هفتمین کنفرانس بین المللی انجمن رمز ایران

۲۴ و ۲۵ شهریورماه ۱۳۸۹، دانشگاه صنعتی خواجه نصیرالدین طوسی

انجمن رمزایران

دانشگاه صنعتی خواجه نصیرالدین طوسی

# بهبود تشخیص نفوذ براساس کاهش ویژگی و با استفاده از داده‌کاوی

مریم معدنی‌پور، حسن ابوالحسنی، حسین شیرازی

تهران، دانشگاه صنعتی مالک اشتر، مجتمع پژوهشی فناوری اطلاعات، ارتباطات و امنیت

madanipour@ece.ut.ac.ir

تهران، دانشگاه صنعتی شریف، دانشکده مهندسی کامپیوتر

abolhassani@sharif.edu

تهران، دانشگاه صنعتی مالک اشتر، مجتمع پژوهشی فناوری اطلاعات، ارتباطات و امنیت

shirazi@mut.ac.ir

## چکیده

در سیستم‌های تشخیص نفوذ، با داده‌های حجیم برای تحلیل مواجه هستند. بررسی مجموعه داده سیستم‌های تشخیص نفوذ نشان می‌دهد که بسیاری از ویژگی‌ها، ویژگی‌های غیرمفید، بی‌تاثیر در سناریوهای حمله و یا ویژگی‌های نامربوط هستند. بنابراین حذف ویژگی‌های نامناسب از مجموعه ویژگی، به عنوان یک راهکار مناسب برای کاهش مجموعه داده سیستم‌های تشخیص نفوذ معرفی می‌شود.
نیازمندی دیگری که در سیستم‌های تشخیص نفوذ مطرح می‌باشد، دانستن مجموعه ویژگی بهینه برای هر نوع حمله است. چرا که در اینصورت، سیستم تشخیص نفوذ قادر خواهد بود برای تشخیص هر نوع حمله، تنها از مجموعه ویژگی متناسب با آن حمله استفاده کند.

در این تحقیق، روشی ارائه می‌شود که قادر است تمام نیازمندیهای فوق را پاسخگو باشد، علاوه براین، این روش نحوه ارتباط بین ویژگی‌ها را برای تحلیل بهتر آنها نشان می‌دهد. روش پیشنهادی از مفاهیم داده‌کاوی و تحلیل شبکه‌های اجتماعی استفاده می‌نماید.

## واژه های! کلیدی

کاهش ویژگی، انتخاب ویژگی، سیستم‌های تشخیص نفوذ، داده‌کاوی، تحلیل شبکه‌های اجتماعی.

## ۱- مقدمه !

راهکارهای حفاظتی از قبیل دیوار آتش، احراز هویت و رمزنگاری معمولا برای حفظ امنیت شبکه کافی نیستند.بنابراین، سیستم‌های تشخیص نفوذ[1] یک مکانیزم دفاعی بسیار مهم برای نقاط آسیب-پذیر شبکه های کامپیوتری خواهد بود.

با توجه به حجم زیاد داده ای که در ترافیک شبکه ای وجود دارد استخراج الگوهایی برای رفتارهای نادرست کاربران و پیش بینی رفتارهای آنان کاری زمانبر و دشوار است. در این بین، برخی داده ها ،در سیستم های تشخیص نفوذ مزاحم عمل تشخیص نفوذ هستند. ممکن است بین ویژگی‌های مختلف داده‌ها، ارتباط اشتباهی وجود داشته باشد که مانع از تشخیص درست نفوذ شود.

احتمال تکراری بودن این ویژگی‌ها نیز وجود دارد، به این معنی که بعضی ویژگی‌ها از ویژگی‌های دیگر قابل استنتاج باشد. انتخاب ویژگی‌های مناسب، می‌تواند سرعت دسته‌بندی و یا خوشه‌بندی را افزایش دهد. در این تحقیق قصد بر آن است که کوچکترین مجموعه ویژگی‌ها و ارتباطات بین آنها با استفاده از روش تحلیل گراف استخراج شود تا با کاهش اندازه مجموعه داده-

[1] Intrusion Detection System

-ها، عملیات دسته‌بندی در داده‌کاوی سریعتر انجام شده و درنهایت دقت و سرعت تشخیص نفوذ بهبود یابد. [۱] اثبات کرده است که انتخاب مجموعه ویژگی مناسب، کارایی دسته‌بندی را بهتر می‌نماید. [۲] نشان می‌دهد حذف این ویژگی‌های بی‌اهمیت و نامربوط، باعث کاهش کارایی سیستم‌های تشخیص نفوذ تاحد زیاد نمی‌شود.به دلیل بالابودن حجم داده‌ها در سیستم‌های تشخیص نفوذ، از روشهای داده‌کاوی[1] برای استخراج الگوی رفتاری داده‌ها استفاده می‌شود. به طور ویژه با بهره گیری از روشهای دسته‌بندی[2] موجود، می‌توان مدل گراف مربوط به ویژگی‌ها را به دست آورد. مدل داده‌ای مجموعه داده سیستم‌های تشخیص نفوذ، مسطح می‌باشد. به این معنی که یک مجموعه داده شامل تعدادی رکورد است که هر رکورد شامل چندین ویژگی می‌باشد. چنانچه مدل گراف از این مدل مسطح استخراج شود، می‌توان از مفاهیم موجود در تحلیل گراف استفاده نمود. در تحقیق حاضر، به کمک روشهای داده‌کاوی و الگوریتم پیشنهادی توضیح داده شده در بخش ۴، مدل مسطح به مدل گراف تبدیل می‌شود. در مدل گراف حاصل، ویژگی‌ها، تشکیل دهنده گره‌های گراف و نحوه ارتباط ویژگی‌ها نشان دهنده یالهای موجود در گراف می‌باشد.

برای تحلیل گراف از روشهای موجود در تحلیل شبکه‌های اجتماعی[3]، استفاده می‌شود. علت این انتخاب، شباهت هدف تحلیل شبکه‌های اجتماعی با تحقیق حاضر می‌باشد.

ساختار این مقاله به این ترتیب می‌باشد: دربخش ۲، کارهای انجام شده معرفی می‌شود. در بخش ۳، مقدمه‌ای بر تحلیل شبکه‌های اجتماعی آورده می‌شود. الگوریتم پیشنهادی در بخش ۴ به تفصیل توضیح داده می‌شود. در بخش ۵ ارزیابی روش پیشنهادی ارائه می‌شود و درنهایت در بخش ۶، نتیجه-گیری کلی بیان می‌گردد.

## ۲-کارهای انجام‌شده!

تحقیقات بسیاری در زمینه کاهش ویژگی‌های سیستم‌های تشخیص نفوذ انجام شده است، که از بین آنها، برخی روشها، مبتنی بر هوش مصنوعی و برخی براساس روشهای داده کاوی می‌باشند.

از بین کل روشهای انتخاب ویژگی می‌توان شبکه عصبی مصنوعی[۳-۵]، ماشین ساپورت و کتور[4][۶-۷]، شبکه‌های بیضین [۸] ، روش ارائه شده در مقاله [۹] که با هزینه نمایی می‌تواند بهترین ویژگی‌ها را به دست آورد و چندین روش دیگر را نام برد که تمام آنها در بسیاری از سیستم‌ها استفاده شده‌اند. در تحقیق حاضر به علت تمرکز بر استفاده از روشهای داده کاوی، تنها بر روشهای مبتنی بر داده کاوی به طور مفصل پرداخته می‌شود.

روش ارائه شده درمقاله [۱۰]، به دنبال یافتن بهترین مجموعه ویژگی بدون حضور ناظر بیرونی است. در این روش، ابتدا براساس الگوریتم جستجوی روبه‌جلو، زیرمجموعه ویژگی‌ها انتخاب می‌شود. سپس مجموعه داده با استفاده از الگوریتم خوشه‌بندی و زیرمجموعه منتخب، خوشه‌بندی می‌شود. درنهایت دقت خوشه‌بندی انجام شده ارزیابی می‌شود. این روند چندین مرتبه تکرار می‌شود تا خوشه‌بندی با بهترین معیار به دست آید. مزیت این روش این است که قابل اعمال بر روی تمام انواع داده‌های عددی و غیرعددی است اما این روش به این دلیل که به زیرمجموعه انتخابی اولیه وابسته است ممکن است همیشه بهترین مجموعه ویژگی را نیابد.

درمقاله [۱۱]، برای رتبه بندی ویژگی ها از روش ارائه شده در مقاله [۱۲] استفاده می شود. به این ترتیب که با حذف یک ویژگی میزان کیفیت خوشه بندی اندازه گیری می شود. هر چه که حذف یک ویژگی میزان کیفیت پایین‌تری به خوشه بندی بدهد آن ویژگی از اهمیت بیشتری برخوردار خواهد بود. برای یافتن میزان کیفیت خوشه بندی از معیارهای تراکم[5] و پراکندگی[6] استفاده می شود.

این روش قادر است ویژگی‌ها را براساس میزان اهمیت، رتبه‌بندی نماید. اما قادر به یافتن ویژگی‌های افزونه و همچنین بیان نحوه ارتباط ویژگی‌ها نمی‌باشد. علاوه براین، این روش تنها قابل اعمال برروی داده‌های عددی است و برای مجموعه داده‌هایی که به صورت جریان داده است مناسب می‌باشد. [۸]

از روش ترکیبی شبکه بیضین و درخت تصمیم استفاده می‌کند. دقت این روش بالاست اما قادر به رتبه‌بندی ویژگی‌ها از نظر اهمیت نمی‌باشد. همچنین تعداد ویژگی‌های بهینه معرفی شده توسط این روش به نسبت روشهای دیگر بیشتر است.

هیچ یک از روشهای ارائه شده، نمی‌توانند ارتباط بین ویژگی‌ها را به درستی نشان دهند، همچنین قادر به یافتن ویژگی‌های بهینه مرتبط با انواع حملات نمی‌باشند درنتیجه لازم است روشی ارائه شود تا علاوه بریافتن اهمیت ویژگی‌ها، قادر به استخراج ارتباط بین ویژگی‌ها و ویژگی‌های مرتبط با هر نوع حمله باشد.

[1] Data Mining
[2] Classification
[3] Social Network Analysis
[4] Support Vector Machine

[5] Compactness
[6] Separateness

## ۳- تحلیل شبکه های اجتماعی!

شبکه‌های اجتماعی متشکل از اعضای یک گروه است که این اعضا با یکدیگر در ارتباط هستند. شبکه تروریستی، نمونه ای از این نوع شبکه‌هاست که در آن اعضای شبکه به یکدیگر پیامهایی ارسال می‌کنند. یک شبکه تروریستی از یک گروه تروریستی که هیچگونه ساختار سلسله مراتبی ندارد متفاوت است، در این نوع شبکه‌ها معمولا یک الگوی ارتباطی بین اعضا برای رسیدن به هدف وجود دارد. هدف از تحلیل این نوع شبکه‌ها، یافتن مهمترین اعضای شبکه و نحوه ارتباط اعضای شبکه می‌باشد. به طرز مشابهی، در این مقاله نیز به دنبال یافتن مهمترین ویژگی‌ها و نحوه ارتباط بین آنها می‌باشیم درنتیجه برای تحلیل ارتباط ویژگی‌ها می‌توان از روشهای ارائه شده در تحلیل شبکه‌های اجتماعی بهره برد. برای این کار ابتدا لازم است مدل گراف ارتباط ویژگی‌ها به دست آید که نحوه ایجاد این مدل در بخش ۴ توضیح داده می‌شود.

درمقاله [۱۳]، شبکه تروریستی با گراف غیرجهت‌دار نشان داده می‌شود، سپس این گراف به کمک مفاهیم اندازه‌گیری مرکزیت به گراف جهت‌دار تبدیل می‌شود[۱۴-۱۵].برای تحلیل شبکه، روشی به کار می‌رود تا به کمک مفهوم مرکزیت وابستگی[1] ،گراف جهت‌دار به نمودار سلسله مراتبی تبدیل شود. با استفاده از ساختار سلسله مراتبی به راحتی می‌توان بین رهبران گروه و پیروان آنها تمایز قایل شد تا شبکه تروریستی ساخته شود. ایده جدید اندازه‌گیری مرکزیت وابستگی برای ایجاد مدل سلسله مراتبی بسیار به کار می‌آید، زیرا این ایده گره‌هایی که کاملا به گره‌های خاصی وابسته هستند را نشان می‌دهد. نحوه محاسبه مرکزیت وابستگی در فرمول ۱ آمده است.

(۱) مرکزیت وابستگی $$DC_{mn} = \sum_{m \neq p, p \in G} \frac{d_{mn}}{N_p} + \Omega$$

با استفاده از اندازه‌گیری کارایی شبکه می‌توان اثر حذف هر گره از گراف را به خوبی نمایش داد. هرچه حذف یک گره، میزان کارایی شبکه را بیشتر کاهش دهد، اهمیت حضور آن گره بیشتر می‌باشد.

محاسبه کارایی شبکه در فرمول ۲ نمایش داده شده است. در فرمول ۲، مقدار $d_{ij}$ بیان کننده کوتاهترین فاصله بین دو گره i,j می‌باشد.

(۲) مرکزیت کارایی $$E(G) = \frac{\sum_{i \neq j} e_{ij}}{N(N-1)} = \frac{1}{N(N-1)} \sum_{i \neq j \in G} \frac{1}{d_{ij}}$$

مرکزیت نمایه نقش سازمانی[2] مشخص کننده نقش افراد شبکه می‌باشد. نمایه نقش سازمانی با استفاده از مرکزیت کارایی شبکه محاسبه می‌شود. چنانچه مرکزیت نمایه نقش سازمانی، بر روی محور مختصات رسم شود، نقاط بالای محور x نمایانگر رهبران گروه و نقاط زیر محور افقی، بیانگر نقش سربازان و اعضای کم‌اهمیت گروه هستند.

نحوه محاسبه نمایه نقش سازمانی در فرمول ۳ نمایش داده شده است.

(۳) نمایه نقش سازمانی

$$PRI = E(G) - E(G\text{-node}_i),\ i=1,2,\ldots,N$$

در فرمول ۳، $G\text{-node}_i$ نشاندهنده وضعیت شبکه پس از غیرفعال کردن گره i ام است.

با استفاده از مرکزیتهای معرفی شده در این بخش، می‌توان گراف ویژگی‌ها را تحلیل و مدل سلسله مراتبی اهمیت ویژگی‌ها را به دست آورد. در بخش ۴ نحوه استفاده از این تحلیلها بیان می‌شود.

## ۴-الگوریتم پیشنهادی

در الگوریتم پیشنهادی سعی برآن است که از تحلیل شبکه‌های اجتماعی برای یافتن مجموعه ویژگی بهینه استفاده شود.

مشکلی که در مدل داده‌ای سیستمهای تشخیص نفوذ وجود دارد این است که در این سیستم‌ها، اطلاعاتی بین ویژگی‌ها مبادله نمی‌شود درنتیجه لازم است مدل داده ای به نحوی به مدل گراف تبدیل شود، که نحوه انجام این تبدیل در بخش حاضر بیان می‌شود.

### ۴-۱- ایجاد مدل گراف !

برای ایجاد مدل گراف ویژگی‌ها، ماتریس دقت معرفی می‌شود. ستونهای ماتریس دقت، نشاندهنده ویژگی‌ها و ردیفهای آن بیانگر انواع کلاسهای حملات در روشهای مختلف کلاس‌بندی می‌باشد. این ماتریس، شامل f ستون و n*m ردیف می‌باشد، به نحوی که f، تعداد ویژگی‌های موجود در مجموعه داده، n تعداد روشهای کلاس‌بندی و m تعداد انواع حملات را نشان می‌دهد. خانه (i,j) مقدار دقت کلاس i، درصورت عدم حضور ویژگی j را می‌باشد. علاوه بر میزان دقت، متغیر دیگری تحت عنوان زیرآستانه[3] نیز در خانه‌های ماتریس قرار می‌گیرد. این متغیر دو مقدار صفر و یا یک را به خود می‌گیرد. درصورتی که حذف یک ویژگی از مجموعه ویژگی‌ها دقت کلاس را از حد آستانه تعیین

[1] Dependence Centrality
[2] Position Role Index (PRI)
[3] BelowThreshold

شده، پایین تر ببرد مقدار این متغیر به یک تنظیم می‌شود. مقدار یک بیان می‌کند که ویژگی j انتخاب شده برای کلاس i از اهمیت بالایی برخوردار است، چرا که حذف آن، دقت کلاس را بسیار کاهش داده است. در شکل ۱ ماتریس دقت نمایش داده شده است.

| روش کلاس بندی | | ویژگی ۱ | ویژگی ۲ | .. | ویژگی n |
|---|---|---|---|---|---|
| روش کلاس بندی ۱ | کلاس ۱ | | | | |
| | کلاس ۲ | | | | |
| | .. | | | - دقت کلاس i درصورت عدم حضور ویژگی j<br>- متغیر زیرآستانه | |
| | کلاس n | | | | |
| روش کلاس بندی ۲ | کلاس ۱ | | | | |
| | کلاس ۲ | | | | |
| | .. | | | | |
| | کلاس n | | | | |
| ... | | | | | |
| روش کلاس بندی m | | | | | |

شکل ۱. ماتریس دقت

برای محاسبه ماتریس دقت، الگوریتمهای دسته‌بندی مختلف به ازای حذف هر ویژگی، بر روی مجموعه داده اجرا می‌شوند. پس از اجرای هر روش دسته‌بندی، دقت دسته‌بندی مربوط به هرنوع حمله با استفاده از ماتریس اغتشاش که دقت هر دسته را نمایش می‌دهد در ماتریس دقت ثبت می‌شود. چنانچه میزان دقت هردسته کمتر از حد آستانه باشد، نشان دهنده اهمیت حضور ویژگی حذف شده می‌باشد. درنتیجه مقدار زیرآستانه آن برابر ۱ می‌شود.

پس از محاسبه ماتریس دقت، مدل گراف با استفاده از فرمول ۴ ایجاد می‌شود.

(۴) محاسبه وزن و جهت یال در گراف

$$F_i \to F_j = P\{(F_j, C) \mid (F_i, C)\} = \frac{C_{ij}}{C_i}$$

فرمول ۴، با استفاده از احتمال شرطی، وزن و جهت یال از ویژگی i به ویژگی j را به دست می‌آورد.

مقدار C نشان‌دهنده کلیه کلاسهایی ( ردیفهای ماتریس دقت، بیان گر کلاسها و یا همان نوع حملات می‌باشد.) است که متغیر زیرآستانه متناظر با آنها در ماتریس دقت، به ازای ویژگی انتخاب شده، دارای مقدار یک هستند. یک بودن متغیر زیرآستانه، میزان اهمیت بالای ویژگی را نشان می‌دهد. به همین ترتیب مقدار $C_i$ نشان دهنده تعداد کلاسهایی است که به ازای ویژگی i برابر یک بوده است و مقدار $C_{ij}$، تعداد کلاسهایی (ردیفهایی) است که به ازای هر دو ویژگی i و j یک می باشد. درنهایت طبق تعریف احتمال شرطی، جهت یال بین دو ویژگی i و j طبق فرمول ۴ به دست می‌آید. خروجی این مرحله، ماتریس گراف می باشد که f ستون و f ردیف دارد و مقدار خانه (i , j) ، وزن یال جهت دار از ویژگی i به ویژگی j را نشان می‌دهد.

چنانچه تنها از ردیفهایی از ماتریس دقت که مربوط به یک نوع حمله خاص است استفاده شود، مدل گراف مربوط به هرنوع حمله نیز به دست می‌آید که این مدل گراف به دست آمده برای استخراج مجموعه ویژگی بهینه مرتبط با هرنوع حمله استفاده می‌گردد.

در گراف ویژگی‌ها، وجود یال، میزان شباهت اهمیت حضور دو ویژگی را نشان می‌دهد.

### ۴-۲- ایجاد مدل سلسله مراتبی !

پس از ایجاد مدل گراف و تحلیل آن با استفاده از اندازه‌گیریهای معرفی شده در بخش ۳ و روش [۱۳]، مدل گراف به مدل سلسله مراتبی تبدیل می‌شود. پیش از ایجاد مدل سلسله مراتبی، به منظور کاهش هزینه زمانی محاسبه مدل سلسله مراتبی، ویژگی‌هایی که یال ارتباطی بین آنها به وزن یک می‌باشد، از گروه ویژگی‌ها حذف می‌شوند و پس از اتمام ایجاد مدل سلسله مراتبی، دوباره به مدل اضافه می‌گردند. علت حذف این ویژگی‌ها در این مرحله این است که این ویژگی‌ها کاملا از نظر اهمیت به هم مشابه هستند و تمام اندازه‌گیری مرکزیتهای آنها یکسان به دست می‌آید.

برای ایجاد مدل سلسله مراتبی، از قواعد زیر استفاده می‌شود:

- در گراف ویژگی‌ها، اگر یالی از گره i به گره j وجود داشته باشد، گره i، به عنوان پدر گره j انتخاب می‌شود.
- هیچ گرهی دارای دو پدر نمی‌تواند باشد. بنابراین تحت شرایطی که از دو گره i,k به گره j یالی وجود داشته باشد، بین دو گره i,k رقابت ایجاد می‌شود. در چنین شرایطی از مرکزیت وابستگی استفاده می‌شود، اگر وابستگی ویژگی i به ویژگی j بیشتر از وابستگی k به ویژگی j باشد، گره i به عنوان پدر گره j انتخاب و یال بین گره k و i حذف می‌شود.

درمدل سلسله مراتبی، ویژگی‌هایی که در بالاترین سطح قرار گرفته‌اند، مهمترین ویژگی‌ها و ویژگی‌های واقع شده در پایینترین سطح، ویژگی‌های غیرمهم و یا نامربوط هستند.

### ۴-۳- تحلیل زمان و حافظه!

زمان اجرای الگوریتم شامل دو قسمت محاسبه ماتریس دقت و محاسبه مدل سلسله مراتبی است. برای ایجاد ماتریس دقت، به ازای هر روش دسته‌بندی، مرحله حذف هریک از ویژگی‌ها

انجام می‌شود. مرحله حذف هر ویژگی و سپس اجرای روش دسته‌بندی، هزینه زمانی در فرمول ۵ نشان داده شده است. در این فرمول f تعداد ویژگی‌ها و C هزینه زمانی روش دسته بندی است.

$$(F+1)\sum C_i \qquad (۵)$$

برای محاسبه مرکزیتها هزینه زمانی $f^2$ صرف می‌شود. از بین این مرکزیتها، تنها محاسبه مرکزیت وابستگی به دلیل یافتن k تا کوتاهترین مسیرهای بین دو گره طبق مقاله [۱۶]، برابر $kf^5$ می باشد.

مقدار حافظه مصرفی این روش برای نگهداری ماتریس دقت در فرمول ۶ بیان شده است.

$$memForMatrix = n*m*f*doubleSize \qquad (۶)$$

در فرمول ۶، n تعداد روشهای کلاس‌بندی، m تعداد انواع حملات، f تعداد ویژگی‌ها و doubleSize سایز عدد از نوع double می‌باشد.

مقدار حافظه‌ای که درطی اجرای الگوریتم لازم است مطابق فرمول ۷ می‌باشد.

$$mem = O(n*m*doubleSize + C_i) \qquad (۷)$$

در فرمول ۷، C مقدار حافظه مصرفی روش دسته‌بندی C را نشان می‌دهد.

### ۴-۴- مزایا و معایب !

برای این روش می‌توان مزایا و معایبی برشمرد که در ادامه به آنها اشاره می‌شود.

مزایای روش ارائه شده عبارت است از :

این روش قابل استفاده بر روی انواع داده عددی و غیرعددی می‌باشد. زیرا برای استخراج مدل گراف ویژگی‌ها، مبتنی بر روشهای کلاس‌بندی می‌باشد.

این روش می‌تواند نحوه ارتباط بین ویژگی‌ها، میزان وابستگی ویژگی‌ها و رتبه بندی ویژگی‌ها براساس اهمیت آنها را ارائه کند.

قادر به یافتن مجموعه ویژگی بهینه مرتبط با هر نوع حمله می‌باشد.

این روش می‌تواند ویژگی‌های مشابه از نظراهمیت حضور را می‌یابد.

معایب این روش عبارت است از :

این روش قادر به یافتن ویژگی‌های افزونه نمی‌باشد.

تحلیلهای بخش ۵ نشان می‌دهد دقت این روش درحد مناسبی است اما این دقت در برخی حملات کمتر از روشهای دیگر می‌باشد.

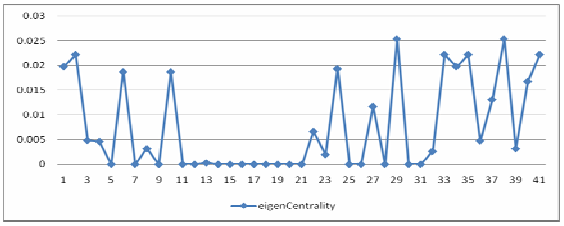
تحلیلهای بخش ۵ نشان می‌دهد مجموعه ویژگی مرتبط با هر نوع حمله، تقریبا مجموعه‌های بزرگی هستند.

## ۵- ارزیابی روش پیشنهادی!

برای ارزیابی روش پیشنهادی از مجموعه داده KDD99 [۱۷] که مجموعه داده جمع‌آوری شده برای تست سیستمهای تشخیص نفوذ می‌باشد استفاده شده است. این مجموعه داده دارای ۴ میلیون رکورد با ۴۱ ویژگی می‌باشد، که از بین آنها، ۱۱۰۰۰رکورد با توزیع مناسب انتخاب شده‌اند. الگوریتم پیشنهادی با استفاده از نرم‌افزار weka و محاسبه مرکزیتها در محیط Matlab پیاده‌سازی شده است.

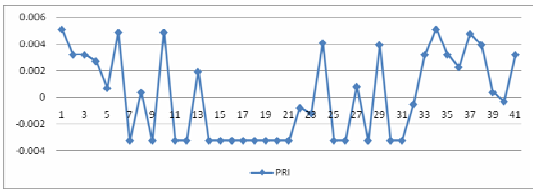
### ۵-۱- نتایج پیاده سازی!

نمودارهای مرکزیت مقادیر مشخصه و نمایه نقش سازمانی در شکل ۲ و ۳ نشان داده شده است. طبق مفهوم مرکزیت مقادیرمشخصه، گره‌های با بالاترین مقدار مشخصه، مهم‌ترین ویژگی محسوب می‌شوند.




شکل ۲. مرکزیت مقادیر مشخصه

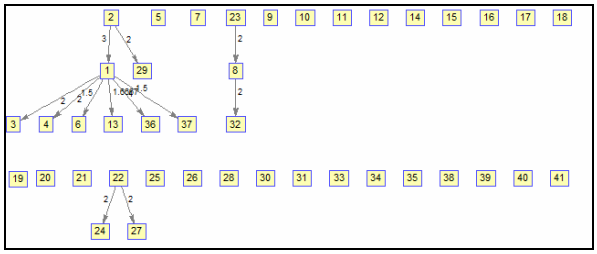
در نمودار نمایه نقش سازمانی، گره‌های با مقادیر مثبت بیان‌کننده گره‌های با اهمیت و رئوس با مقادیر منفی، سربازان و یا همان گره‌های کم اهمیت می‌باشند.




شکل۳. نمایه نقش سازمانی

در دو نمودار مرکزیت مقادیر مشخصه و نمایه نقش سازمانی، ویژگی‌های با اهمیت مشابه دارای مقادیر یکسان می‌باشند.

پس از محاسبه مرکزیتها، با استفاده از روش توضیح داده شده در بخش ۴-۲، مدل سلسله مراتبی ویژگی‌ها طبق شکل ۴ به دست می‌آید.



شکل ۴. مدل سلسله مراتبی

در مدل سلسله مراتبی حاصل، سه ویژگی ۲، ۲۲ و ۲۳ در سطح اول قرار دارند، بنابراین، این سه ویژگی مهمترین ویژگی‌ها می‌باشند. با توجه به اینکه ویژگی۲ از نظر اهمیت مشابه ویژگی‌های ۳۳، ۳۵ و ۴۱ می‌باشد، درنتیجه ۶ ویژگی به عنوان مهمترین ویژگی‌های معرفی شده توسط روش پیشنهادی می‌باشد.

## ۵-۲- ارزیابی نتایج!

پس از محاسبه مدل سلسله مراتبی و به دست آوردن مجموعه ویژگی بهینه، لازم است این مجموعه بهینه حاصل با مجموعه ویژگی بهینه روشهای دیگر مقایسه شود. برای این مقایسه، مجموعه ویژگی‌های بهینه روشهای مختلف به روش دسته‌بندی j48 داده می‌شود. سپس این روش دسته‌بندی، کل مجموعه داده منتخب از KDD99 را تنها درحضور مجموعه ویژگی بهینه، دسته بندی می‌کند. درنهایت دقت دسته‌بندی‌های حاصل باهم مقایسه می‌شوند. ابتدا مجموعه ویژگی‌های بهینه روشهای مختلف ذکر می‌شود.

روش بیضین :۱ ، ۲، ۳، ۵، ۷، ۸، ۱۱، ۱۲، ۱۴، ۱۷، ۲۲، ۲۳، ۲۴، ۲۵، ۲۶، ۳۰ و ۳۲.

روش CART : ۳، ۵، ۶، ۱۲، ۲۳، ۲۴، ۲۵، ۲۸، ۳۱، ۳۲، ۳۳ و ۳۵.

روش Rough-PSO : این روش قادر است مجموعه ویژگی بهینگی مرتبط با هرنوع حمله را بیابد. اما هزینه اجرای این روش از نوع نمایی است.

داده نرمال : ۱۲، ۳۱، ۳۲، ۳۳، ۳۵، ۳۶، ۳۷ و ۴۱.

حمله وارسی : ۲، ۳، ۲۳، ۳۴، ۳۶ و ۴۰.

حمله عـدم پـذیرش سـرویس : ۵، ۱۰، ۲۴، ۲۹، ۳۳، ۳۴، ۳۸ و ۴۰.

حمله کاربر به ریشه : ۳، ۴، ۶، ۱۴، ۱۷ و ۲۲.

حمله دور به محلی:۳، ۴، ۱۰، ۲۳، ۳۳ و ۳۶.

مجموعه ویژگی بهینه کلی عبارت است از : ۲، ۴، ۲۴، ۲۷، ۳۴ و ۳۵.

روش SVDF : ۲، ۴، ۵، ۲۳، ۲۴ و ۳۳.

برنامه خطی ژنتیک LGP : ۳، ۵، ۱۲، ۲۷، ۳۱ و ۳۵.

روش پیشنهادی: این روش قادر است مجموعه ویژگی بهینه مرتبط با هرنوع حمله را نیز بیابد و هزینه آن توانی است. درنتیجه به نسبت روش Rough-PSO بهتر است.

داده نرمـــال :۱، ۲، ۴، ۶، ۱۰، ۲۹، ۳۲، ۳۳، ۳۴، ۳۵، ۳۶، ۳۸، ۳۹، ۴۱.

حملـه وارسـی : ۱، ۳، ۶، ۱۰، ۱۳، ۲۴، ۲۹، ۳۴، ۳۶، ۳۷، ۳۸ و ۴۰.

حمله عدم پـذیرش سـرویس : ۶، ۱۰، ۱۳، ۲۲، ۲۴، ۲۷ و ۳۶.

حمله دور به محلی:۱، ۳، ۴، ۵، ۱۳، ۲۷، ۳۴، ۳۶، ۳۷.

همانطور که مشاهده می‌شود تعداد اعضای مجموعه ویژگی بهینه معرفی شده توسط روش پیشنهادی برابر تعداد اعضای معرفی شده توسط دیگر روشها می‌باشد.

برای مقایسه روشهای ذکرشده، دو مجموعه داده که یکی دارای ۵۰۰۰ رکورد و دیگری دارای ۱۱۰۰۰ رکورد می‌باشد استفاده شده است.

نتایج حاصل از اجرای j48 در جدول ۱ آمده است.

جدول ۱. مقایسه روشهای کاهش ویژگی

| class | LGP | | SVDF | | Rough-POS | | My Algorithm | |
|---|---|---|---|---|---|---|---|---|
| Data | 5000 | 11000 | 5000 | 11000 | 5000 | 11000 | 5000 | 11000 |
| Normal | 0.986 | 0.987 | 0.994 | 0.994 | 0.983 | 0.986 | 0.955 | 0.97 |
| Prob | 0.952 | 0.96 | 0.95 | 0.965 | 0.889 | 0.942 | 0.937 | 0.944 |
| DOS | 0.998 | 0.997 | 0.992 | 0.991 | 0.971 | 0.974 | 0.979 | 0.978 |
| U2R | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| R2L | 0.463 | 0.798 | 0.39 | 0.607 | 0 | 0.048 | 0.516 | 0.642 |

همانطور که در جدول ۱ قابل مشاهده است روش پیشنهادی در مقایسه با برخی روشها و به ازای برخی حملات دقت بیشتری داشته است و در باقی موارد دقتی مناسبی را ارائه کرده است.

## ۶- جمع‌بندی و کارهای آتی!

هدف این تحقیق، یافتن مجموعه ویژگی بهینه برای سیستم-های تشخیص نفوذ بوده است. با توجه به اینکه سیستم های تشخیص نفوذ، با داده های حجیم برای تحلیل مواجه هستند، کاهش مجموعه داده می‌تواند راه حل مناسبی برای افزایش دقت و سرعت تشخیص آنها باشد. نیازمندی دیگری که در سیستم‌های تشخیص نفوذ مطرح می‌باشد، دانستن مجموعه

ویژگی بهینه برای هر نوع حمله است. چرا که در اینصورت، سیستم تشخیص نفوذ قادر خواهد بود برای تشخیص هر نوع حمله، تنها از مجموعه ویژگی متناسب با آن حمله استفاده کند. روش ارائه شده در تحقیق حاضر، قادر به پاسخگویی به تمام نیازمندیهای مطرح شده می‌باشد. علاوه براین ، روش پیشنهادی نحوه ارتباط ویژگی‌ها و رتبه‌بندی آنها براساس میزان اهمیت را نیز بیان می‌کند. این روش برای پاسخگویی به نیازهای مذکور، از روشهای داده‌کاوی و تحلیل شبکه‌های اجتماعی بهره می‌برد. در این مقاله، روشهای موجود برای انتخاب ویژگی ذکر شدند که هرکدام مزایا و معایبی دارند. تنها برخی از آنها قادر به رتبه بندی ویژگی‌ها از نظر اهمیت هستند و هیچ یک قادر به بیان نحوه ارتباط ویژگی‌ها نمی‌باشند. تنها روش Rough-PSO می‌تواند مجموعه بهینه مرتبط با هرنوع حمله را بیان کند که هزینه اجرای آن نمایی است. درنتیجه روش ارائه شده در این مقاله می‌تواند روش مناسبی برای انتخاب مجموعه بهینه محسوب شود.

ازجمله کارهایی که در ادامه این تحقیق می‌توان انجام داد، یافتن ویژگی‌های افزونه می‌باشد. چرا که روش پیشنهادی قادر به بیان آنها نمی‌باشد. به علاوه، می‌توان قابلیتهایی را به روش پیشنهادی اضافه نمود که قادر باشد ارتباطات خطی و یا غیرخطی ویژگی‌ها را نیز نشان بدهد.

## مراجع

[1] A. Hassan, M.S. Nabi Baksh, A.M. Shaharoun, and H.Jamaluddin. "Improved SPC Chart Pattern Recognition Using Statistical Feature."International Journal of Production Research 41(7), 2003, pp. 1587-1603.

[2] A.H. Sung and S. Mukkamala. “Identifying Important Features for Intrusion Detection Using Support Vector Machines and Neural Networks." In Proceedings of the Symposium on Applications and Internet (SAINT’03), pp. 209-216.

[3] F Sung,A.H. and Mukkamala, S.: Identifying Important Features for Intrusion Detection Using Support Vector Machines and Neural Networks. Proceedings of the 2003 Symposium on Applications and the Internet (SAINT’03).(2003) 209-216.

[4] Li, J., Zhang, G.Y and Gu, G.C.: The Research and Implementation of Intelligent Intrusion Li, J., Detection System Based on Artificial Neural Network. IEEE Proceedings of the 3rd Zhang, G.Y International Conference on Machine Learning and Cybernetics. (2004) 3178-3182.

[5] Zhang, C., Jiang, J. and Kamel, M.: Intrusion Detection using Hierarchical Neural Networks. Pattern Recognition Letters Vol. 26 (2005) 779-791.

[6] Xu, X. and Wang, X.: An Adaptive Network Intrusion Detection Method Based on PCA and Support Vector Machines. Proceedings of First International Conference on Advanced Data Mining and Applications ADMA:, Wuhan, China, Volume 3584 (2005) 696-703.

[7] Gao, H., Yang H. and Wang, X.: Kernel PCA Based Network Intrusion Feature Extraction and Detection Using SVM. Lecture Notes in Computer Science, Vol. 3611. Springer- Verlag, Berlin Heidelberg New York (2005) 89-94.

[8] Chebrolu, S., Abraham, A. and Thomas, JP.: Feature Deduction and Ensemble Design of Intrusion Detection Systems. Journal of Computers and Security. Vol 24, Issue 4 (2005) 295-307.

[9] A. Zainal, M.A. Maarof and S.M. Shamsuddin, "Feature Selection Using Rough-DPSO in Anomaly Detection" LNCS 4705, Part 1 Springer Hiedelberg 2007, pp. 512-524.

[10] Jennifer G. Dy , Carla E. Brodley, Feature Selection for Unsupervised Learning, The Journal of Machine Learning Research, 5, p.845-889, 2004.

[11] Mohsen Jafari Asbagh, Hassan Abolhassani, "Feature-Based Data Stream Clustering," icis, pp.363-368, 2009 Eigth IEEE/ACIS International Conference on Computer and Information Science (icis 2009), 2009.

[12] M. Dash, K. Choi, P. Scheuermann, and H. Liu, "Feature selection for clustering – A filter solution", In Proceeding of the 2002 IEEE International Conference on Data Mining(ICDM'02) December 2002, pp. 115-124.

[13] Nasrullah Memon and Henrik Legind Larsen: Investigative Data Mining Toolkit: A Software Prototype for Visualizing, Analyzing and Destablizing Terrorist Networks. (pp. 14-1 – 14-24). Meeting Proceedings RTO-MP-IST-063, Paper 14. Neuilly-sur-Seine, France: RTO, 2006

[14] Memon Nasrullah and Henrik Legind Larsen.: Practical Approaches for Analysis, Visualization and Destabilizing Terrorist Networks. In the proceedings of ARES 2006: The First International Conference on Availability, Reliability and Security, Vienna, Austria, IEEE Computer Society, pp. 906-913,2006.

[15] Memon Nasrullah and Henrik Legind Larsen: Practical Algorithms of Destabilizing Terrorist Networks. In the proceedings of IEEE Intelligence Security Conference, San Diego, Lecture Notes in Computer Science, Springer-Verlag, Vol. 3976: pp. 398-411 , 2006.

[16] J.Y. yen. Finding thr k shortest loopless paths in a network. Management Science, 17:712 – 716, 1971.

[17] KDD Cup 1999.Available on: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, Ocotber 2007.