الگوریتم Firefly برای تشخیص نفوذ شبکه

چکیده

تشخیص نفوذ شبکه فرآیند شناسایی فعالیت مخرب در یک شبکه با تحلیل رفتار ترافیک شبکه است. تکنیک های داده کاوی به طور گسترده ای در سیستم تشخیص نفوذ (IDS) برای تشخیص ناهنجاری ها استفاده می شود. کاهش ابعاد نقش حیاتی در IDS بازی می کند، زیرا شناسایی ناهنجاری ها از ویژگی ترافیک شبکه با ابعاد بزرگ، فرایندی زمان بر است. انتخاب مشخصه بر سرعت تحلیل و کار پیشنهادی، استفاده از فیلتر و روش مبتنی بر پوشش با استفاده از الگوریتم firefly در انتخاب ویژگی تاثیر می گذارد. ویژگی های حاصل از آن به طبقه بندی C4.5 و شبکه های بیزین (BN) با مجموعه داده KDD CUP 99 منتهی می شود. نتایج تجربی نشان می دهد که 10 ویژگی برای تشخیص نفوذ، نشان دهنده کفایت دقت بهبود یافته است. کار پیشنهادی با کارهای موجود نشان داده شده است که نشان دهنده پیشرفت قابل ملاحظه ای است.

کلمات کلیدی - امنیت شبکه، سیستم تشخیص نفوذ شبکه (NIDS) ، انتخاب ویژگی، الگوریتم Firefly، اطلاعات متقابل و دوطرفه.

1. مقدمه

سیستم تشخیص نفوذ (IDS) یکی از اجزای مهم سیستم های اطلاعات امن است. مزاحمان ومخلان در شبکه در حال تلاش برای دسترسی به منابع غیر مجاز در شبکه هستند. این امر برای نظارت و تجزیه و تحلیل فعالیت های کاربر و رفتار سیستم بسیار مورد نیاز است. به سادگی با اصلاح پیکربندی پارامترهای سیستم، رفتار سیستم می تواند بی وقفه باشد. از این رو سیستم باید از ویژگی های نظارت دوره ای و الگوهای رفتاری آن برای فعالیت های

عادی و غیر طبیعی ارائه شود. دو نوع IDS [13]وجود دارد که مبتنی بر استقرار در زمان واقعی و مکانیزم تشخیص است. IDS مبتنی بر استقرار به IDS مبتنی بر میزبان (HIDS) و IDS مبتنی بر شبکه (NIDS) طبقه بندی شده است. HIDS فعالیت های داخلی یک سیستم محاسباتی را نظارت می کند. NIDS به طور پویا log های ترافیک شبکه را در زمان واقعی برای شناسایی نفوذهای بالقوه در یک شبکه با استفاده از الگوریتم های تشخیص مناسب کنترل می کند. IDS بر اساس مکانیزم تشخیص به تشخیص سوء استفاده، تشخیص غیرمتعارف و IDS ترکیبی طبقه بندی شده است. تشخیص سو مصرف از مجموعه قوانین و یا امضاهای از پیش تعیین شده برای شناسایی حملات شناخته شده استفاده می شود. تشخیص ناهنجاری یک پروفایل فعالیت نرمال، برای تشخیص حملات ناشناخته با چک کردن این موضوع می باشد که آیا وضعیت سیستم پروفایل از فعالیت نرمال تثبیت شده تغییر می کند یا خیر. هیبرید IDS حملات شناخته شده وناشناخته را تشخیص می دهد. امروزه انواع IDS از تکنیک های داده کاوی برای تشخیص نفوذ استفاده می کنند. اکثر NIDS های موجود، با استفاده از تمام ویژگی های ساخته شده از ترافیک شبکه، حملات را شناسایی می کنند. اما برای تشخیص حملات، تمام خصوصیات لازم نیست. کاهش تعداد خواص یا ویژگی ها می تواند زمان تشخیص را کاهش دهد و میزان تشخیص را نیز افزایش می دهد. در این کار، ما روشی مبتنی بر فیلتر و پوشش را برای انتخاب ویژگی های مناسب برای تشخیص نفوذ شبکه ترکیب کردیم. انگیزه کار کاهش تعداد ویژگی ها با عملکرد بهبود یافته برای نرخ تشخیص غیر ترکیبی است. کار پیشنهادی بر روی NIDS تمرکز دارد. اگر چه تکنیک های مختلف در مکتوبات برای NIDS در مورد انتخاب ویژگی ها، طبقه بندی ها وجود دارد، روش پیشنهادی بر روی رویکرد اکتشافی موسوم به تکنیک firefly برای انتخاب ویژگی و ۴C ۵ در مقایسه با طبقه بندی کننده شبکه بیزین متمرکز می باشد.

باقی مانده این مقاله به شرح زیر است. بخش دوم، کار مرتبط در مکتوبات را مشخص می کند. شرح مجموعه داده ها در بخش سوم ارائه شده و بخش چهارم ارائه کار پیشنهاد شده برای تشخیص نفوذ است. نتایج و بحث ها در بخش V و به دنبال آن اظهارات نهایی در بخش VI صورت می گیرد.

2. کار مرتبط

NIDS فعالیت شبکه را بر اساس اطلاعات بارگیری و ویژگی های آماری ترافیک شبکه را نظارت می کند. یک نظرسنجی دقیق از روش های موجود در ارائه روش ها و کاربرد آنها در ابزارهای NIDS توسط Monowar و همکاران انجام شده است. [13]. همچنین آنها حملات کامل مربوط به HIDS و NIDS را ذکر کرده اند. علاوه بر این، آنها بر نیاز به استخراج ویژگی های موثر تأکید کردند که نقش مهمی در تشخیص نفوذ دارند. روش های تشخیص همراه با معیارهای مورد استفاده برای ارزیابی عملکرد NIDS مورد بحث قرار گرفت. یک شبکه عصبی مبتنی بر NIDS توسط Gowrison و همکاران [7] همراه با الگوریتم تقویت شده با پیچیدگی محاسباتی کمتر پیشنهاد شده است. همچنین آنها ارتباط میان ترکیبی از ویژگی ها و حملات را در قالب دستور نشان دادند [29]. آزمایشات انجام شده بر روی KDDCUP'99 انجام شد. یک کار مشابه نیز توسط Weiming و همکاران [22] با روش های پارامتری مبتنی بر Adaboost آنلاین انجام شده است.

عدم نظارت سیستم تشخیص ناهنجاری برای تشخیص نفوذ توسط Jungsuk و همکاران [9] با داده های بدون برچسب انجام شد. علیرغم مزایا، هنوز هم سخت است که آنها را در یک محیط شبکه واقعی قرار دهیم. برای غلبه بر معایب کار مبتنی بر خوشه بندی، Deepak و همکاران. [6] یک رویکرد ترکیبی را پیشنهاد می کند که ترکیبی از خوشه بندی K-Medoids و طبقه بندی Naie-Bayes است. در ابتدا کارشنان جمع آوری داده ها برای تشکیل یک گروه انجام شد و پس از آن یک طبقه بندی برای شناسایی نفوذ در شبکه استفاده شد. تکنیک های داده کاوی توسط Vaishali و همکاران [18] و Uday و همکاران. [3] برای شناسایی الگوهای شناخته شده و ناشناخته حملات مورد استفاده قرار گرفت.

با توجه به ترکیبات مختلف ویژگی های پرونده های ترافیک شبکه، روش های بهینه سازی توسط محققان معرفی شده است. Revathi و همکاران [16] کار خود را با استفاده از تکنیک هوشمندانه برای حل مشکل پیچیده بهینه سازی و پیش پردازش داده انجام دادند. الگوریتم های مبتنی بر ژنتیک توسط Skalak و همکاران [۴] مورد استفاده قرار گرفتند که در آن جهش تصادفی برای انتخاب ویژگی ها با رویکرد ابتکاری صعود برای IDS بکار گرفته شد. بیش از یک طبقه بندی ضعیف توسط Akhilesh و همکاران [1] توسط مجموعه ای از شبکه عصبی مصنوعی (ANN) و شبکه بیسیم با نسبت افزایش (GR) ویژگی انتخاب برای تشخیص نفوذ استفاده می شود.

تجزیه و تحلیل مولفه اصلی (PCA) یکی از روش های ابزار انتخاب ویژگی ها بود. یکی از این روش ها توسط Keerthi et al [10]برای کاهش ابعاد استفاده شد. آنها با آزمایش PCA با استفاده از الگوریتم های طبقه بندی تصادفی و C4.5 با KDD CUP و UNB ISCX آزمایش های انجام شده را انجام دادند. در کار خود، دقت طبقه بندی به دست آمده توسط 10 مولفه اصلی با 41 ویژگی با استفاده از طبقه بندی C4.5 مقایسه شد.

روش انتخاب ویژگی مبتنی بر پوشش توسط وی و همکاران [۲۰] پیشنهاد شد و آزمایش ها بر روی داده های " ۹۹KDD" انجام شد. در کار آن ها، به جای ساختن تعداد زیادی از مشخصه ها از ترافیک شبکه گسترده، هدف نویسندگان انتخاب بهترین ویژگی ها و استفاده از آن ها برای تشخیص حملات نفوذ به شیوه ای سریع و موثر است. آن ها ابتدا از انتخاب ویژگی براساس روش فیلتر و روش پوشش استفاده کردند. انتخاب ویژگی مبتنی بر فیلتر برای ویژگی های مهم براساس ارتباط بین خصوصیت و ویژگی های مهم براساس رتبه بندی انتخاب می شود. انتخاب ویژگی مبتنی بر پوشش از برخی از روش های جستجو برای انتخاب زیر مجموعه از ویژگی ها و زیر مجموعه های انتخاب شده با استفاده از شبکه C4.5 و بیزین استفاده کرد. با این حال، Siva و همکاران. [26] از جستجوی ژنتیکی به عنوان یک استراتژی جستجو برای انتخاب ویژگی های مبتنی بر پوشش استفاده کرد تا زیرمجموعه مطلوب را انتخاب کند. اما Lei Yu و همکاران [12] یک مدل همبستگی مبتنی بر فیلتر ایجاد کرده اند تا ویژگی ها را سریعتر انتخاب کنند بدون اینکه از کارایی استفاده کنند. طبقه بندی ها بر اساس ماشین های بردار پشتیبانی و شبکه های عصبی توسط Sung و همکاران [2] با 13 ویژگی انتخاب شده مورد استفاده قرار گرفت. یک مدل محاسباتی موازی و تکنیک انتخاب ویژگی الهام گرفته از طبیعت توسط Natesan و همکاران [27] پیشنهاد انتخاب و طبقه بندی کارآمد برای به دست آوردن میزان تشخیص بهینه سازی شده انجام شد. همچنین نقشه کاهش مدل برنامه نویسی برای انتخاب زیر مجموعه های بهینه با پیچیدگی محاسباتی کم استفاده می شود. IDS با استفاده از تئوری مجموعه ((Rough (rst)همراه با SVM توسط چن و همکاران ساخته شد که در آن rst برای انتخاب ویژگی های مهم استفاده شد [۱۵]. مجموعه داده NSL-KDD که تنوع KDDCUP'99 است توسط Dhanabal et al استفاده شده است. [11] که کار آنها از نظر دیگران از نظر استفاده از داده ها متفاوت است.

3. توصیف DATASET

در این کار، ما از مجموعه داده KDD CUP 99 [14]استفاده کردیم که شامل انواع عادی و حمله (22 نوع مختلف) است. هر رکورد از داده ها از گروهی از بسته ها که بیش از ۲ پنجره دوم اتصال که به یک مقصد ایجاد شده است، ساخته شده است. هر رکورد داده دارای 41 ویژگی (34 عدد، 4 باینری و 3 عدد) است. 9 ویژگی اول نشان دهنده اطلاعات آماری اولیه بسته ها بر روی یک اتصال است، ویژگی های بعدی 13 ویژگی محتوای بسته ها را نشان می دهد، و 9 ویژگی دیگر نشان دهنده اطلاعات ترافیکی است. 9 ویژگی های دیگر ویژگی های میزبان را نشان می دهد. انواع مختلفی از حمله وجود دارد که در طی یک زمان وارد شبکه می شوند و حملات به چهار طبقه اصلی زیر تقسیم می شوند. آنها به طور خلاصه شرح داده می شوند:

- Denial of Service (Dos): مهاجمان سعی می کند از استفاده از یک سرویس از کاربران قانونی جلوگیری کند.
- از راه دور به محلی (R2L): مهاجم دارای حساب کاربری در دستگاه قربانی نیست، بلکه تلاش می کند دسترسی به آن را داشته باشد.
- کاربر به ریشه (U2R): مهاجم دارای دسترسی محلی به دستگاه قربانی است و تلاش می کند تا امتیازات فوق العاده کاربر را به دست آورد.
- Probe: مهاجم تلاش می کند اطلاعاتی در مورد میزبان هدف به دست آورد.

4. کار پیشنهاد شده

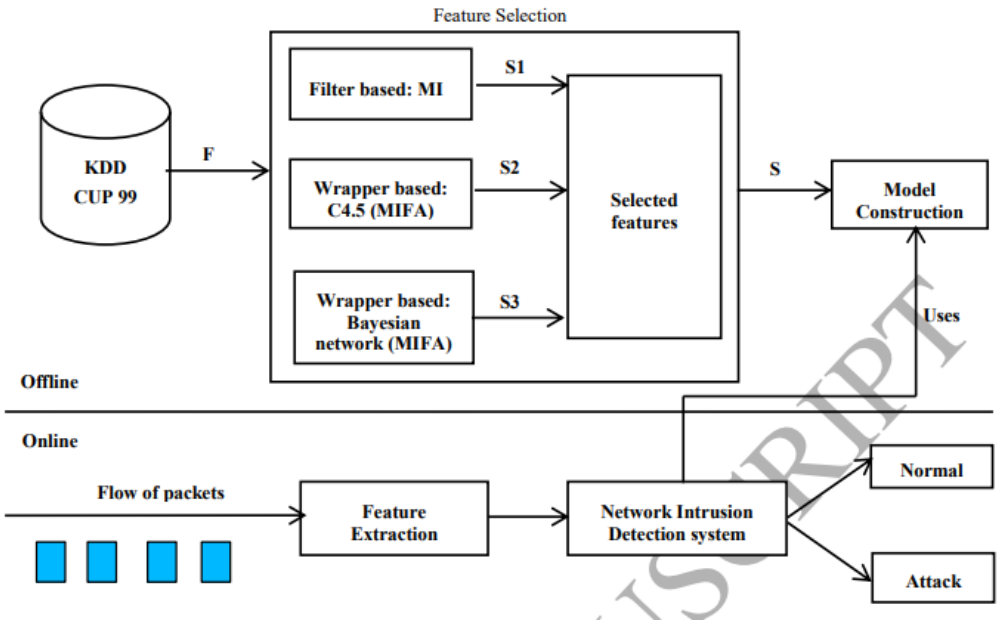
در طول چند سال گذشته تعداد بیشتری از کارهای تحقیقاتی تکنیک های داده کاوی را به اشکال مختلفی اعمال کرده اند. در کار پیشنهادی، ما آنها را در سیستم تشخیص نفوذ پذیرفته ایم. شکل 1 معماری کار پیشنهادی را نشان می دهد. انتخاب ویژگی های مهم اولین گام برای تشخیص نفوذ است. انتخاب ویژگی فرآیند انتخاب زیر مجموعه ای از ویژگی های اصلی با توجه به معیارهای خاص است و برای داده هایی با ابعاد بالاتر اهمیت دارد. اجازه دهید F مجموعه ای از ویژگی های دارای تعداد " "nباشد. زیر مجموعه ای از ویژگی های پیچیده، n-12 است. مجموعه ای از زیر مجموعه های ویژگی به عنوان S مشخص شده و توسط:

$S = \{S_1, S_2, \ldots S_{2^n-1}\}$ که در اینجا n = 41 در مجموعه داده KDD.

تعداد زیر مجموعه ها بسیار بزرگ و جامع است. کار با تمام زیر مجموعه ها و گرفتن راه حل های شمارنده فراتر از راه حل عملی است و از این رو استراتژی های مختلف باید اقتباس شوند. الگوریتم برای انتخاب ویژگی می تواند به دو دسته تقسیم شود: انتخاب ویژگی های مبتنی بر فیلتر و انتخاب ویژگی های مبتنی بر پوشش .[17] الگوریتم firefly اکتشافی که ابتدا توسط Xin-She Yang توسعه یافت [25] و در رویکرد پوشش قرار گرفته است که تاکنون در هیچ یک از کارهای موجود در NIDS مورد توجه قرار نگرفته است. ساختن ویژگی های کمتر باعث بهبود کارایی تشخیص نفوذ شبکه می شود. اگرچه هر کار بر روی IDS با مجموعه داده مبنا تمرکز داشت، وی وانگ و همکاران [۱۹] این ویژگی ها را از محیط زمان واقعی ساخته و ویژگی های را با استفاده از آنالیز KNN و اصل مولفه های اصلی طبقه بندی کردند.

A. انتخاب ویژگی مبتنی بر فیلتر

ویژگی ها بر اساس ویژگی های کلی داده های آموزشی بدون تکیه بر الگوریتم های داده کاوی ارزیابی می شوند. این زیرمجموعه را با محتوای اطلاعات خود یا با اطلاعات متقابل یا با افزایش اطلاعات ارزیابی می کند. ما این ویژگی را با بزرگترین اطلاعات متقابل (MI) انتخاب کرده ایم. اطلاعات متقابل دو متغیر تصادفی با استفاده از آنتروپی محاسبه می شود که قادر به تعیین عدم قطعیت متغیرهای تصادفی و مقادیر اطلاعات به اشتراک گذاشته شده توسط آنها می باشد [24].



شکل 1 معماری کار پیشنهادی

فرض کنید X یک متغیر تصادفی گسسته است و عدم اطمینان آن را می توان با انتروپی H (X) اندازه گیری کرد که به صورت زیر محاسبه می شود:

$$H(X)=\sum_{i} p(x_i)\log_2(p(x_i)) \qquad (1)$$

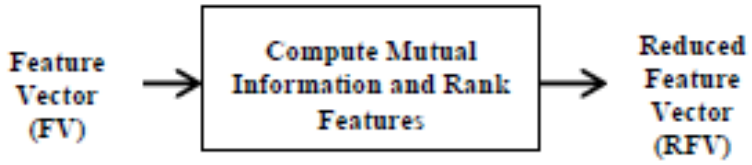
که در اینجا آنتروپی شانون با توزیع احتمال p (x) برای هر رویداد احتمالی x Ω نشان داده می شود (تمام وقایع ممکن). اجازه دهید Y برچسب طبقه X باشد و ما آنتروپی مشترک H (X)، Y را داریم:

$$H(X,Y)=-\sum_{y\in Y}\sum_{x\in X} p(x,y)\log_2(p(x,y)) \qquad (2)$$

که در اینجا p (x)، y توابع توزیع احتمال از X و Y است. اطلاعات متقابل . I(X; Y) بین متغیر نشان دهنده مجموعه داده X و برچسب طبقه Y به عنوان زیر است

$$I(X;Y)=\sum_{y\in Y}\sum_{x\in X} p(x,y)\log_2\frac{p(x,y)}{p(x)p(y)} \qquad (3)$$

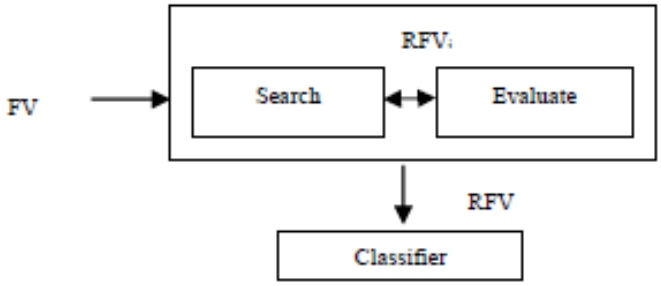
فرض کنید Ss یک زیر مجموعه از ویژگی های F، و C برچسب های طبقه است. اگر اطلاعات ارائه شده در مورد طبقه C ارائه شده توسط این ویژگی دارای بیشترین اطلاعات متقابل بین تمام ویژگی های انتخاب شده در زیرمجموعه Ss باشد، سپس ویژگی Fi انتخاب می شود. شکل 2 روش انتخاب ویژگی مبتنی بر فیلتر را نشان می دهد.



شکل 2 انتخاب ویژگی مبتنی بر فیلتر

B. انتخاب ویژگی های مبتنی بر پوشش

از یک طبقه بندی کننده برای ارزیابی زیرمجموعه ای از ویژگی ها با دقت پیش بینی آن ها استفاده می کند (بر روی داده های آزمون). مقاله بررسی شده توسط Monowar و همکاران (13) در مورد بسیاری از روش های جستجو برای بهترین زیر مجموعه، که در آن یکی از روش ها انتخاب ویژگی های مبتنی بر پوشش است، مورد بحث است. در کار پیشنهادی ما، الگوریتم Firefly اطلاعات متقابل (MIFA) به عنوان یک استراتژی انتخاب ویژگی در انتخاب ویژگی پوشش با استفاده از ۴C ۵ [۸] و شبکه بیزین (۵)به عنوان یک طبقه بندی کننده انتخاب شده است. شکل 3، روش انتخاب ویژگی مبتنی بر پوشش را نشان می دهد.



شکل 3. انتخاب ویژگی مبتنی بر پوشش

به طور کلی، دو نوع الگوریتم تصادفی وجود دارد: اکتشافی و فرااکتشافی. اکتشافی به معنی "یافتن" یا "کشف توسط آزمون و خطا" و فرا اکتشافی، نسخه بهبود یافته الگوریتم اکتشافی و firefly است که در اصل توسط ژین جیانگ یانگ توسعه داده شد [۲۵]، که در آن فرض می شود fireflies با وجود خود به یکدیگر جذب می شوند. تابع هدف هر مساله بهینه سازی را می توان در وجود یک firefly ترسیم کرد.

در الگوریتم firefly دو مسئله مهم وجود دارد: تنوع روشنایی و فرمول جاذبیت. بنابراین جاذبیت بین دو firefly i و j با توجه به فاصله متفاوت است که با فاصله از منبع آن کاهش می یابد. یک عامل دیگر ضریب جذب به دلیل

رسانه ای است که بر جاذبیت تاثیر می گذارد. از این رو روشنایی یک firefly در شعاع ((rاز یک firefly با منبع روشنایی B بصورت زیر است:

$$B(r) = B_0 e^{-\gamma r} \qquad (4)$$

که در اینجا B0 مولفه اصلی است؛ r فاصله بین هر دو firefly است و ضریب جذب نور است که کاهش شدت آن را کنترل می کند. از آنجا که جاذبیت firefly با روشنی مشاهده شده توسط یک firefly دیگر متناسب است، جاذبیت یک firefly به صورت زیر ارایه می شود:

$$A(r) \quad = A_0 e^{-\gamma r} \qquad (5)$$

که در اینجا 0 جاذبیت در r = 0 است. در این صورت firefly i توسط j firefly جذب می شود و این حرکت توسط فرمول زیر نشان داده می شود

$$v_i^{t+1} = A_0 e^{-\gamma r_i^2} (v_j^t - v_i^t) + \beta(R - 0.5) \qquad (6)$$

که در اینجا β پارامتر تصادفی و R یک مولد عدد تصادفی است که به طور یکنواخت بین 0 و 1 توزیع شده است. اصطلاح ' 't'تعداد تکرار را نشان می دهد. تعداد ابعاد D (d = 1 ... D) و rij فاصله بین firefly i و j firefly است که توسط معادله (7) تعریف شده است.

$$r_{ij} = \| v_i - v_j \| = \sqrt{\sum_{d=1}^{D} \left(v_{id} - v_{jd} \right)^2} \qquad (7)$$

در کار پیشنهادی، تعداد ابعاد 41 (D) است که نشان دهنده تعداد کل ویژگی های مربوط به تشخیص نفوذ شبکه است. پیچیدگی انتخاب ویژگی ها D2 است، که نوعی از چند جمله ای غیر قطعی است. از این رو نیاز به انتخاب ویژگی های موثر برای کاهش پیچیدگی محاسبات و ذخیره سازی برای استقرار زمان واقعی وجود دارد. شبه کد برای انتخاب ویژگی های مبتنی بر اطلاعات متقابل و الگوریتم firefly به NIDS داده شده است در الگوریتم ارائه شده است.

در این کار برای ارزیابی انتخاب ویژگی، از C4.5 و شبکه بیزین استفاده شده است. هر firefly به عنوان یک بردار دودویی با تعداد زیادی از ویژگی های D نمایش داده می شود و توسط (vi1 = vi، vi2، vi3، ... vid، viD) i = 1 ... n مشخص می شود که در آن ' 'nتعداد firefly است. هر عنصر vi به 0 یا 1 محدود می شود که نشان می دهد که آیا این ویژگی ترافیک انتخاب شده است یا نه. به عبارت دیگر، هر vi firefly به عنوان یک نقطه در فضای بردار D بعدی قرار می گیرد.

زیر مجموعه ای از ویژگی های (41 ویژگی) توسط ترکیب های مختلف از حضور 0 یا 1 در مجموعه ویژگی های نشان داده شده است.

هر firefly در یک جهت در فضای جستجو حرکت می کند تا زیرمجموعه ویژگی بهینه را بر اساس دقت مدل طبقه بندی با زیر مجموعه ای از ویژگی های انتخاب شده را پیدا کند. دقت ارزیابی کننده (مدل طبقه بندی) شامل ویژگی انتخاب شده به عنوان یک تابع هدف یا وجود firefly در نظر گرفته می شود. Firefly با استفاده از معادله دیفرانسیل،از دقت روشنی کمتری به دقت روشنایی بالاتر حرکت خواهد کرد Eq.6 و فاصله بین دو Firefly با استفاده از معادل 7 محاسبه می شود. تعدادی از ویژگی های حاصل از الگوریتم Firefly متفاوت است. برای داشتن تعدادی از ویژگی های ثابت برای اجرای موثر در الگوریتم Firefly مبتنی بر اطلاعات متقابل (MIFA) ، استراتژی انطباقی در کار فعلی پیشنهاد شده است که نوآوری کار پیشنهادی در NIDS است. این کنترل فرآیند اضافه کردن یا حذف ویژگی های حاصل توسط الگوریتم Firefly است که تعداد مشخصی از ویژگی ها را ایجاد می کند. Long Zhang et.al [28]در انتخاب ویژگی ها با استفاده از الگوریتم firefly برای تعیین داده های مختلف معیار بدون اصلاح تعدادی از ویژگی های مورد نیاز انتخاب شد. در کار پیشنهادی، تعداد ویژگی های انتخاب شده در k ثابت است. اگر تعداد حاصل از ویژگی ها | | vd == 1می گویند " "mکمتر از k است و سپس (k-m) تعداد ویژگی های باقی مانده به آن بر اساس اطلاعات متقابل (MI) از ویژگی های انتخاب نشده اضافه شده است. آن m بزرگتر از k است، MI برای ویژگی های حاصل محاسبه شده و (m-k) تعداد ویژگی های با کمترین MI حذف می شوند. این استراتژی به عنوان استراتژی سازگاری مبتنی بر اطلاعات متقابل (MIAS) معنا می یابد.

موقعیت i firefly در MIFA از قانون به روز شده در معادله (8) استفاده می کند.

$$v_{id}^{t+1} = \begin{cases} 1 & if \ p_{id} > rand \\ 0 & otherwise \end{cases} \qquad (8)$$

Where,

$$p_{id} = \frac{1}{1+e^{v_{id}^t}}$$

v_{id}^{t+1} -نشان دهنده مقدار فعلی ویژگی dth در firefly i است، v_{id}^t نشان دهنده مقدار قبلی از ویژگی dth در firefly i است و rand یک عدد تصادفی یکنواخت توزیع بین 0 و 1 است.

الگوریتم شبیه ساز مبتنی بر firefly (MIFA) به شرح زیر است، که در اینجا الگوریتم تکراری است و برای زمانهای Tmax اجرا می شود

الگوریتم MIFA (n، L، C، k، Tmax)

```
//Input:n - number of fireflies,
//Input: L - Attack class labels, C - Classifier,
// Input: k- number of features to be selected
//Output v - modified position of the firefly

// set of selected network flow features
// T_max – Maximum number of iterations
    {
    v_{i,i=1..n} = InitRand(D choose k)
    θ_{i,i=1..n} = evaluateclassifier(v_i)
    sort(θ)
    t = 1
    while(t < T_max)
    {
     for  i = 1 to n
     {
      for  j = 1 to n    // i ≠ j
      {
      if(θ_i < θ_j)
      {
       move(i, j)  //use equation 6
       updateAttractivenss(i, j) //use eqn.5
       v_id = update(v_id)  //use equation 8
       v_i = MIAS(v_i, k)   //returns modified v
        //as discussed  in section IV
       θ_i = evaluateclassifier(v_i)
      }
      }
     }
      sort(θ)
      t = t + 1
    }
    return(v)
   }
```

در کار پیشنهادی، سه نوع مختلف استراتژی انتخاب ویژگی به شرح زیر است:

• ویژگی مجموعه بر اساس اطلاعات متقابل (S1)

• مجموعه ویژگی های به دست آمده از روش پوشش شده MIFA با C4.5 به عنوان ارزیاب (S2)

• مجموعه ویژگی های به دست آمده از روش پوشش شده MIFA با شبکه بیزی به عنوان ارزیاب (S3)

رأی مبتنی بر انتخاب ویژگی ها از این مجموعه ویژگی ها (یکی از ویژگی های این سه مجموعه انتخاب شده است که در حداقل دو مجموعه موجود است) همانطور که در معادله (9) مشاهده می شود

$$S=\{f: f\in((S1\cap S2)\cup(S1\cap S3)\cup(S2\cap S3))\} \qquad (9)$$

مجموعه ای از ویژگی های نهایی حاصل از آن به عنوان ورودی به طبقه بندی C4.5 استفاده می شود.

5. نتایج و بحث

A. آزمایشات مبتنی بر KDD CUP 99

در این مقاله مجموعه داده KDD CUP 99 برای راه اندازی تجربی مورد استفاده قرار می گیرد که یکی از مجموعه های محبوب برای تشخیص نفوذ است. همانطور که ذکر شد، پرونده ها به خوبی به عنوان عادی یا به عنوان نوع دقیق حمله در NSL-KDD برچسب گذاری شده اند.

جدول 1 توزیع نمونه های حمله که در آزمایش ما استفاده می شود را توصیف می کند.

Attack Category	Types	Training Size	Testing Size
DoS	Normal	40,000	40000
	smurf	10000	10000
	neptune	5000	5000
	back	1000	1203
	land	10	11
	teardrop	100	579
	pod	400	164
	Subtotal	**56510**	**56957**
Probe	Normal	40000	40000
	satan	800	789
	portsweep	500	540
	nmap	110	121
	ipsweep	600	647
	Subtotal	**42010**	**42097**
R2L	Normal	40000	40000
	ftp_write	4	4
	guess_passwd	23	30
	multihop	7	5
	imap	23	4
	warezclient	520	500
	warezmaster	10	10
	phf	4	0
	spy	2	0
	Subtotal	**40573**	**40553**
U2R	Normal	40000	40000
	buffer_overflow	15	15
	rootkit	4	6
	loadmodule	4	5
	Perl	0	3
	Subtotal	**40023**	**40029**

جدول .1 شرح داده ها

A. نتایج کار پیشنهادی

در KDD CUP 99، تمام 22 نوع حمله به طور مساوی توزیع نمی شوند. این ممکن است عملکرد تشخیص نفوذ را کاهش دهد. برای جلوگیری از تاثیر در توزیع نامتقارن داده، داده های آموزشی و داده های تست را که در جدول 1 توضیح داده می شوند، تشکیل می دهیم. جدول 2 ویژگی های مهم را با روش های MIFA مبتنی بر فیلتر و بسته بندی نشان می دهد. جدول 3 مجموعه ای از ویژگی های انتخاب شده توسط روش رای گیری پیشنهاد شده برای انواع مختلف حملات را نشان می دهد. این امر نشان می دهد که تنها 10 ویژگی برای تشخیص نفوذ کافی است. از جدول دوم مشاهده شده است، برخی از ویژگی های همپوشانی بین روش پیشنهاد شده و روش های موجود وجود دارد و برجسته می شوند. در بیشتر موارد، کار پیشنهادی دارای ویژگی های منحصر به فرد نسبت به روش های موجود است. الگوریتم firefly مبتنی بر اطلاعات، $\alpha = 0.1$ (پارامتر تصادفی)، 0 = 1 (جاذبه پایه)، =

1 (ضریب جذب)، n = 10 (تعداد)، Tmax = 100 (حداکثر تعداد تکرار) که پارامترهای اولیه الگوریتم می باشند.

Type	Methods	Important features selected
DOS	MI	**$f_{41}, f_{40}, f_{13}, f_{10}, f_5, f_6, f_{23}$**, f_{28}, **f_{24}, f_{27}**
	Wrapper(C4.5)	f_2, f_3, **f_5, f_6**, f_{11}, **$f_{12}, f_{23}, f_{24}, f_{27}, f_{41}$**
	Wrapper(BN)	f_1, **f_5, f_{12}**, f_{22}, **f_{23}**, f_{25}, **f_{27}**, f_{31}, f_{34}, **f_{40}**
PROBE	MI	**f_{41}, f_{28}, f_{27}**, f_{40}, **f_5, f_6, f_{33}**, f_4, f_{35}, **f_3**

	Wrapper(C4.5)	f_1, **f_2, f_3, f_5**, $f_{10}, f_{16}, f_{31}, f_{39}$, **$f_{40}, f_{41}$**
	Wrapper(BN)	**f_2, f_5, f_6**, f_{19}, f_{22}, f_{26}, **f_{27}**, f_{29}, **f_{31}**, f_{38}
R2L	MI	**f_{41}, f_{40}**, f_{27}, **f_{28}**, f_3, f_{33}, **f_5, f_6**, f_{11}, **f_{24}**
	Wrapper(C4.5)	**f_5, f_6, f_7, f_{13}**, f_{14}, f_{18}, f_{21}, **f_{22}, f_{25}, f_{28}**
	Wrapper(BN)	**f_5, f_6, f_7, f_{13}**, f_{15}, **f_{22}, f_{24}, f_{25}**, f_{32}, f_{36}
U2R	MI	**$f_{41}, f_{27}, f_{28}, f_{40}, f_{33}, f_3, f_5, f_6$**, f_{24}, f_{23}
	Wrapper(C4.5)	**f_3, f_5**, f_8, f_{13}, f_{14}, **f_{15}**, f_{16}, **f_{25}**, f_{35}, **f_{40}**
	Wrapper(BN)	**f_5**, f_{10}, f_{11}, **f_{15}**, f_{20}, **f_{25}**, $f_{26}, f_{29}, f_{32}, f_{39}$

جدول II ویژگی های مهم برای شناسایی انواع حملات با استفاده از روش های مختلف

Attack Type	Important features selected
DOS	$f_5, f_6, f_{10}, f_{12}, f_{13}, f_{23}, f_{24}, f_{27}, f_{40}, f_{41}$
PROBE	$f_2, f_3, f_5, f_6, f_{27}, f_{28}, f_{31}, f_{33}, f_{40}, f_{41}$
R2L	$f_5, f_6, f_7, f_{13}, f_{22}, f_{24}, f_{25}, f_{28}, f_{40}, f_{41}$
U2R	$f_3, f_5, f_6, f_{15}, f_{25}, f_{27}, f_{28}, f_{33}, f_{40}, f_{41}$

جدول III مشخصه مهم انتخاب شده توسط روش پیشنهاد شده ما

تمام آزمایش ها بر روی کامپیوتر با GHZ i5 CPU 3.00 و حافظه RAM 8.00GB انجام می شود. آزمایش ها با ویژگی های انتخاب شده (10 عدد) و تمام 41 ویژگی انجام می شود. بسیاری از طبقه بندی ها مانند C4.5، Naive Bayes، شبکه بیزین و توده تصادفی بوده و نتایج امیدوار کننده برای C4.5 و شبکه بیزین به دست آمده است. دقت طبقه بندی بر روی ویژگی های حاصل شده با استفاده از شبکه C4.5 و شبکه بیزین مقایسه شده و در جدول 4 نشان داده شده است. از جدول آمده است که روش پیشنهادی بهبود عملکرد در مقایسه با طبقه بندی با تمام 41 ویژگی را نشان می دهد. همچنین میزان هشدار اشتباه، اندازه گیری F کار پیشنهادی در جدول V و VI نشان داده شده است.

AttackType	Methods	DR	
		With 41 features	With 10 features
DOS	BN	99.78	**99.95**
	C4.5	99.95	**99.98**
PROBE	BN	87.74	**93.42**
	C4.5	63.04	**63.85**
R2L	BN	99.90	97.83
	C4.5	92.95	**98.73**
U2R	BN	75.86	68.97
	C4.5	31.03	17.24

جدول IV مقایسه ی میزان تشخیص حملات توسط C4.5 و طبقه بندی BN با 10 ویژگی برجسته و 41 ویژگی

Attack Type	Methods	FPR	
		With 41 features	With 10 features
DOS	BN	0.06	**0.01**
	C4.5	0.02	0.03
PROBE	BN	0.05	**0.01**
	C4.5	0.04	**0.00**
R2L	BN	0.018	**0.01**
	C4.5	0.00	**0.00**
U2R	BN	0.29	**0.00**
	C4.5	0.00	0.00

جدول .Vمقایسه ضرایب حملات با کلاس C4.5 و BN با 10 ویژگی های مهم و 41 ویژگی

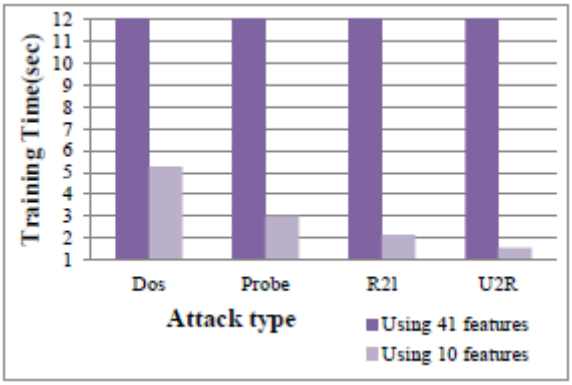
مقایسه عملکرد تشخیص حملات توسط c4.5 و طبقه بندی شبکه بیزین با 10 ویژگی و 41 ویژگی نیز در جدول VI، VII و VIII نشان داده شده است. مقادیر با فونت برجسته در جدول به این معنی است که عملکرد تشخیص حمله با 10 ویژگی عملکرد بهتری نسبت به 41 ویژگی دارد.

Attack Type	Methods	F-Measure	
		With 41 features	With 10 features
DOS	BN	0.93	**0.99**
	C4.5	0.97	0.97
PROBE	BN	0.63	**0.92**
	C4.5	0.52	**0.76**
R2L	BN	0.61	**0.74**
	C4.5	0.96	**0.99**
U2R	BN	0.26	**0.44**
	C4.5	0.47	0.29

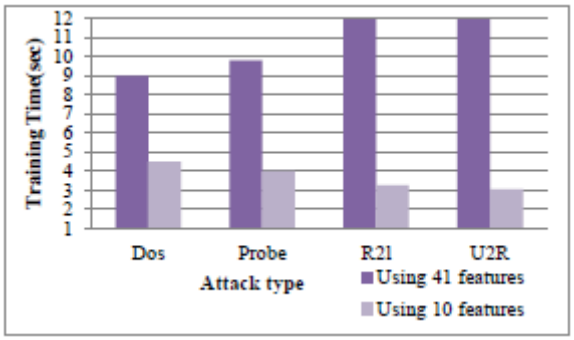
جدول VI مقایسهٔ اندازه گیری F-ATTACKS توسط C4.5 و طبقه بندی BN با 10 ویژگی مهم و 41 ویژگی

مقایسه زمان جابجایی توسط c4.5 و طبقه بندی شبکه بیزین برای 10 ویژگی انتخاب شده با 41 ویژگی در شکل 4 و شکل 5 نشان داده شده است. این نشان می دهد که زمان صرف شده برای ساخت یک مدل با ۱۰ ویژگی کم تر از مدل ساختمان با ۴۱ ویژگی است. مقایسه زمان تشخیص توسط c4.5 و طبقه بندی شبکه بیزین

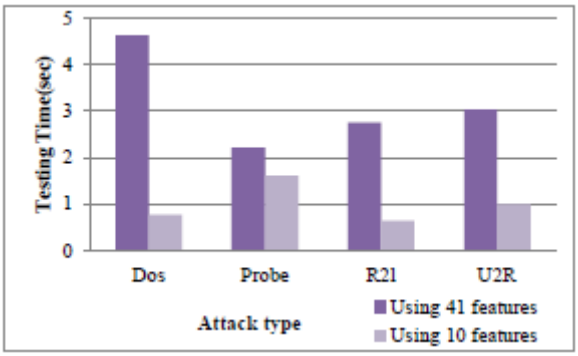
برای 10 ویژگی انتخاب شده با 41 ویژگی در شکل 6 و شکل 7 نشان داده شده است. زمان برای تشخیص نفوذ با ۱۰ ویژگی کم تر از ۴۱ ویژگی طول می کشد. در نتیجه انتخاب ویژگی، زمان محاسبات در هر دو تمرین و تست ذخیره می شود. یک کار اخیر توسط Chuanlong Yin et.al [30] در رویکرد یادگیری عمیق برای تشخیص نفوذ با استفاده از شبکه عصبی مکرر با همان مجموعه داده KDD کار کرده و دقت برای حملات DoS، Probe، R2L و U2R نشان داده شده است که 83.5٪ ، 24.7٪، 11.5٪ و 83.4٪ بود. نرخ های مثبت کاذب برای این حملات 2.1، 0.8، 0.1، و 2.2 نشان داده شده است. در کار پیشنهادی، دقت و هشدارهای کاذب بهبود یافته است. نتایج بهبود یافته برای دقت 99.98٪، 93.42٪، 98.73٪، 68.97٪ و نرخ های بهبود یافته مثبت کاذب عبارتند از: 0.01، 0.01، 0، و 0 که برای حملات DoS، Probe، R2L و U2R می باشند.



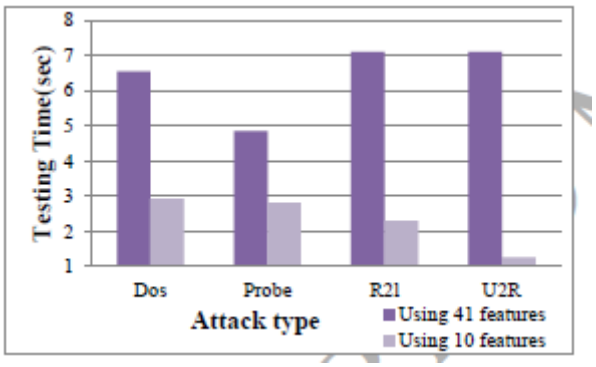
شکل 4 مقایسه زمان آزمایش حملات با طبقه بندی C4.5 برای10 ویژگی و 41 ویژگی انتخاب شده



شکل 5 مقایسه زمان آزمایش حملات توسط طبقه بندی BN برای10 ویژگی و 41 ویژگی انتخاب شده



شکل 6 مقایسه زمان آزمایش حمله با طبقه بندی C4.5 برای 10 ویژگی و 41 ویژگی انتخاب شده



شکل 7 مقایسه زمان آزمایش حمله توسط طبقه بندی BN برای 10 ویژگی و 41 ویژگی انتخاب شده

6. نتیجه گیری و کار آینده

یکی از بیشترین چالش های تشخیص نفوذ شبکه، رسیدگی به اطلاعات گسترده برای تشخیص نفوذ است. میزان تشخیص NIDS بر اساس تعداد نمونه ها و همچنین تعداد ویژگی ها است. کاهش اندازه، افزایش دقت تشخیص و کاهش میزان مثبت کاذب، وظیفه حیاتی تکنیک های داده کاوی داده برای تشخیص نفوذ است. بیشتر روش موجود نمی تواند تمام یا بیشتر از 41 ویژگی را برای شناسایی نفوذ در شبکه و بر اساس مجموعه داده KDD CUP 99 و NSL-KDD استفاده کند. در این کار الگوریتم انتخاب جدید ویژگی برای انتخاب ویژگی با استفاده از KDD CUP 99 ارائه شده است. ما ویژگی های مناسب از تعداد کل ویژگی ها (41) را برای تشخیص نفوذ در شبکه انتخاب کردیم. چندین روش انتخاب ویژگی، بر اساس اطلاعات متقابل (MI) و پوشش با شبکه بیزین، C4.5 برای انتخاب ویژگی استفاده می شود. با تنها 10 ویژگی مناسب، عملکرد تشخیص بهتر از 41 ویژگی است و هزینه های محاسباتی برای طبقه بندی را کاهش می دهد. بهره وری تشخیص با ویژگی های مناسب بهبود می

یابد. روش پیشنهادی ما برای انتخاب ویژگی، نتیجه بهتر را به جای روش موجود برای انتخاب ویژگی فراهم می کند. توسعه کار با استفاده از امکانات GPU در حال پیشرفت است تا زمان صرف شده برای محاسبه و بهبود نتایج را کاهش دهد.

References

[1] Akhilesh Kumar Shrivas and Amit Kumar Dewangan,"An Ensemble Model for Classification of Attacks with Feature Selection based on KDD99 and NSL-KDD Data Set," International Journal of Computer Applications (0975 – 8887) Vol. 99 – No.15,August 2014.

[2] Andrew H. Sung, SrinivasMukkamala, "Identifying important features for intrusion detection using support vector machines and neural networks," Sympon Applications and the Internet, 2003.

[3] B.UdayBabu,C.G.Priya and Vishakh,"Survey on intrusion detection techniques using data-mining domain," IJERT, 2014. Vol. 3.

[4] David B.Skalak, "Protopype and feature selection by sampling and Random Mutation Hill Climbing algorithms".

[5] David Heckerman,"A Tutorial on Learning with Bayesian Networks," Microsoft Research, Technical Report MSRTR-95-06, March 1995.

[6] Deepak Upadhyaya and Shubha Jain, "Hybrid Approach for Network Intrusion Detection System Using K-Medoid Clustering and Naïve Bayes Classification," IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 3, No 1, pp 231-236, May 2013.

[7] G.Gowrison,K.Ramar,K.Muneeswaran and K.Revathi, "Minimal complexity attack classification intrusion detection system," Appl. Soft Comput., 2013, 13, (2), pp. 921–927 .

[8] J.Ross Quinlan."C4.5: Programs for Machine Learning," Morgan Kaufmann Publishers, 1993.

[9] Jungsuk Song, HirokiTakakurb, yasuo Okabe, and Koji Nakao,"Toward a more practical unsupervised anomaly detection system," Inf. Sci., 2013, 231, (10), pp. 4–14.

[10] K. KeerthiVasan and B.Surendiran, "Dimensionality reduction using Principal Component analysis for network intrusion detection," Elsevier, 2016.

[11] L.Dhanabal and Dr.S.P. Shantharajah,"A study on NSL-KDD dataset for intrusion detection system based on classification algorithms," IJARCCE,Vol. 5,6,June 2015.

[12] Lei Yu, Huan Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," ICML, 2003, pp. 856–863.

[13] Monowar H. Bhuyan,D.K.Bhattacharyya and J.K. Kalita "Network anomaly detection: Methods,systems and tools," IEEE Commun. Surv. Tutor. 2014, 16, (1), pp. 303–336.

[14] NSL-KDD data set,"https://github.com/defcom17/NSL_KDD".

[15] Rung-Ching Chen, Kai-Fan Cheng and Chia-Fen Hsieh, "Using Rough Set And Support Vector Machine For Network Intrusion Detection," International Journal of Network Security & Its Applications (IJNSA), Vol 1, No 1, April 2009.

[16] S. Revathi and A. Malathi, "Data Preprocessing for Intrusion Detection System using Swarm Intelligence Techniques," International Journal of Computer Applications , Volume 75– No.6, August 2013.

[17] SwathiV.Jadhav, VishwakamaPinki, "A survey on feature selection methods for High dimensional data," IJRITCC, 2016, pp. 83-86.

[18] Vaishali B Kosamkar and Sangita S Chaudhari, "Data Mining Algorithms for Intrusion Detection System: An Overview," International Conference in Recent Trends in Information Technology and Computer Science (ICRTITCS), 2012.

[19] Wei Wang, Xiangliang Zhang and Sylvain Gombault "Constructing attribute weights from computer audit data for effective intrusion detection," J. Syst. Softw., 2009, 82, (12), pp. 1974–1981.

[20] Wei Wang, Yongzhong He, Jiqiang Liu and Sylvain Gombault,"Constructing important features from massive network traffic for lightweight intrusion detection," IET, 2015, pp. 374-379.

[21] Wei Wang,SylvainGombault,"Efficient detection of DDoS attacks with important attributes," CRiSIS, 2008, pp. 61–67.

[22] Weiming Hu, Jun Gao, Yanguo Wang, Ou Wu, and Stephen Maybank, "Online adaboost-based parameterized methods for dynamic distributed network intrusion detection," IEEE Trans. Cybern., 2014, 44, (1), pp. 66–82.
[23] Wenke Lee and Salvatore J. Stolfo, "A framework for constructing features and models for intrusion detection systems," ACM Trans. Inf. Syst. Sec., 2000, 3, (4), pp. 227–261.
[24] Cover TM, Thomas JA (2006) Elements of information theory (Wiley series in telecommunications and signal processing). WileyInterscience, London.
[25] X.-S. Yang,"Firefly algorithm, Levy flights and global optimization", in: Research and Development in Intelligent Systems XXVI (Eds M. Bramer, R. Ellis, M. Petridis), Springer London, pp. 209-218 (2010)
[26] Siva S., Sivatha Sindhu,Geetha S.,Kannan a.,"Decision tree based light weight intrusion detection using a wrapper approach", Elsevier, Expert Systems with Applications,pp. 129-141,2012.
[27] Natesan P., Rajalaxmi R.R., and Gowrison G., "Hadoop based parallel Binary Bat Algorithm for Network Intrusion Detection", Springer,Int J Parallel Prog,PP. 1-20,2016.
[28] Long Zhang, LinlinshanandJianhuaWang,"Optimal feature selection using distance–based firefly algorithm with mutual Information criterion",Springer,2016.
[29] G. Gowrison., et.al "Efficient context-free grammar intrusion detection system" International Journal of Innovative Computing, , Volume 7, Number 8, pp.1-20, August 2011
[30] Chuanlong Yin , Yuefei Zhu, Jinlong Fei, And Xinzheng He, A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks, DOI: 1 0.1109/ACCESS.2017.2762418, 2017