |  |  |  |  |  |  |  |  |  dst_bytes <= 1
29.  |  |  |  |  |  |  |  |  |  |  |  serror_rate <= 0.75
30.  |  |  |  |  |  |  |  |  |  |  |  |  flag_cost <= 8
```

شکل (۴): بخشی از درخت تصمیم J48 تولیدشده توسط وکا

بر این مبنا جدول (۸) نشان‌دهنده ترتیب سرشاخه بودن هر خصیصه در درخت J48 است. در این جدول، ۶ خصیصه آخر در مدل دسته‌بندی J48 جایگاهی نداشته‌اند. لذا به‌صورت تصادفی به انتهای جدول اضافه شده‌اند.

همان‌گونه که در مقدمه مطرح شد در مدل کلی الگوریتم کلونی مورچگان که متناسب با مسئله یافتن کوتاه‌ترین مسیر تنظیم شده است، پس از مقداردهی مؤلفه‌های الگوریتم (از جمله $\alpha, \beta, \tau_0, \rho$)، هر مورچه با شروع از یک گره، احتمال انتخاب گره بعدی را با استفاده از تابع احتمال (۱۳) به‌دست می‌آورد.

$$P_{ij}^{k}(t)=\begin{cases}\frac{[\tau_{ij}]^{\alpha}.[\eta_{ij}]^{\beta}}{\sum_{l\epsilon N_i^k}[\tau_{il}]^{\alpha}*[\eta_{il}]^{\beta}} & if\ \ j\epsilon N_i^k\\ 0 & otherwise\end{cases} \qquad (۱۳)$$

اما در مسئله کاهش خصایص، ازآنجاکه فاصله بین خصایص تعریف نشده و عملاً از یک خصیصه به خصیصه دیگر جهش می‌شود، لذا مقدار فرومون و مقدار ابتکاری را به‌جای یال، به گره اختصاص می‌یابد و درنتیجه رابطه (۱۴) جایگزین رابطه (۱۳) می‌شود.

$$P_{j}^{k}(t)=\begin{cases}\frac{[\tau_{j}]^{\alpha}.[\eta_{j}]^{\beta}}{\sum_{l\epsilon N_k}[\tau_{l}]^{\alpha}*[\eta_{l}]^{\beta}} & if\ \ j\epsilon N_k\\ 0 & otherwise\end{cases} \qquad (۱۴)$$

در تابع احتمال (۱۴)، N_k مجموعه گره‌هایی است که توسط مورچه k ملاقات نشده و τ_j مقدار فرومون باقی‌مانده در گره j است و η_j را مطابق رابطه (۱۵) تعریف می‌شود.

$$H_j = \text{Classify}(S_{k_j}(t)) \qquad (۱۵)$$

که در آن $S_{k_j}(t)$ آرایه گره‌هایی است که مورچه kاُم در مرحله t با انتخاب گره j از ابتدا تاکنون طی کرده و تابع Classify(S) نشان‌دهنده نرخ صحت دسته‌بندی ستون‌هایی از مجموعه داده

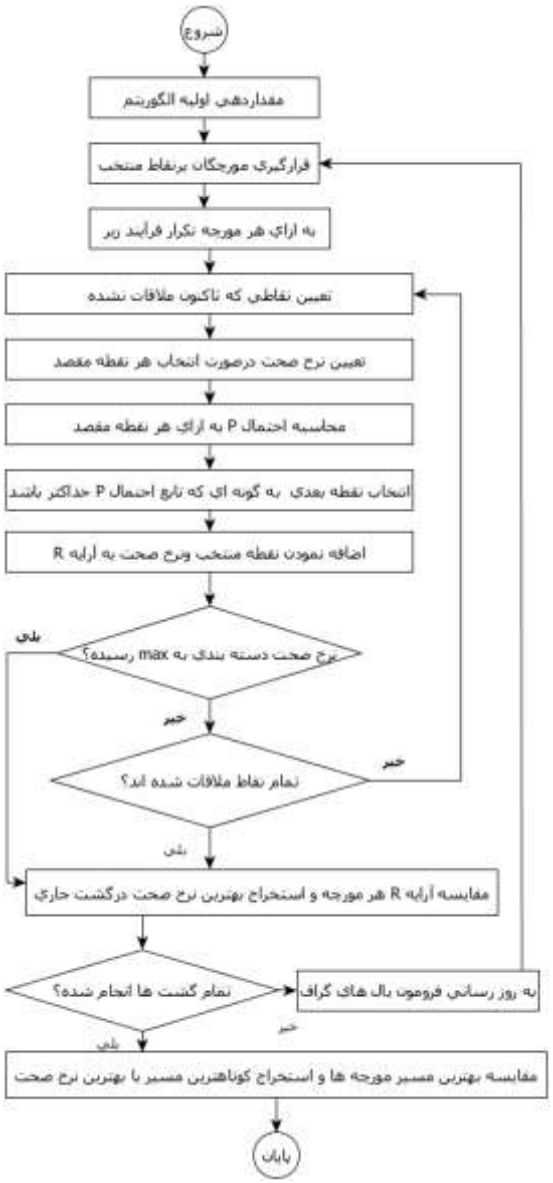
است که در آرایه خصایص S تعیین شده‌اند. برای این منظور تابع Classify در هر مرحله ابتدا با استفاده از کل داده‌های آموزشی، مدل دسته‌بندی را ساخته و سپس به ازای کل داده‌های آزمایشی نرخ صحت دسته‌بندی را تعیین می‌کند. الگوریتم دسته‌بندی نیز، الگوریتم منتخب مرحله قبل یعنی الگوریتم PART می‌باشد.

جدول (۸): اولویت‌بندی خصایص بر اساس درخت تصمیم J48

شماره خصیصه	اولویت	خصیصه
۴	۱	src_bytes
۱۲	۲	num_compromised
۲۱	۳	Counts
۱	۴	protocol_cost
۳۴	۵	dst_host_same_src_port_rate
۳۳	۶	dst_host_diff_srv_rate
۷	۷	wrong_fragment
۹	۸	hot
۳۵	۹	dst_host_srv_diff_host_rate
۳۶	۱۰	dst_host_serror_rate
۱۱	۱۱	logged_in
۳	۱۲	flag_cost
۳۷	۱۳	dst_host_srv_serror_rate
۳۱	۱۴	dst_host_srv_count
۳۰	۱۵	dst_host_count
۲۸	۱۶	diff_srv_rate
۲۹	۱۷	srv_diff_host_rate
۲	۱۸	service_cost
۱۷	۱۹	num_shells
۳۸	۲۰	dst_host_rerror_rate
۵	۲۱	dst_bytes
۱۰	۲۲	num_failed_logins
۳۲	۲۳	dst_host_same_srv_rate
۲۴	۲۴	srv_serror_rate
۲۵	۲۵	rerror_rate
۰	۲۶	duration
۲۶	۲۷	srv_rerror_rate
۲۲	۲۸	srv_count
۱۳	۲۹	root_shell
۶	۳۰	Land
۱۶	۳۱	num_file_creations
۱۸	۳۲	num_access_files
۲۳	۳۳	serror_rate
۲۰	۳۴	is_guest_login
۸	۳۵	Urgent
۱۴	۳۶	su_attempted
۱۵	۳۷	num_root
۱۹	۳۸	is_host_login
۲۷	۳۹	same_srv_rate
۳۹	۴۰	dst_host_srv_rerror_rate

با توجه به توضیحات ارائه‌شده، هر مورچه در ابتدا یکی از خصایص (غیر از نوع حمله) را انتخاب می‌کند و خصیصه بعدی را به‌گونه‌ای از میان خصایص ملاقات نشده، انتخاب می‌کند که با اضافه شدن آن خصیصه به آرایه خصایص ملاقات شده، نرخ صحت دسته‌بندی خصایص، ارتقای بهتری داشته باشد.

هر مورچه در هر مرحله، نرخ صحت دسته‌بندی خود را به همراه گره انتخاب‌شده در آرایه $R_k(t)$ قرار می‌دهد. ازآنجاکه ترتیب خصایص در الگوریتم دسته‌بندی بی‌تأثیر است لذا چنانچه در مرحله t، گره‌های انتخابی مورچه y مشابه گره‌های انتخابی مورچه x شود، باید گره دیگری توسط مورچه x انتخاب شود.



شکل (۵): روند نمای فرآیند کلی طرح ACFSM

در پایان هر گشت، که شامل یک دور کامل توسط تمام مورچه‌ها است، میزان فرومون یال‌ها مطابق رابطه (۱۶) بهنگام می‌شود.

$$\tau_j(t+1) = (1-\rho).\tau_j(t) + \sum_{k=1}^{m} \Delta\tau_j^k(t) \qquad (۱۶)$$

برای محاسبه $\Delta\tau_j^k(t)$ به‌جای رابطه (۳) از رابطه (۱۷) استفاده می‌شود.

$$\Delta\tau_j^k(t)=\begin{cases}\frac{\tau_j^k(t)+\tau_{j-1}^k(t)+R_j^k(t)}{3} & if\ j\ visited\ by\ ant_k\\ 0 & otherwise\end{cases} \quad (۱۷)$$

که رابطه (۱۷)، $R_j^k(t)$ مقدار نرخ صحت در صورت انتخاب گره j و τ_{j-1}^k فرومون باقی‌مانده در گره قبلی مسیر مورچه k اُم است.

دلیل جایگذاری رابطه (۱۷) جای رابطه (۳) را می‌توان در موارد زیر برشمرد:

۱) مبنای ارائه رابطه (۳)، یافتن کوتاه‌ترین مسیر بود لذا در محاسبه $\Delta\tau_{ij}^k$ که نشان‌دهنده جذابیت مسیر است از عکس طول مسیر استفاده شد. اما در مسئله کاهش خصایص، موضوع یافتن زنجیره‌ای از خصایص است که دسته‌بندی بهتری از مجموعه داده را حاصل کند. لذا اگر نرخ صحت خصایص انتخاب‌شده (یعنی مقادیر آرایه $R_j^k(t)$) با انتخاب گره جاری بیشتر شود، انتخاب این گره برای مورچه جذاب‌تر است.

۲) در مسئله کاهش خصایص، نرخ صحتی که در یک گره حاصل می‌شود نتیجه انتخاب آن گره به‌تنهایی نیست بلکه گره‌های قبلی که از ابتدای مسیر انتخاب شده‌اند نیز تأثیرگذارند لذا τ_{j-1}^k تأثیر انتخاب گره‌های قبلی را به گره جاری منتقل می‌کند.

۳) باقی‌مانده فرومون گره جاری (یعنی τ_j^k) نیز منطقاً باید در مقدار فرومون جدید آن گره تأثیرگذار باشد و جهت جلوگیری از افزایش غیرمعمول فرومون یک گره و ایجاد جذابیت کاذب، این مقدار را بر تعداد فاکتورهای تأثیرگذار (یعنی۳) تقسیم می‌کنیم.

در پایان جستجو، آرایه بهینه بر اساس قواعد زیر تعیین می‌شود:

۴) چنانچه مورچه‌ای موفق شود در میانه راه به نرخ صحت ۱۰۰٪ برسد در این صورت قطعاً این مورچه، قهرمان بلامنازع این تورنمنت بوده و مسیر پیشنهادی وی بهترین آرایه خصایص جهت مرحله آزمایش است.

۵) در صورتی‌که بند یک محقق نشود، آنگاه با پویش آرایه R_k به ازای هر مورچه، گره‌ای که دارای حداکثر نرخ صحت دسته‌بندی است، تعیین‌شده و مسیر طی‌شده تا این گره به‌عنوان بهترین مسیر پیشنهادی هر مورچه انتخاب می‌شود.

۶) در پایان تمام گشت‌ها، بهترین مسیرهای پیشنهادی مورچه‌ها با هم مقایسه شده و بین آن‌ها، بهترین مسیر به‌عنوان خصایص برگزیده تعیین می‌گردد.

طرح ACFSM با زبان برنامه‌نویسی جاوا پیاده‌سازی شده است که در بخش بعد به جزئیات آن پرداخته‌می‌شود.

۴- ارزیابی و نتایج

جهت بررسی میزان موفقیت تحقیق لازم است که نتایج حاصل‌شده با ابزار مناسب ارزیابی گردد. برای این منظور مراحل زیر انجام گردید:

- ذخیره‌سازی و پیش‌پردازش مجموعه داده در SQL-Server
- استفاده از ۵ الگوریتم منتخب دسته‌بندی به‌نام‌های MLP, J48, PART, BeaysNet, LogitBoost جهت دسته‌بندی داده‌هایی که در مرحله قبل تولیدشده‌اند در نرم‌افزار وکا.
- پیاده‌سازی الگوریتم کلونی مورچگان و ادغام آن با الگوریتم منتخب داده‌کاوی با زبان برنامه‌نویسی جاوا
- آموزش و آزمایش مجموعه داده با نرم‌افزار تولیدشده.

۴-۱- تشریح جزئیات ارزیابی طرح ACFSM

برای دسته‌بندی اولیه مجموعه داده با استفاده از الگوریتم‌های منتخب، از نرم‌افزار داده‌کاوی وِکا[۱] نسخه ۳/۹ استفاده شد.

با استفاده از ابزار دسته‌بندی این نرم‌افزار، داده‌های آموزشی (شامل ۱۲۵۹۷۳ ردیف ۴۱ ستونی) با ۵ الگوریتم زیر مدل‌سازی شد:

۱) الگوریتم J48 که یک درخت تصمیم و تعمیم‌یافته درخت تصمیم C4.5 است و در تنظیم مؤلفه‌های آن می‌توان تعیین کرد که برگ‌های کم اثر، هرس گردند یا خیر.

۲) الگوریتم PART که یک الگوریتم قانون پایه و تعمیمی دیگری از درخت تصمیم C4.5 است. در این الگوریتم، در هر مرحله یک درخت تصمیم مقطعی C4.5 تشکیل شده و بهترین برگ مبنای تعریف یک سطر قانون قرار می‌گیرد. در هر سطر قانون، شرایط انتخاب با تابع منطقی AND از یکدیگر تفکیک می‌گردند.

۱- وکا (WEKA) یک نرم‌افزار متن باز است که توسط دانشگاه Wailkato نیوزیلند جهت آزمایش پژوهش‌های داده‌کاوی توسعه داده‌شده است و امروزه توسط بیش‌تر محققان جهت ارائه نتایج داده‌کاوی استفاده می‌شود.

۳) الگوریتم BayesNet با تابع تخمین SimpleEstimator و تابع جستجوی K2

۴) شبکه عصبی چندلایه پرسپترون (MLP)

۵) الگوریتم LogitBoost که به روش‌های مبتنی بر رگرسیون اقدام به دسته‌بندی مجموعه داده می‌کند.

مؤلفه‌های تأثیرگذار هر الگوریتم که در نرم‌افزار وکا قابل‌تغییر است، در جدول (۹) و نتایج دسته‌بندی هر الگوریتم در جدول (۱۰) آمده است.

اطلاعات جدول (۱۱) نشان‌دهنده پیشی گرفتن الگوریتم PART در این مرحله است.

در مرحله دوم داده‌های آزمایشی که باید متفاوت با داده‌های آموزشی باشند، با استفاده از مدل آماده‌شده، آزمایش می‌شود تا نوع حمله احتمالی پیش‌بینی گردد.

جدول (۹): مقادیر مؤلفه‌های اصلی الگوریتم‌های دسته‌بندی

PART	ConfidenceFacor=0.25 MinNumObject=2 Seed=1 unpruned=false useMDLocation=true
J48	ConfidenceFacor=0.25 MinNumObject=2 Seed=1
MLP	LearningRate=0.3 Momentum=0.2 TrainingTime=500 ValidationSetSize=0 Seed=0 ValidationThreshold=20 HidenLayers=a
BayesNet	NumDecimalPlaces=2 Estimator=SimpleEstimator(alpha=0.5) SearchAlgorithm=K2(maxParent=1)
LogitBoost	LikelihoodThreShold=-1.7976 ZMax=3.0 WeightThreshol=100 classifier=decisionStump Shrinkage=1.0

جدول (۱۰): نتایج دسته‌بندی در مرحله آموزش

الگوریتم	زمان ساخت مدل(ثانیه)	نرخ صحت
BayesNet	۵.۹۲	۹۵.۷۳۸۰%
J48	۳۲.۵۷	۹۹.۹۱۴۳%
MLP	۱۱۶۹.۱۳	۹۸.۶۱۸۰%
PART	۳۷.۹۱	۹۹.۹۴۸۴%
LogitBoost	۶۲.۹۳	۹۸.۴۳۷۸ %

جدول (۱۱): نتایج دسته‌بندی در مرحله آزمایش

الگوریتم	زمان ساخت مدل(s)	زمان آزمایش(s)	نرخ صحت
BayesNet	**۵/۶۵**	۰/۴۵	۷۳/۳۱۳۲%
J48	۳۲/۳	۰/۳۴	۷۵/۱۵۴۱%
MLP	۱۱۵۱/۷۱	۰/۴۲	۷۲/۷۱۸۸%
PART	۳۸/۲۲	**۰/۳۱**	**۷۶/۲۲۷۷%**
LogitBoost	۶۱/۸۳	۰/۵۵	۷۴/۰۴۵۲ %

داده‌های آزمایشی مجموعه داده NSL-KDD شامل ۲۲۵۴۳ ردیف ۴۱ خصیصه‌ای می‌باشند. نتایج پیش‌بینی نوع حمله با استفاده از ۵ الگوریتم فوق در جدول (۱۱) آمده است.

نتایج به‌دست‌آمده در این مرحله نشان می‌دهد که نرخ صحت پیش‌بینی نوع حمله حداکثر ۷۶/۲۲۷۷٪ است که مربوط به الگوریتم PART است. هم‌چنین در این روش زمان تشخیص نمونه‌های آزمایشی ۰/۳۱s است که بهتر از سایر روش‌ها است. لذا در این مرحله سعی شده است با تغییر مؤلفه‌های پیش‌فرض الگوریتم PART، عملکرد آن بهبود یابد. بنابراین با تغییر unpruned=true و useMDLocation=false نرخ صحت تشخیص مرحله آموزش به ۹۹/۹۸۷۳٪ و نرخ صحت پیش‌بینی مرحله آزمایش به ۷۸/۱۷۹۵٪ افزایش می‌یابد. به این معنا که در مرحله آموزش از بین ۱۲۵۹۷۳ ردیف تنها ۱۶ ردیف اشتباه دسته‌بندی‌شده‌اند که نتیجه بسیار خوبی برای مرحله آموزش است. لذا این الگوریتم مبنای بهبود دسته‌بندی ردیف‌های NSL-KDD در مرحله بعد قرار می‌گیرد.

برای ارزیابی مرحله اصلی طرح ACFSM نرم‌افزار مناسبی یافت نشد. بنابراین با استفاده از زبان برنامه‌سازی جاوا در محیط NetBeans8.1 الگوریتم کلونی مورچگان پیاده‌سازی گردید و برای پیاده‌سازی الگوریتم PART از کتابخانه آماده نرم‌افزار وکا استفاده شد. در این برنامه مؤلفه‌های الگوریتم کلونی مورچگان به‌صورت زیر مقداردهی اولیه شد:

$n=۴۰$ $m=۶$ $\alpha=۰/۵$ $\beta=۰/۵$ $\rho=۰/۳$ $\tau_0=۰/۴$ $T=۵$

با اجرای الگوریتم کلونی مورچگان مسیر انتخابی هر مورچه و نرخ صحت در هر مرحله مطابق جدول(۱۸) مشخص گردید.

همان‌گونه که از اطلاعات جدول (۱۸) مشخص است بالاترین نرخ صحت دسته‌بندی متعلق به مورچه شماره یک است که در گام شماره ۱۷ موفق به کسب حداکثر نرخ صحت دسته‌بندی ۸۵/۳۵۲۴۴٪ شده است. ذکر این نکته لازم است، با توجه به این‌که در بخش پیش‌پردازش بردار خصایص، خصیصه شماره ۲۰ حذف گردید، لذا جهت تطبیق شماره خصایص استخراج‌شده با شماره خصایص تعیین‌شده توسط تدوین‌کنندگان مجموعه داده NSL-DKK، در زنجیره‌های فوق به خصایصی که بزرگ‌تر از ۲۰ هستند باید یک عدد افزود. زنجیره بهینه خصایص بر اساس نام و شماره‌اصلی در جدول (۱۲) آمده است. در جدول (۱۲) شماره هر خصیصه بر اساس جدول (۱۷) تنظیم شده و ۷ خصیصه انتهایی که شماره آن‌ها بالاتر از ۲۰ بوده به‌اضافه یک گردیده است.

حال یک‌بار دیگر با استفاده از این زنجیره خصایص، مراحل آموزش و آزمایش مجدداً تکرار می‌شود. نتایج حاصل از خصایص انتخابی طرح ACFSM برای مرحله آموزش در جدول (۱۳) و برای مرحله آزمایش در جدول (۱۴) آمده است.

جدول (۱۲):خصایص انتخاب‌شده بر اساس نام و شماره اصلی

۳	Service Network	۱۹	Num-access-Files
۵	Src-bytes	۲۱	Is-host-login
۷	Land	۲۲	Is-guest-login
۹	Urgent	۲۷	Rerror-rate
۱۱	Num-failed-logins	۲۸	Srv-rerror-rate
۱۲	Logged-in	۳۰	Diff-srv-rate
۱۳	Num-compromised	۳۲	Dst-host-count
۱۴	Root-shell	۳۷	Dst-host-srv-diff-host-rate
۱۸	Num-shells	رده	Attack_category

برای مقایسه نتایج ACFSM با نتایج سایر مقالاتی که از مجموعه داده NSL-KDD استفاده کرده و اصول انعکاس نتایج علمی را نیز رعایت کرده باشند از مقاله اقدم در [۹] استفاده می‌شود. اقدم در مقاله خود نتایجش را با ۳ مقاله دیگر با مجموعه داده KDD-Cup99 نیز مقایسه کرده است. لذا می‌توان به‌صورت غیرمستقیم نتایج خود را با آن ۳ مقاله نیز مقایسه نمود.

جدول (۱۳): نتایج ACFSM در مرحله آموزش

Class	Accuracy	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
normal	۰/۹۹۹	۰/۹۹۹	۰/۰۰۱	۰/۹۹۹	۰/۹۹۹	۰/۹۹۹	۰/۹۹۸	۱	۱
dos	۰/۹۹۸	۰/۹۹۹	۰/۰۰۱	۰/۹۹۸	۰/۹۹۹	۰/۹۹۹	۰/۹۹۸	۱	۱
r2l	۰/۹۸۵	۰/۹۸۴	۰	۰/۹۸۸	۰/۹۸۴	۰/۹۸۶	۰/۹۸۶	۱	۰/۹۹۹
probe	۰/۹۹۵	۰/۹۹۵	۰	۰/۹۹۸	۰/۹۹۵	۰/۹۹۶	۰/۹۹۶	۱	۱
u2r	۰/۸۶۵	۰/۸۶۵	۰	۰/۹۳۸	۰/۸۶۵	۰/۹	۰/۹۰۱	۱	۰/۹۷۲
avg	۰/۹۹۸	۰/۹۹۸	۰/۰۰۱	۰/۹۹۸	۰/۹۹۸	۰/۹۹۸	۰/۹۹۷	۱	۱

جدول (۱۴): نتایج ACFSM در مرحله آزمایش

Class	Accuracy	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
normal	۰/۹۶۴	۰/۹۶۳	۰/۲۱۵	۰/۷۷۲	۰/۹۶۳	۰/۸۵۷	۰/۷۴۲	۰/۸۷۷	۰/۷۶۹
dos	۰/۹۴۲	۰/۹۴۲	۰/۰۱	۰/۹۷۸	۰/۹۴۲	۰/۹۶	۰/۹۴۱	۰/۹۶۶	۰/۹۴
r2l	۰/۲۵	۰/۲۵	۰/۰۰۱	۰/۹۷۷	۰/۲۵	۰/۳۹۸	۰/۴۶۸	۰/۶۳۱	۰/۳۵
probe	۰/۸۴۹	۰/۸۴۹	۰/۰۲۱	۰/۸۳	۰/۸۴۹	۰/۸۳۹	۰/۸۲	۰/۹۲۵	۰/۷۳۵
u2r	۰/۲۲۴	۰/۲۲۴	۰	۰/۶۲۵	۰/۲۲۴	۰/۳۳	۰/۳۷۳	۰/۶۱۹	۰/۱۴۱
avg	۰/۸۵۳	۰/۸۵۱	۰/۰۹۹	۰/۸۷۲	۰/۸۵۱	۰/۸۲۹	۰/۷۸	۰/۸۸	۰/۷۶۷

در جدول (۱۵) مقایسه نتایج ACFSM با اقدم آمده و در جدول (۱۶) مقایسه نتایج اقدم با نتایج سایر مقالات آمده است.این دو مقایسه نشان می‌دهد که ACFSM در تشخیص حملات DoS,Probe,R2L در مجموعه داده NSL-KDD بهتر از اقدم عمل کرده و اقدم نیز در شناسایی حملات DoS,U2R,R2L در مجموعه داده KDD99 بهتر از سایر تحقیقات عمل کرده است. لذا می‌توان نتیجه‌گیری کرد که ACFSM حداقل در شناسایی دو نوع حمله DoS,R2L بهتر از سایر تحقیقات عمل کرده است.

جدول (۱۵): مقایسه ACFSM با نتایج اقدم در مجموعه داده NSL-KDD

	Accuracy	Normal	DoS	Probe	R2L	U2R
ACFSM	۸۵/۳۵	۹۶/۴	۹۴/۲۵	۸۴/۸۸	۲۴/۹۹	۲۲/۳۹
Aghdam [9]	۸۴/۱	۹۷/۶۵	۷۵/۲۶	۶۸/۸۶	۲۴/۶۶	۲۶/۴۵

جدول (۱۶): مقایسه اقدم با سایر مقالات در مجموعه داده KDD99 [9]

Model	Normal	DoS	U2R	R2L	Probe	Accuracy
Aghdam [9]	۹۷/۴۱	۹۹/۷۸	۳۱/۱	۹۹/۱۷	۷۴/۶۵	۹۸/۹
PLSSVM [21]	۹۵/۶۹	۷۸/۷۶	۳۰/۷	۸۴/۸۵	۸۶/۴۶	NoRep
C_feature [22]	۹۹/۳	۹۹/۵	۱۹/۷	۲۸/۸	۹۷/۵	۹۵/۷
ESC-IDS [23]	۹۸/۲	۹۹/۵	۱۴/۱	۳۱/۵	۸۴/۱	۹۵/۳

۵- نتیجه‌گیری و کارهای آینده

طرح ACFSM با استفاده از الگوریتم کلونی مورچگان زنجیره‌ای از خصایص را معرفی می‌کند که استفاده از این زنجیره در الگوریتم اصلی تشخیص نفوذ، موجب افزایش دقت در تشخیص نوع حمله و کاهش زمان تشخیص نفوذ، می‌گردد.

نتایج تحقیق نشان می‌دهد که طرح ACFSM نرخ صحت تشخیص نوع حمله را از متوسط ۸۴/۱٪ در سایر تحقیقات موفق به، ۸۵/۳۵٪ افزایش و زمان تشخیص نفوذ برای مجموعه داده آزمایشی NSL-KDD حدود ۲۰٪ کاهش‌یافته است. این نتایج در مقایسه با نتایج سایر محققین، نشان‌دهنده این است که استفاده از الگوریتم کلونی مورچگان جهت کاهش زنجیره خصایص به‌صورت بارزی سبب کاهش فضای تصمیم، افزایش دقت تشخیص و کاهش زمان تشخیص می‌گردد.

در پایان برای محققینی که مایل هستند نتایج این تحقیق را توسعه دهند موارد زیر پیشنهاد می‌گردد:

۱) بررسی بیش‌تر خصایص ۴۱ گانه جهت اضافه نمودن خصایص مؤثری که در مجموعه داده NSL-KDD پیش‌بینی نشده است.
۲) استفاده از سایر روش‌های داده‌کاوی همچون تصویر کردن چند خصیصه در یک خصیصه جهت کاهش هدفمند خصایص.
۳) استفاده از سایر الگوریتم‌های جستجوی ابتکاری یا تلفیقی از آن‌ها، جهت افزایش سرعت و دقت جستجو باهدف ارتقای سطح تشخیص نفوذ.

جدول (۱۷): فهرست خصایص هر ثبت اتصال [۹]

۱	خصایص پایه	Duration	مدت‌زمانی که اتصال برقرار بوده برحسب ثانیه
۲		Protocol-type	نوع پروتکل(ICMP،UDP،TCP) که بعد از اتمام ارتباط پیش‌بینی می‌شود و از داخل پاکت مشتق نشده است
۳		Service Network	aol, auth, bgp, courier, csnet_ns, ctf, daytime, discard, domain, domain_u, echo, eco_i, ecr_i, efs, exec, finger, ftp, ftp_data, gopher, harvest, hostnames, http, http_2784, http_443, http_8001, imap4, IRC, iso_tsap, klogin, kshell, ldap, link, login, mtp, name, netbios_dgm, netbios_ns, netbios_ssn, netstat, nnsp, nntp, ntp_u, other, pm_dump, pop_2, pop_3, printer, private, red_i, remote_job, rje, shell, smtp, sql_net, ssh, sunrpc, supdup, systat, telnet, tftp_u, tim_i, time, urh_i, urp_i, uucp, uucp_path, vmnet, whois, X11, Z39_50
۴		Flag	پرچم حالت که خلاصه‌ای از وضعیت اتصال است و نباید با Flagهای موجود در بسته TCP اشتباه شود.
۵		Src-bytes	کل بایت‌های مبدأ که به مقصد ارسال‌شده است
۶		Dst-bytes	کل بایت‌های مقصد که به مبدأ ارسال‌شده است
۷		Land	اگر آدرس و پورت آدرس مقصد و مبدأ یکسان است مقدار ۱ و در غیراین‌صورت صفر
۸		Wrong-fragment	تعداد fragment های اشتباه در این اتصال
۹		Urgent	تعداد پاکت‌های اضطراری (پاکت‌هایی که urgent bit آن‌ها یک است)
۱۰	خصایص مربوط به محتوای بسته	Hot	Number of hot indicators in the content such as: entering a systemdirectory, Creating programs and executing programs
۱۱		Num-failed-logins	Number of failed login attempts
۱۲		Logged-in	1 if successfully logged in 0 otherwise
۱۳		Num-compromised	Number of compromised conditions
۱۴		Root-shell	1 if root shell is obtained 0 otherwise
۱۵		Su-attempted	1 if su root command attempted or used 0 otherwise
۱۶		Num-root	Number of root accesses or number of operations performed as a root in the connection
۱۷		Num-File-creations	Number of File creation operations in the connection
۱۸		Num-shells	Number of shell prompts
۱۹		Num-access-Files	Number of operations on access control Files
۲۰		Num-outbound-cmds	Number of outbound commands in an ftp session
۲۱		Is-host-login	1 if the login belongs to the hot list i.e., root or admin 0 otherwise
۲۲		Is-guest-login	1 if the login is a guest login 0 otherwise
۲۳	خصایص وابسته به زمان	Count	Number of connections to the same host as the current connection in the past two seconds
۲۴		Srv-count	Number of connections to the same service(port number) as the current connection in the past two seconds
۲۵		Serror-rate	The percentage of connections that have activated the flag (4) s0, s1,s2 or s3, among the connections aggregated in count (23)
۲۶		Srv-serror-rate	The percentage of connections that have activated the flag (4) s0, s1,s2 or s3, among the connections aggregated in count (24)
۲۷		Rerror-rate	The percentage of connections that have activated the flag (4) REJ, among the connections aggregated in count (23)
۲۸		Srv-rerror-rate	The percentage of connections that have activated the flag (4) REJ, among the connections aggregated in srv_count (24)
۲۹		Same-srv-rate	The percentage of connections that were to the same service, among the connections aggregated in count (23)
۳۰		Diff-srv-rate	The percentage of connections that were to different services, among the connections aggregated in count (23)
۳۱		Srv-diff-host-rate	The percentage of connections that were to Different destination machines , among the connections aggregated in srv_count (24)
۳۲	خصایص ترافیکی وابسته به میزبان	Dst-host-count	Number of connections having the same destination host IP address
۳۳		Dst-host-srv-count	Number of connections having the same port number
۳۴		Dst-host-same-srv-rate	The percentage of connections that were to the same service, among the connections aggregated in dst_host_count (32)
۳۵		Dst-host-diff-srv-rate	The percentage of connections that were to different services, among the connections aggregated in dst_host_count (32)
۳۶		Dst-host-same-src-port-rate	The percentage of connections that were to the same source port, among the connections aggregated in dst_host_srv_count (33)
۳۷		Dst-host-srv-diff-host-rate	The percentage of connections that were to different destination machines, among the connections aggregated in dst_host_srv_count (33)
۳۸		Dst-host-serror-rate	The percentage of connections that have activated the flag (4) s0, s1, s2 or s3, among the connections aggregated in dst_host_count (32)
۳۹		Dst-host-srv-serror-rate	The percent of connections that have activated the flag (4) s0, s1, s2 or s3, among the connections aggregated in dst_host_srv_count (33)
۴۰		Dst-host-rerror-rate	The percentage of connections that have activated the flag (4) REJ, among the connections aggregated in dst_host_count(32)
۴۱		Dst-host-srv-rerror-rate	The percentage of connections that have activated the flag (4) REJ, among the connections aggregated in dst_host_srv_count (33)

جدول (۱۸): مسیر و نرخ صحت پیش‌بینی خصایص انتخابی هر مورچه

مورچه	خصایص انتخاب‌شده توسط هر مورچه به همراه نرخ صحت دسته‌بندی تا هر خصیصه				
1	7: 83.19212%	18: 82.17185%	29: 81.41773%	3: 76.78215%	5: Start
	27: 82.12749%	13: 82.65537%	11: 81.66171%	9: 82.08313%	20: 83.19212%
	36: 84.12811%	26: 83.755486%	14: 84.11037%	21: 83.23648%	31: 82.739655%
	2: 84.14142%	10: 84.25232%	15: 83.515945%	**19: 85.35244%**	12: 85.01974%
	8: 82.91709%	28: 81.09835%	17: 82.81063%	35: 83.61797%	16: 84.451935%
	23: 80.37528%	38: 80.54385%	40: 79.55907%	33: 81.04955%	30: 82.53116%
	22: 77.70483%	6: 77.57619%	4: 78.454506%	25: 79.80305%	1: 80.41964%
	32: 78.1573%	24: 77.57175%	37: 76.21434%	39: 78.38797%	34: 76.95959%
	بالاترین نرخ صحت	خصایص انتخاب‌شده			تعداد گام
	85.35244%	5-3-29-18-7-20-9-11-13-27-31-21-14-26-36-12-19			17
2	11: 79.08885%	2: 78.45007%	26: 76.2454%	5: 72.190926%	13: Start
	20: 80.61926%	7: 80.61926%	16: 80.61926%	21: 80.517235%	12: 79.63891%
	27: 80.628136%	17: 80.15792%	8: 80.166794%	19: 80.50393%	9: 80.59264%
	03: 80.21558%	36: 80.24664%	18: 82.31824%	15: 82.3138%	14: 82.03433%
	31: 80.48174%	28: 82.02103%	30: 82.9304%	**10: 83.728874%**	29: 81.22255%
	06: 81.187065%	1: 78.87592%	24: 80.10912%	23: 79.740944%	35: 81.426605%
	33: 79.96717%	39: 77.868965%	37: 79.19088%	25: 79.23524%	22: 80.086945%
	34: 77.89558%	4: 77.727005%	40: 78.04196%	32: 79.56794%	38: 79.714325%
	بالاترین نرخ صحت	خصایص انتخاب‌شده			تعداد گام
	83.728874%	13-5-26-2-11-12-21-16-7-20-9-19-8-17-27-14-15-18-36-3-29-10			22
3	06: 80.74347%	11: 79.6123%	03: 78.6009%	05: 73.18458%	35: Start
	18: 81.320145%	17: 81.26691%	09: 81.09391%	20: 81.15157%	19: 81.15157%
	14: 81.533066%	08: 81.00519%	07: 81.09835%	26: 81.11609%	15: 81.617355%
	24: 78.782776%	30: 80.12687%	31: 79.878456%	27: 80.07807%	**13: 81.697205%**
	10: 81.04955%	02: 80.53054%	16: 80.623695%	25: 81.01849%	29: 79.72763%
	01: 80.490616%	36: 79.16427%	39: 80.11356%	38: 79.88732%	21: 80.53054%
	34: 78.54323%	37: 78.28594%	40: 79.62116%	28: 78.5521%	12: 79.96274%
	04: 78.991264%	22: 78.050835%	32: 79.22193%	23: 77.970985%	33: 77.59393%
	بالاترین نرخ صحت	خصایص انتخاب‌شده			تعداد گام
	81.697205%	35-5-3-11-6-19-20-9-17-18-15-26-7-8-14-13			16
4	18: 78.94247%	17: 78.884796%	3: 78.28594%	5: 73.26443%	22: Start
	9: 78.60977%	19: 78.62752%	7: 78.60977%	15: 78.63638%	20: 78.94247%
	16: 79.040054%	28: 79.22193%	25: 79.27516%	21: 78.88924%	6: 77.80686%
	27:82.606575%	1: 81.10278%	29: 79.90507%	35: 80.23777%	11: 79.27073%
	37: 79.55019%	23: 81.81254%	40: 80.56603%	**4: 83.48489%**	13: 82.602135%
	39: 79.10216%	33: 79.29734%	24: 80.22446%	30: 80.255516%	26: 82.30049%
	32: 80.9076%	14: 80.49949%	10: 80.21115%	8: 79.77643%	31: 81.12496%
	34: 77.23462%	36: 77.811295%	2: 78.25046%	12: 79.02675%	38: 78.71179%
	بالاترین نرخ صحت	خصایص انتخاب‌شده			تعداد گام
	80.23777%	22-5-3-17-18-20-15-7-19-9-6-21-25-28-16-11-35-29-1-27-13-4			22
5	7: 79.69658%	12: 79.69658%	26: 77.77137%	5: 73.92095%	2: Start
	11: 79.92281%	20: 80.04702%	13: 80.04702%	21: 79.70545%	15: 79.69658%
	27: 84.026085%	14: 83.946236%	29: 82.415825%	31: 80.730156%	19: 80.58821%
	28: 83.02355%	**10: 84.691475%**	30: 84.451935%	9: 84.52291%	3: 84.642685%
	36: 82.61988%	35: 82.282745%	8: 82.78401%	16: 83.3252%	17: 82.85499%
	24: 80.73904%	1: 79.07555%	33: 79.66553%	25: 80.14905%	18: 82.681984%
	22: 77.34995%	34: 79.11103%	4: 79.99822%	40: 79.54132%	23: 80.637%
	39: 78.74285%	38: 77.18139%	32: 77.96655%	37: 77.39875%	6: 78.35248%
	بالاترین نرخ صحت	خصایص انتخاب‌شده			تعداد گام
	84.642685%	2-5-26-12-7-15-21-13-20-11-19-31-29-14-27-3-9-30-10			19
نتیجه	بهترین نرخ صحت	بهترین خصایص(به ترتیب)			تعداد گام
	85.35244%	3-5-7-9-11-12-13-14-18-19-20-21-26-27-29-31-36			17

۶- مراجع

[1] M. Hosseinzadeh Aghdam and P. Kabiri, "Feature Selection for Intrusion Detection System Using Ant Colony Optimization," International Journal of Network Security, vol. 18, pp. 420-432, 2016.

[2] F. Amiri, M. Rezaei.Yousefi, and C. Lucas, "Mutual information-based feature selection for intrusion detection systems," International Journal of Network and Computer Applications, vol. 34, pp. 1184-1199, 2011.

[3] S. Horng, M. Su, Y. Chen, and T. Kao, "A novel intrusion detection system based on hierarchical clustering and support vector machines," International Journal of Expert Systems with Applications, vol. 38, pp. 306-3313, 2011.

[4] A. Toosi and M. Kahani, "A new approach to intrusion detection based on an evolutionary soft computing model using neuro-fuzzy classifiers," International Journal of Computer Communications, vol. 30, pp. 2201-2212, 2007.

[5] H.-H. Gao, H.-H. Yang, and X.-Y. Wang, "Ant colony optimization based network intrusion feature selection and detection," in Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, 2005.

[6] D. M. Powers, "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation," School of Informatics and Engineering Flinders University, Adelaide-Australia , December 2007.

[7] A. S. Al-Aziz, A. T. Azar, M. Al-Salama, A. E. Hassanien, and S. E. Hanafy, "Genetic Algorithm with Different Feature Selection Techniques for Anomaly Detectors Generation," in Computer Science and Information Systems, Krakow, 2013.

[8] A. Alazab, M. Hobbs, J. Abawajy, and M. Alazab, "Using Feature selection for intrusion detection system," in Communications and Information Technologies (ISCIT), Gold Coast of Australia, 2012.

[9] M. Ambusaidi, H. Xiangjian, and N. Priyadarsi, "Building an Intrusion Detection System Using a Filter Based Feature selection algorithm," IEEE Transactions on Computers, vol. 65, pp. 2986 - 2998, 2016.

[10] E. Amoroso, "Intrusion Detection: An Introduction to Internet Surveillance," Correlation, Trace Back, Traps, and Response, Sparta, Intrusion.Net, 1999.

[11] S. Zargari and D. Voorhis, "Feature Selection in the Corrected KDD-dataset," in Emerging Intelligent Data and Web Technologies, Third International Conference, 2012.

[12] F. Zhang and D. Wang, "An Effective Feature Selection Approach for Network Intrusion Detection," in Networking, Architecture and Storage(NAS), IEEE Eighth International Conference, 2013.

[13] A. Tesfahun and L. Bhaskari, "Intrusion Detection using Random Forests Classifier with SMOTE and Feature Reduction," in Cloud & Ubiquitous Computing & Emerging Technologies(CUBE), International Conference, 2013.

[14] M. Tavallaee and E. Bagheri, "A Detailed Analysis of the KDD CUP 99 Data Set," in Computational Intelligence for Security and Defense Applications (CISDA), Second IEEE Symposium, 2009.

[15] S. Tabakhi, P. Moradi, and F. Akhlaghian, "An unsupervised feature selection algorithm based on ant colony optimization," Engineering Applications of Artificial Intelligence, vol. 32, pp. 112-123, 2014.

[16] A. Sepahi and J. Rasool, "A Hybrid Approach of Similarity-based and Scenario-based Algorithms in Alert Correlation, Tehran, Sharif University of Technology, 2014. (In Persian)

[17] R. Lippmann, J. Haines, D. Fried, J. Korba, and K. Das, "Analysis and Results of the 1999 DARPA Off-Line Intrusion Detection Evaluation," Lecture Notes in Computer Science(LNCS), vol. 1097, pp. 162-182, 2000.

[18] H. S. Chae, B. O. Jo, S. H. Choi, and T. K. Park, "Feature Selection for Intrusion Detection using NSL-KDD," in Applied Computing Conference(ACC), China, 2014.

[19] M. Mirzaei and M. Bashiri, "Ant Colony Optimization," Tehran, Bazagani, 2010. (In Persian)

[20] Durigo and Marco, "Ant Colony Optimization," Tehran, Naghoos, 2016. (In Persian)

[21] O. Namadchian, "Anomaly-Based Intrusion Detection using Memetic algorithm," Tehran, Malek Ashtar University, 2010. (In Persian)

[22] M. Ghazanfari and S. Alizadeh, "Data mining and knowledge discovery," Tehran, ElmoSanat University, 2013. (In Persian)

[23] S. Parsa and S. H. R. Arabi, "Provide a new approach based on a combination method to detect network intrusion," Electronic and cyber defense, vol. 3, pp. 79-93, 2017. (In Persian)

Journal of Electronical & Cyber Defence
Vol. 6, No. 2, 2018, Serial No. 22

Sequential Forward Feature Selection for Intrusion Detection System, Using Ant Colony Algorithm

M. Abbasi*, S. Bejani

*Imam Hossein University

(Received: 07/06/2017, Accepted: 16/12/2017)

ABSTRACT

Intrusion detection system (IDS) is one of the most important security tools, which is used for detecting computer attacks. This System reacts based on two methods: misuse-based and anomaly-based detection. The time limitation to responding and using low efficiency algorithm is the biggest challenge for researchers to promote detection of attacks in IDS. One of the most significant stages in intrusion detection process is the accurate selection of features of IDS to promote the detection, based on these features. In this article, a new method is presented to determine the most effective features in IDS, based on misuse detection method. In this method, the features of NSL-KDD data set have been reduced by ant colony optimization in sequential forward feature selection algorithm, utilizing PART classification algorithm. For evaluating success rate of this method, a specific software in Java language was implemented, using the functions of the library of WEKA. The results compared with other successful methods show that this method increases detection accuracy rate, with concurrent detection of attack category, from 84.1% to 85.35%. Also, the detection time decreases from 0.31 seconds to less than 0.25 seconds in a data set of approximately twenty thousand members.

Keywords: Intrusion Detection System, Feature Selection, Data Mining, Ant Colony Algorithm, Part Algorithm

* Corresponding Author Email: m.abbasi@sndu.ac.ir