اولین همایش ملی پیشرفت های تکنولوژی در مهندسی برق، الکترونیک و کامپیوتر
First National Conference of Technology Developments on Electronical, Electronics and Computer Engineering
. . . W W W . T D E C O N F . I R . . .

بهره‌گیری از تکنیک‌های خوشه بندی داده‌کاو محور جهت افزایش تشخیص نفوذ در سیستم‌های اطلاعاتی ابری

فاطمه اسحقی[۱]، مهدی خلیلی[۲]، مجید انجیدنی[۳]

دانشگاه پیام نور مرکز تهران شمال[۱] ، دانشگاه پیام نور مرکز تهران شمال[۲]، دانشگاه پیام نور مرکز نیشابور[۳]

مسئول مکاتبات: فاطمه اسحقی

چکیده

با گسترش روز افزون تبادل اطلاعات و استفاده از سیستم‌های بر خط، میزان حملات و نفوذ در سیستم‌های اطلاعاتی افزایش یافته است. رایانش ابری گام بعدی تکامل سرویس های فناوری اطلاعات برحسب تقاضا می‌باشد. فایروال ها که توسط بسیاری از سازمان ها، به عنوان سیستم های تشخیص نفوذ به منظور حفاظت از امنیت سیستم های اطلاعاتی به کار گرفته می شوند اغلب در تشخیص حملاتی که از درون این سازمان ها اتفاق می افتد با شکست مواجه می‌شوند.

در این مقاله، به منظور غلبه بر این مشکل در فایروال‌ها، استفاده از تکنیک‌های خوشه‌بندی داده‌کاو محور شامل تکنیک‌های خوشه‌بندی K-Means، Y-Means و Fuzzy C-Means جهت برعهده گرفتن مسئولیت رسیدگی به نفوذ از درون سازمان‌ها، ارتقاء سیستم‌های تشخیص نفوذ و افزایش سطح امنیت اطلاعات در ابرها پیشنهاد می گردد تا نقشی حیاتی در تشخیص نفوذ با استفاده از تجزیه و تحلیل حجم بزرگ داده‌های شبکه و دسته‌بندی آن ها به صورت عادی و یا غیرعادی در فایروال‌ها بازی کند. همچنین با توجه به اینکه پیش‌تر تکنیک‌های داده‌کاوی با موفقیت برای تشخیص نفوذ در حوزه‌های کاربردی مختلف از جمله بیوانفورماتیک، بازار سهام، تجزیه و تحلیل وب و غیره مورد استفاده قرارگرفته است؛ از این روش استخراج روابط قبلی و ناشناخته در پایگاه داده‌های بزرگ، الگوبرداری نموده و سپس از الگوهای استخراج شده به عنوان پایه‌ای برای شناسایی حملات جدید استفاده خواهیم نمود.

کلمات کلیدی

رایانش ابری- حملات امنیتی به ابر، سیستم‌های تشخیص نفوذ، داده‌کاوی، سیستم‌های تشخیص نفوذ داده‌کاو محور، الگوریتم K-Means، الگوریتم Y-Means، الگوریتم Fuzzy C-Means.

۱- مقدمه

با پیشرفت فناوری اطلاعات نیاز به انجام کارهای محاسباتی در همه جا و همه زمان و همچنین نیاز به این که افراد بتوانند کارهای محاسباتی سنگین خود را بدون داشتن سخت‌افزارها و نرم‌افزارهای گران، از طریق خدماتی انجام دهند، به وجود آمده است که رایانش ابری آخرین پاسخ فناوری به این نیازها می‌باشد (آپارنا اس. وارد، ۲۰۰۹). با تکیه بر تکنولوژی ابر، میلیون‌ها نفر از کاربران داده‌های خود را در یک ابر که دارای فضای زیادی است ذخیره می‌کنند. ذخیره سازی در ابرها خطرات بسیاری نظیر دسترسی غیرمجاز، از دست دادن داده‌ها و غیره را به همراه دارد براساس بررسی های انجام شده در سال ۲۰۱۰ میلادی، امنیت به عنوان مهمترین چالش رایانش ابری شناخته شده است (دی. جی. برآون، بی. سوکو و تی. وانگ، ۲۰۱۰). جدیدترین بررسی ها در سال ۲۰۱۱ حاکی از کاهش ۳۳/۵ درصدی چالش امنیت و رسیدن این درصد به عدد ۵۵ می‌باشد (جی. هویس منز، بی. بیسنز، دی. مارتنیز، کی. دنیس و جی. وانسینن، ۲۰۱۱).

حفظ حریم خصوصی داده‌ها عمده نگرانی افرادی است که از خدمات ابرهای عمومی استفاده می‌کنند. تهدیدهای امنیتی بر روی کابران ابر به دو دسته داخلی و خارجی تقسیم بندی می شود. تهدیدهای خارجی شامل تهدید مراکز داده بزرگ می باشد که این نگرانی امنیتی در میان کاربران ابر و فراهم آورندگان (اشخاص ثالث) در حصول اطمینان از نرم افزارهای امن موجود، امکان پذیر است. علاوه بر مسائل مربوط به امنیت خارجی، ابر دارای برخی از مسائل مربوط به امنیت داخلی نیز می باشد که در آن کاربران باید در مقابل حملات از طرف یکدیگر، محافظت شوند. (انجمن کامپیوتر IEEE، ۲۰۱۰).

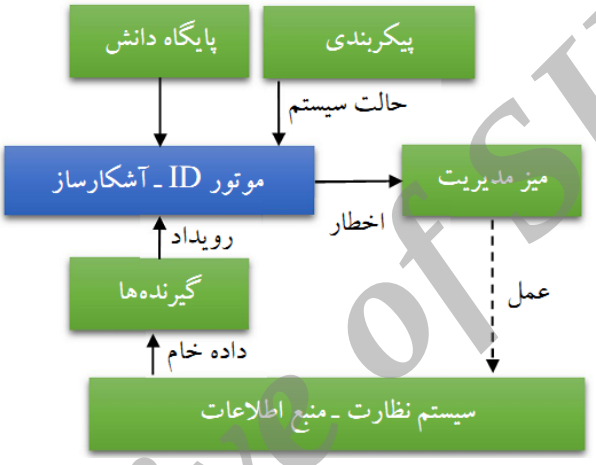
شکست در امنیت رایانش ابری به دلایل: الف- سخت افزاری و در لایه زیرساخت به عنوان سرویس ابر، ب- نفوذ کدهای مخرب در نرم‌افزار و در لایه نرم‌افزار کاربردی و ج- نفوذ کدهای مخرب در حال اجرا، توسط برنامه کاربردی کاربر یا تزریق اطلاعات ساختگی به برنامه توسط شخص ثالث؛ رخ می‌دهد که راه حل‌های: ۱- تنظیم تغییر رمز عبورکاربران به صورت اجباری در زمان‌های مشخص. ۲- پشتیبان‌گیری از داده ها در فواصل منظم ۳- تست نفوذ در سیستم در بازه‌های زمانی منظم ۴- نصب فایروال و ضدویروس‌های بروز شده ۵- استفاده از ناوگانی از سرورها که هر سرور جهت انجام کار خاصی در نظر گرفته شده برای رفع این مشکلات مورد استفاده قرار می‌گیرد (دیپسی کی. دناشز و آنیتا جان، ۲۰۱۳).

در این مقاله و به منظور غلبه بر این مشکل در فایروال‌ها، استفاده از تکنیک های خوشه بندی داده کاو محور شامل تکنیک های خوشه‌بندی K-Means، Y-Means و Fuzzy C-Means جهت برعهده گرفتن مسئولیت رسیدگی به نفوذ از درون سازمان‌ها، ارتقاء سیستم‌های تشخیص

نفوذ و افزایش سطح امنیت اطلاعات در ابرها پیشنهاد می گردد تا نقشی حیاتی در تشخیص نفوذ با استفاده از تجزیه و تحلیل حجم بزرگ داده‌های شبکه و دسته‌بندی آن ها به صورت عادی و یا غیرعادی در فایروال‌ها بازی کند. همچنین با توجه به اینکه پیش‌تر تکنیک‌های داده‌کاوی با موفقیت برای تشخیص نفوذ در حوزه‌های کاربردی مختلف از جمله بیوانفورماتیک، بازار سهام، تجزیه و تحلیل وب و غیره مورد استفاده قرارگرفته است؛ از این روش استخراج روابط قبلی و ناشناخته در پایگاه داده‌های بزرگ، الگوبرداری نموده و سپس از الگوهای استخراج شده به عنوان پایه‌ای برای شناسایی حملات جدید استفاده خواهیم نمود.

۲- نفوذ و سیستم تشخیص نفوذ

یک سیستم تشخیص نفوذ (IDS)، یک دستگاه یا برنامه نرم‌افزاری است که بر شبکه یا فعالیت های سیستم جهت فعالیت های مخرب یا نقض سیاست‌ها نظارت می‌کند و تولید گزارش‌ها برای یک ایستگاه مدیریت را بر عهده دارد که در شکل ۱ نشان داده شده است. (تشخیص نفوذ، ۱۹۹۳)



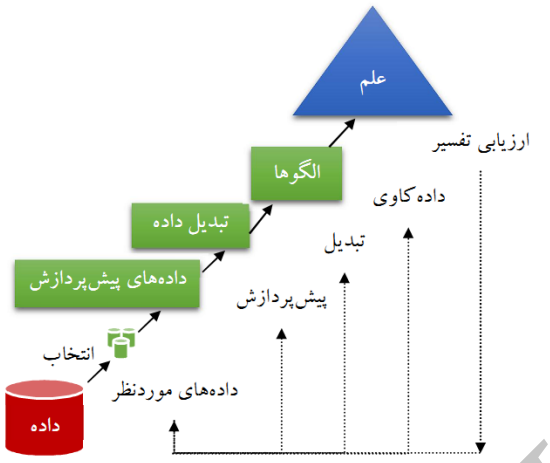
شکل ۱- مدل **IDS**

روش های متعددی در پیاده‌سازی یک ID براساس روش IDS (سیستم تشخیص نفوذ) وجود دارند:

- **تشخیص ناهنجاری:** این روش براساس تشخیص ناهنجاری ترافیکی است.
- **سوءاستفاده/تشخیص امضاء:** این روش الگوها و امضاهای حملات شناخته شده پیشین در ترافیک شبکه را جستجو می‌کند. سبک این روش با تشخیص نفوذ سر و کار دارد که شبیه به روش نرم‌افزار ضدویروس عمل می‌کند.
- **تشخیص ترکیبی:** این روش ترکیبی از هر دو روش تشخیص ناهنجاری و تشخیص امضاء را بکار می‌برد.

۳- داده‌کاوی و سیستم تشخیص نفوذ

همانگونه که در شکل ۲ نشان داده شده است؛ داده‌کاوی، فرایند استخراج الگوها از داده‌ها می باشد. داده‌کاوی به عنوان یک ابزار مهم رو به افزایش برای کسب و کار مدرن که اطلاعات را به آگاهی شغلی تبدیل می‌کند، یک برتری اطلاعاتی را نمایش می‌دهد. در حال حاضر داده‌کاوی در طیف گسترده‌ای از شیوه‌های پروفایل، مانند بازاریابی، نظارت، تشخیص تقلب و اکتشاف علمی به کار می‌رود. (سی. یو. پی. مجموعه داده ها، ۲۰۱۰) دلیل اولیه برای استفاده از داده‌کاوی کمک به تجزیه و تحلیل مجموعه مشاهدات رفتار است. یک واقعیت غیرقابل اجتناب داده‌کاوی این است که تجزیه و تحلیل زیرمجموعه داده‌ها ممکن است نماینده کل دامنه و بنابراین شامل نمونه‌هایی از روابط بحرانی و رفتاری خاص که در دیگر بخش‌های دامنه وجود دارد، نباشد. (یان فاستر، یونگ ژائو، ایوان رایکو و شیانگ لو، ۲۰۰۸).

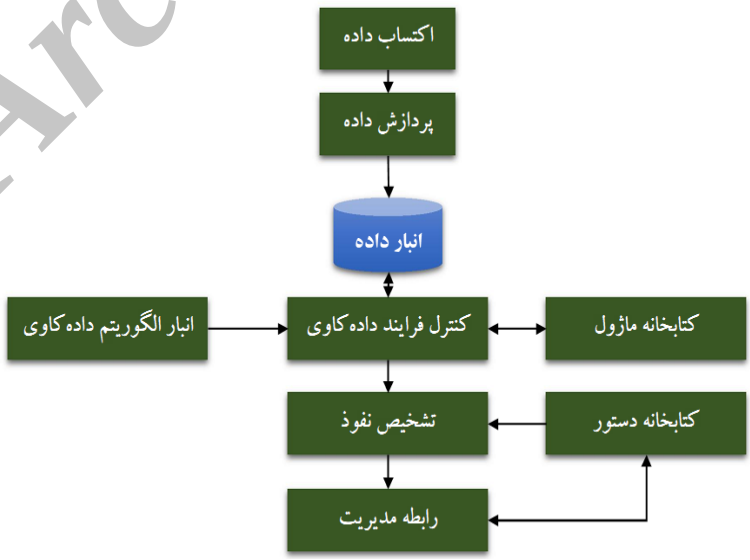


شکل ۲ـ مفهوم داده‌کاوی

چارچوب داده‌کاوی به طور خودکار الگوهای موجود در مجموعه داده‌های ما را تشخیص می‌دهد و از این الگوها برای پیدا کردن مجموعه‌ای از فایل‌های دودویی مخرب استفاده می‌کند. تکنیک‌های داده‌کاوی می‌تواند الگوهای مقادیر زیادی از داده‌ها را تشخیص دهد (مانند بایت کُد) و از این الگوها برای تشخیص داده‌های مشابه در نمونه‌های آینده استفاده کند. در سیستم تشخیص نفوذ، اطلاعات از منابع مختلف مانند: داده‌های میزبان، اطلاعات ورود به سیستم شبکه، پیام های هشدار و غیره به دست می‌آید. (داده کاوی "، کارگاه آموزشی بین المللی IEEE دوم در فناوری آموزش و پرورش و سیستم های تشخیص در پایگاه داده اوراکل ۱۰g، 1998)

تکنولوژی داده‌کاوی قابلیت استخراج پایگاه داده‌های بزرگ را دارد و از اهمیت زیادی برای استفاده در تشخیص نفوذ برخوردار است. با به کاربردن تکنولوژی داده‌کاوی، سیستم تشخیص نفوذ می‌تواند به طور گسترده‌ای داده‌ها را برای به دست آوردن یک مدل بررسی کند، بدین گونه برای مقایسه بین الگوی غیرطبیعی و الگوی رفتار طبیعی کمک می‌کند. تجزیه و تحلیل دستی برای این روش لازم نیست (اوما شانکار، کانیکا لاخانی و منیش موندرا، ۲۰۱۰).

یک مسأله مهم در تشخیص نفوذ این است که چگونه می‌توان به طور مؤثر الگوهای حمله و الگوهای داده‌های طبیعی را از تعداد زیادی از داده‌های شبکه جدا کرد و چگونه می‌توان به طور مؤثر قوانین نفوذ خودکار را پس از جمع‌آوری داده‌های خام شبکه به وجود آورد. برای انجام این کار، از تکنیک‌های داده‌کاوی مختلفی مانند: طبقه‌بندی، دسته‌بندی و استخراج قانون رابطه استفاده می‌شود. در شکل ۳، یک سیستم که براساس الگوی تشخیص نفوذ طراحی گردیده است را نشان می دهد. (دیپسی کی. دناشز و آنیتا جان، ۲۰۱۲).



شکل ۳- داده‌کاوی براساس سیستم تشخیص نفوذ

داده کاوی براساس سیستم تشخیص نفوذ به دانش کمتر کارشناس نیاز دارد با این وجود اجرای خوب و امنیت را تأمین می‌کند. این سیستم‌ها به صورت حملات شناخته‌شده به اندازه حملات ناشناخته شبکه قابل تشخیص هستند. تکنیک‌های مختلف داده‌کاوی می تواند شبیه رده بندی، خوشه بندی و کاوش قواعد وابستگی، برای تحلیل ترافیک شبکه و در نتیجه تشخیص نفوذها استفاده شود. از میان موارد فوق، الگوریتم‌های خوشه‌بندی در بسیاری از جاها برای تشخیص نفوذ استفاده می‌شوند چون آنها به رده بندی داده آموزش به صورت دستی نیاز ندارند (میشل ام. جی.، ۲۰۱۳).

اکثر نفوذها برای حمله به سیستم های مقصد از راه شبکه با استفاده از پروتکل‌های شبکه اتفاق می‌افتد. این نوع ارتباطات به عنوان ارتباطات غیرعادی (ناهنجار) و سایر ارتباطات، به عنوان ارتباطات عادی طبقه‌بندی می‌شوند. کلاً چهار نوع حمله به شرح ذیل وجود دارد (یه چینگ، وو شیائوپنگ و هوانگ گاوفنگ، ۲۰۱۲):

۱) **DOS – عدم قبولی سرویس:** حمله‌کننده سعی می‌کند مانع دسترسی کاربران مجاز به خدمات در رایانه شود. برای مثال: فرمان مرگ، طغیان SYN و غیره

۲) **بررسی – نظارت و کاوش:** حمله‌کننده یک شبکه را برای کشف آسیب‌پذیری‌های شناخته شده رایانه آزمایش می‌کند. بررسی‌های این شبکه به طور معقول برای حمله‌کننده ای که یک حمله را در آینده برنامه‌ریزی می‌کند، باارزش هستند. برای مثال: درگاه اسکن، فرمان پاکسازی و غیره

۳) **R2L – کنترل از راه دور به محلی:** حمله‌کننده‌های غیرمجاز، دسترسی محلی رایانه را از یک ماشین کنترل از راه دور به دست می‌آورند و سپس از آسیب‌پذیری‌های ماشین استفاده می‌کنند. برای مثال: حدس رمز ورود به سیستم.

۴) **U2R – کاربر با ریشه:** به ماشین اخیراً حمله شده، ولی حمله‌کننده تلاش می‌کند تا دسترسی با امتیازات کاربر بسیاری را به دست آورد. به عنوان مثال، حافظه موقت از حملات پُر می‌شود.

۴- تکنیک‌های خوشه‌بندی داده کاومحور و سیستم تشخیص نفوذ

در این بخش، چند الگوریتم خوشه‌بندی برای تشخیص نفوذ استفاده شده است. همه این الگوریتم‌ها میزان خطای واقعی (مثبت‌های کاذب) را کاهش و میزان تشخیص نفوذ را افزایش می‌دهند. میزان تشخیص مانند تعداد نمونه‌های تشخیص نفوذ تعریف می‌شود تا سیستم تقسیم‌شده توسط جمع کل تعداد نمونه‌های تشخیص نفوذ در مجموعه داده را نشان دهد. میزان خطای واقعی به صورت تعداد کل نمونه‌های عادی که به اشتباه به صورت نفوذهای تعریف‌شده با تعداد کل نمونه‌های عادی، تعریف می‌شود. تعدادی از تکنیک‌های خوشه‌بندی نظیر الگوریتم‌های خوشه‌بندی داده کاو محور k-Means، Y-Means و Fuzzy C-Means در ذیل بحث می‌شوند.

۴-۱ الگوریتم خوشه بندی K-Means

الگوریتم k-Means، الگوریتمی است که به سختی قسمت‌بندی شده و در بسیاری جاها به سادگی و با سرعت به کار می‌رود. این الگوریتم فاصله اقلیدسی را با اندازه مشابه استفاده می‌کند. مفهوم خوشه بندی سخت این است که یک آیتم در یک مجموعه داده می‌تواند متعلق به یکی و فقط یک خوشه در یک زمان باشد. این الگوریتم، یک تحلیل خوشه‌بندی است به طوری که آیتم‌های گروه‌ها براساس مقادیر مشخص در داخل K به خوشه‌هایی که آیتم‌ها در همان خوشه ویژگی‌های مشابه دارند ملحق نمی‌شوند چون آن ها در خوشه‌های متفاوت، مشخصه‌های مختلفی دارند. عملکرد فاصله اقلیدسی با محاسبه فاصله بین دو آیتم ارائه شده به صورت:

$$d(p,q) = \sqrt{(p1 - q1)^2 + (q2 - q2)^2}$$

استفاده می‌شود.

در اینجا؛ P=(P1,P2,…Pm) و q=(q1,q2,…qm) دو بردار ورودی با مشخصه‌های کمیت m هستند. الگوریتم اجرا می‌شود تا مجموعه‌های آموزش داده که ممکن است شامل ترافیک عادی و یا غیرعادی شود بدون اینکه قبلاً برچسب خورده باشد.

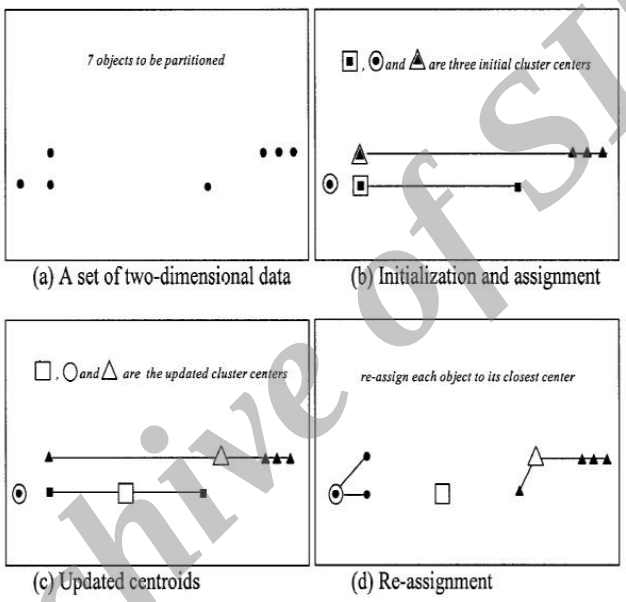
ایده اصلی این دیدگاه براساس فرضیه‌ای است که ترافیک عادی و غیرعادی از خوشه‌های مختلف می‌باشد. همچنین داده ممکن است شامل بخش های مجزایی شود که کدام آیتم‌های داده از آیتم‌های دیگر در خوشه خیلی متفاوت هستند و کدام متعلق به هیچ خوشه‌ای نیستند (جیبان کی. پال، ۲۰۱۱). یک بخش مجزا با مقایسه شعاع آیتم‌های داده درک می‌شود به طوری که اگر شعاع یک آیتم داده بزرگتر از یک آستانه ارائه‌شده باشد، پس به صورت یک بخش مجزا دیده می‌شود. اما این موضوع، پردازش خوشه بندی K-means را از وقتی که تعداد بخش مجزا کوچک است، برهم نمی‌زند (ونکاتادری ام. لوکاناسا سی. ردی، ۲۰۱۳).

الگوریتم خوشه‌بندی K-means به صورت زیر است:

۱. تعریف تعداد خوشه‌های K. برای مثال، اگر K=2 باشد ما در آموزش داده از دو خوشه متفاوت، ترافیک عادی و غیرعادی را فرض می‌کنیم.
۲. مقداردهی اولیه مراکز ثقل خوشه K . این کار با انتخاب اتفاقی آیتم‌های داده K از مجموعه داده انجام می‌شود.

۳. محاسبه فاصله هر آیتم تا مراکز ثقل همه خوشه‌ها با به‌کارگیری فاصله اقلیدسی متریک که برای یافتن شباهت بین آیتم‌ها در مجموعه داده به کار می‌رود.
۴. اختصاص هر آیتم با نزدیک‌ترین مرکز ثقل خوشه. در این روش همه آیتم‌ها به خوشه‌های مختلف اختصاص خواهند یافت، به طوری که هر خوشه آیتم‌هایی با ویژگی‌های مشابه خواهد داشت.
۵. پس از اختصاص یافتن همه آیتم‌ها به خوشه‌های مختلف، میانگین خوشه‌های تغییریافته مجدداً محاسبه می‌شود. میانگین اخیراً محاسبه شده به عنوان مرکز ثقل جدید اختصاص می‌یابد.
۶. تکرار مرحله ۳ تا زمانی که مرکز ثقل خوشه تغییر نکند.
۷. برچسب زدن خوشه‌ها به صورت عادی و غیرعادی، به تعداد آیتم‌های داده هر خوشه بستگی دارد.

یک مشکل اساسی روش خوشه‌بندی K-means، تعیین بخش اولیه و تعداد مناسب خوشه‌های K می‌باشد. همچنین گاهی اوقات منجر به هم‌ترازی می‌شود که میانگین کدام پردازش خوشه‌بندی ممکن است با چند خوشه خالی به پایان برسد. در شکل ۴، تولید خوشه‌های خالی را نشان داده شده است.



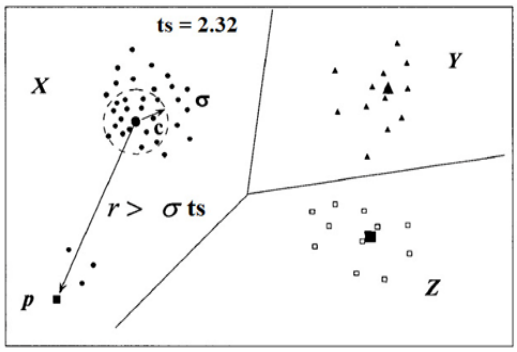
شکل ۴- تولید خوشه‌های خالی

۴-۲ الگوریتم خوشه بندی Y-Means

Y-Means، الگوریتم دیگری از خوشه‌بندی است که برای تشخیص نفوذ استفاده می‌شود. این تکنیک به طور خودکار یک مجموعه داده را درون تعداد قابل قبولی از خوشه‌ها قسمت‌بندی می‌کند به طوری که آیتم‌های داده درون خوشه‌های عادی و غیرعادی رده‌بندی می‌شوند. امتیاز اصلی الگوریتم خوشه‌بندی Y-Means آن است که به سه کمبود الگوریتم K-Means یعنی وابستگی مراکز ثقل اولیه، تعداد خوشه‌ها و هم‌ترازی برتری دارد. خوشه‌بندی Y-Means موانع خوشه‌های خالی را رفع می‌کند.

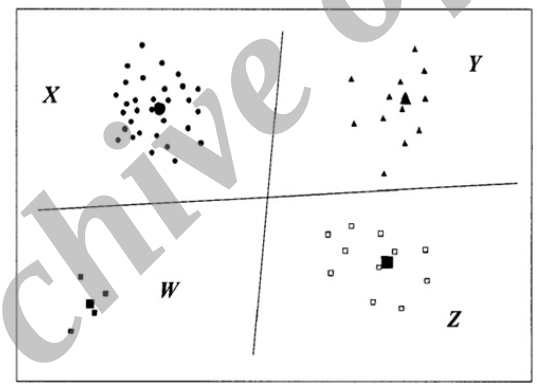
تفاوت اصلی بین خوشه‌بندی‌های Y-Means و K-Means تعداد خوشه‌های (k) در Y-Means است (یک خودتعریف متغیر به جای کاربرتعریف ثابت). اگر میزان K خیلی کوچک باشد، Y-Means تعداد خوشه‌ها را با چنددسته‌ای کردن آن ها افزایش می‌دهد. از طرف دیگر، اگر میزان K خیلی بزرگ باشد، Y-Means تعداد خوشه‌ها را با ادغام کردن در نزدیکی آن ها کاهش می‌دهد. Y-Means میزان مناسبی از K را به وسیله چنددسته‌ای کردن و ارتباط بین خوشه‌ها حتی بدون داشتن هیچ دانشی از آیتم توزیع‌شده تعیین می‌کند. این کار باعث می‌شود Y-Means یک تکنیک خوشه بندی موثر برای تشخیص نفوذ باشد چون ثبت وقایع شبکه داده به صورت اتفاقی توزیع شده و به دست آوردن میزان K به صورت دستی کار مشکلی است. Y-Means فاصله اقلیدسی را برای ارزیابی تشابه بین دو آیتم در مجموعه داده استفاده می‌کند. خوشه‌بندی سه مرحله اصلی دارد:

- **مرحله اول- تخصیص آیتم‌ها به خوشه‌های K**: وابستگی میزان K مشخص شده توسط کاربر، آیتم‌هایی را در یک مجموعه داده به نزدیک‌ترین خوشه‌های وابسته به فاصله بین آیتم و مرکز ثقل هر خوشه اختصاص می‌دهد.



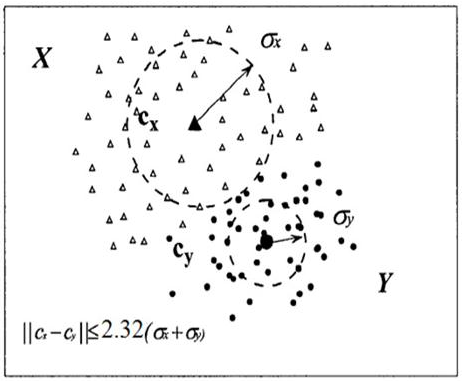
شکل ۵- تخصیص آیتم‌ها به خوشه‌های K

- **مرحله دوم- چنددسته‌ای کردن خوشه‌ها:** برطبق عملکرد توزیع عادی جمعی استانداردشده، ۹۹٪ نمونه‌های خوشه در داخل شعاع دایره σ 2.32 قرار دارد جایی که σ انحراف استاندارد داده می‌باشد. بنابراین آستانه t= 2.32 σ انتخاب می‌شود. منطقه درون دایره، منطقه مطمئن خوشه نامیده می‌شود. بدین ترتیب، همه نقاط در خوشه که خارج از منطقه مطمئن قرار دارند، به صورت بخش‌های مجزا در نظر گرفته می‌شوند. سپس این بخش‌های مجزا از خوشه‌های فعلی دور شده و به مراکز ثقل جدید اختصاص می‌یابند. مراکز ثقل جدیداً شکل یافته ممکن است چند آیتم را از خوشه‌های مجاور و به موجب آن از خوشه‌های جدید جذب کنند. چنددسته‌ای کردن خوشه‌ها تا وقتی که هیچ بخش مجزایی وجود نداشته باشد ادامه پیدا می‌کند. چند دسته ای کردن، خوشه‌ها را به شاخه‌های کوچک‌تر تبدیل، تعداد خوشه‌ها را افزایش و سپس آیتم‌های درون همان خوشه را به یکدیگر شبیه‌تر می‌سازد.



شکل ۶- چنددسته‌ای کردن خوشه‌ها

- **مرحله سوم**- پیوند بین خوشه‌ها: وقتی دو خوشه نزدیک به هم یک اشتراک دارند، آنها می‌توانند درون یک خوشه بزرگ‌تر ادغام شوند. ادغام آستانه، مجموعه‌ای است با σ 2.32 یعنی هرجایی که آن ها هستند هیچ آیتم داده‌ای در یک منطقه مطمئن نیست، همچنین اگر در منطقه مطمئن خوشه دیگری قرار گیرد، دو خوشه می‌توانند ادغام گردند. وقتی دو خوشه ادغام شدند مراکز ثقلشان کامل و سالم مانده و مرکز ثقل جدیدی ایجاد نمی‌شود یعنی خوشه جدید دو مرکز ثقل دارد. مزیت آن است که خوشه‌ها می‌توانند از شکل‌های دلخواه نظیر شکل‌های زنجیره مانند باشند.



شکل ۷- پیوند بین خوشه‌ها

الگوریتم خوشه‌بندی Y-means به صورت زیر است:

۱. تعریف تعداد خوشه‌های K
۲. مقداردهی اولیه مراکز ثقل خوشه K. این کار توسط انتخاب اتفاقی آیتم‌های K از مجموعه داده انجام می‌شود.
۳. محاسبه فاصله هر آیتم تا مراکز ثقل همه خوشه‌ها با استفاده از فاصله اقلیدسی متریک که برای یافتن شباهت بین آیتم‌ها در مجموعه داده به کار می‌رود.
۴. اختصاص هر آیتم با نزدیک‌ترین مرکز ثقل خوشه. در این روش همه آیتم‌ها به خوشه‌های مختلف اختصاص خواهند یافت، به طوری که هر خوشه آیتم‌هایی با ویژگی‌های مشابه خواهد داشت.
۵. پس از اختصاص یافتن همه آیتم‌ها به خوشه‌های مختلف، میانگین خوشه‌های تغییریافته مجدداً محاسبه می‌شود. میانگین اخیراً محاسبه شده به عنوان مرکز ثقل جدید اختصاص می‌یابد.
۶. کنترل کنید آیا هم ترازی وجود دارد ؟ اگر پاسخ بله است خوشه‌های خالی را حذف کرده و به مرحله ۷ بروید.
۷. اگر پاسخ خیر است، بخش‌های مجزا را پیدا کنید. اگر در هیچ یک از خوشه‌ها پیدا نشدند پس خوشه‌ها را چند دسته ای کنید.
۸. کنترل کنید آیا هم ترازی وجود دارد ؟ اگر پاسخ بله است خوشه‌های خالی را حذف کرده و به مرحله ۹ بروید.
۹. اگر پاسخ خیر است، خوشه‌های دارای اشتراک را لینک کنید.
۱۰. خوشه‌ها را به صورت عادی و غیرعادی برچسب بزنید.

۳-۴ الگوریتم خوشه بندی Fuzzy C-Means

FCM یک الگوریتم بازبینی‌نشده خوشه بندی براساس مجموعه تئوری fuzzy است که اجازه می‌دهد یک عامل متعلق به بیش از یک خوشه باشد. درجه ارتباط هر آیتم داده با خوشه، محاسبه شده و تصمیم می‌گیرد که کدام خوشه با کدام آیتم داده وابسته شود. برای هر آیتم، یک عامل مشترک داریم که مشخصه‌های درجه عضویت (uij) از بودن در خوشه Kth به صورت

$$uij = \sum_{k=1}^{n} (dij/dik)^{(2im-1)}$$

می‌باشد. در اینجا؛ dij – فاصله آیتم ith از خوشه jth ، dik – فاصله آیتم ith از خوشه kth و m- عامل نادقیق‌سازی.

وجود یک آیتم داده در بیش از یک خوشه وابسته به مقدار m یعنی نادقیق‌سازی تعریف شده توسط کاربر در حد [0,1] می‌باشد که درجه نادقیقی را در خوشه تعیین می‌کند. بدین‌ترتیب، آیتم‌های روی گوشه یک خوشه ممکن است در خوشه‌ای با درجه کمتر از آیتم‌ها در مرکز خوشه باشد. وقتی m به مقدار ۱ برسد الگوریتم شبیه یک الگوریتم قسمت‌بندی حلقوی کار می‌کند و برای مقادیر بزرگ‌تر از m خوشه‌های متمایل به بیشتر بودن به اشتراک گذاشته می‌شوند.

هدف اصلی الگوریتم خوشه‌بندی fuzzy قسمت‌بندی داده درون خوشه‌هاست به طوری که تشابه آیتم‌های داده درون هر خوشه افزایش یافته و تشابه آیتم‌های داده در خوشه‌های مختلف کم می‌شود. به علاوه، این خوشه‌بندی، کیفیت قسمت‌بندی را طوری اندازه‌گیری می‌کند که یک مجموعه داده درون خوشه‌های C تقسیم شوند. الگوریتم FCM به کاهش میزان عملکرد عینی به شرح $J(U,V) \sum_{i=1}^{n} \sum_{j=1}^{c} (uij)^m :: xi - vj ::^2$ توجه می‌کند. در اینجا؛ m – هر شماره واقعی بیش از ۱ باشد، uij – درجه ارتباط xi در خوشه j ، ith – xi از داده اندازه‌گیری شده ابعادی، vj – فاصله مرکز ابعاد از خوشه و $||*||^2$ – هر مقیاس که بیان‌کننده تشابه بین هر داده اندازه‌گیری شده و مرکز است.

الگوریتم FCM به شرح زیر است:

۱. انتخاب به صورت اتفاقی مراکز خوشه cll -

۲. مقداردهی اولیه عضویت fuzzy ماتریکس uij با استفاده از:

$$uij = \sum_{k=1}^{n} (dij/dik)^{(2im-1)}$$

۳. محاسبه مراکز ,vj در fuzzy با استفاده از:

$$\forall j = 1,2, \ldots . c$$
$$vj = (\sum_{i=1}^{n} (uij)mxi) / \sum_{i=1}^{n} (uij)m),$$

۴. به روزکردن ماتریس عضویت یعنی به مرحله ۲ بروید.

۵. اگر ||u(k+1)-u(k)|| کوچکتر از آستانه باشد پس پایان کار و اگر نباشد به مرحله ۳ برگردید. جایی که k مرحله تکراری است.

۵- مقایسه الگوریتم های خوشه‌بندی داده کاومحور

این بخش مقایسه سه تکنیک خوشه‌بندی یعنی K-Means، Y-Means و Fuzzy C-Means را نشان می‌دهد. مقایسه با محاسبه معیار مختلف شبیه اجرا، کارایی، میزان تشخیص، میزان خطای واقعی، خالص بودن و ... انجام می‌شود. هر تکنیک تعدادی خصوصیت خوب برای رفع موانع سایر تکنیک‌ها دارد.

جدول ۱. مقایسه تکنیک‌های خوشه‌بندی‌های K-Means، Y-Means و Fuzzy C-Means

ورودی	تعداد خوشه‌های K درصورتی که K کوچکتر از m باشد. مجموعه آیتم‌های داده (X1,X2,… Xm)	تعداد خوشه‌های Y درصورتی که Y کوچکتر از m باشد. مجموعه آیتم‌های داده (X1,X2,… Xm)	تعداد خوشه‌های C درصورتی که C کوچکتر از m باشد. مجموعه آیتم‌های داده (X1,X2,… Xm) مجموعه مراکز خوشه (V1,V2,… VC)
خروجی	مجموعه خوشه‌های K، جایی که هرخوشه آیتم‌های مشابه دارد.	مجموعه‌ای از خوشه‌های غیرخالی، جایی که هرخوشه آیتم‌های مشابه دارد.	مجموعه‌ای از خوشه‌های C، جایی که هرخوشه آیتم‌های مشابه بیشتری دارد.
میزان عضویت	وجود ندارد.	وجود ندارد.	یک میزان عضویت به صورت "uij" دارد.
زمان محاسبه	ساده و مستقیم به جلو به طوری که به زمان کمتری نیاز دارد.	چنددسته‌ای شدن و ارتباط خوشه‌ها را شامل می‌شود، به طوری که به زمان بیشتری نیاز دارد.	محاسبه چند فرمول را شامل می‌شود، به طوری که به زمان بیشتری نیاز دارد.
خالص بودن خوشه	پایین	بالا	بالا
تولید خوشه خالی	ممکن است تولید داشته باشد شاید هم خیر	خیر	خیر
کارایی/بازده	برای مجموعه داده کوچک خوب کارمی‌کند.	برای مجموعه داده‌های کوچک به اندازه مجموعه داده‌های بزرگ خوب کارمی‌کند.	برای مجموعه داده‌های کوچک به اندازه مجموعه داده‌های بزرگ خوب کارمی‌کند.
تعداد خوشه‌های وابسته به یک آیتم	یک	یک	یک یا بیشتر از یک خوشه

اجرای کامل	به تعداد اولیه خوشه‌های K بستگی دارد.	به تعداد اولیه خوشه‌های K بستگی ندارد.	به تعداد اولیه خوشه‌های C بستگی دارد.
شکل خوشه	برای فشردن و کروی نمودن خوشه‌ها خوب کار می‌کند.	هم برای کروی کردن و هم غیر کروی کردن خوشه‌ها خوب کار می‌کند.	هم برای کروی کردن و هم غیر کروی کردن خوشه‌ها خوب کار می‌کند.
میزان تشخیص	بالاترین	بالاتر	بالا
میزان خطای واقعی	پایین‌ترین	پایین	پایین‌تر
امتیازات	ساده و سریع کارها برای فشردن و کروی نمودن خوشه‌ها خوب انجام می‌شود.	هم‌ترازی نیست. به تعداد اولیه خوشه‌های K بستگی ندارد. با خوشه‌های کروی و غیرکروی خوب کار می‌کند.	کیفیت قسمت‌بندی اندازه‌گیری‌می‌شود. آیتم‌ها می‌توانند متعلق به بیش از یک خوشه باشند. برای خوشه‌های کروی و غیرکروی خوب کار می‌کند.
معایب	نیاز به تعیین تعداد مناسبی از خوشه‌ها دارد. هم‌ترازی با خوشه‌های غیرکروی خوب کار نمی‌کند. کیفیت خوشه‌ها اندازه‌گیری نمی‌شود.	کیفیت خوشه‌ها اندازه‌گیری نمی‌شود. زمان‌بر بودن	اجرا به تعداد اولیه خوشه‌ها بستگی دارد. زمان‌بر بودن

۶- بحث و نتیجه‌گیری

همان طور که می‌دانید، روش های تشخیص سوءاستفاده برای شناسایی حملات ناشناخته کافی نیست لذا برای تشخیص نفوذ ناشناخته، ما باید به سمت تشخیص ناهنجاری برویم. تکنیک‌های مختلف داده‌کاوی شبیه رده بندی، خوشه‌بندی و کاوش قواعد وابستگی در تحلیل شبکه داده بسیار مفید است و از آن جایی که میزان وسیعی از ترافیک شبکه برای تشخیص نفوذ نیاز به جمع‌آوری دارد، پس خوشه‌بندی مناسب‌تر از رده‌بندی در دامنه تشخیص نفوذ می‌باشد، به طوری که جهت برچسب زدن مجموعه داده به صورت دستی، به کم کردن تلاش نیاز ندارد (کیزی، آونورهین، سوزان و وبسکی، ۲۰۰۹).

تکنیک‌های داده‌کاوی می‌تواند حملات شناخته شده را به خوبی حملات ناشناخته تشخیص دهند. فناوری داده‌کاوی به درک رفتار عادی در یک طرف داده و استفاده از این دانش برای تشخیص نفوذهای ناشناخته کمک می‌کند. در این مقاله، سه الگوریتم خوشه‌بندی داده کاومحور یعنی K-Means، Y-Means و Fuzzy C-Means بحث و بررسی شده‌اند. هر یک از آنها هم امتیازات و هم معایبی دارند و همدیگر را اصلاح می‌کنند:

الف- الگوریتم خوشه‌بندی k-Means در هم‌ترازی نتیجه می‌دهد و برای پایگاه‌های داده بزرگ مناسب نیست.

ب- الگوریتم خوشه بندی Y-Means انجام یک تغییر بر روی k-Means است که خوشه‌بندی‌های خالی را حذف می‌کند.

ج- الگوریتم خوشه بندی Fuzzy C-Means برپایه یک منطق نامعلوم است که اجازه می‌دهد یک آیتم به بیش از یک خوشه بند وابسته باشد و روی به حداقل رساندن تابع هدف که کیفیت قسمت‌بندی را بازرسی می‌کند، تمرکز می‌کند. اجرا و کارآیی خوشه‌بندی Fuzzy C-Means در دوره‌های تشخیص نفوذ بهتر از دو تکنیک دیگر است. در انتها، سه الگوریتم خوشه‌بندی را مقایسه می‌کنیم.

از میان خوشه بندی‌ها، Fuzzy C-Means را می‌توان یک الگوریتم مؤثر برای تشخیص فرض کرد چون اجازه می‌دهد یک آیتم متعلق به بیش از یک خوشه باشد و همچنین کیفیت قسمت‌بندی را اندازه می‌گیرد. تکنیک می‌تواند برای مجموعه داده بزرگ به خوبی مجموعه‌های داده که آیتم های اشتراکی دارند، استفاده شود. به علاوه، هیچ خوشه‌خالی را ایجاد نمی‌کند و در هنگام ایجاد خوشه بالاترین خلوص را دارد.

مزیت اصلی خوشه‌بندی Fuzzy C-Means برای تشخیص نفوذ، پیشنهاد میزان تشخیص بالا و میزان خطای واقعی پایین‌تر است. اگرچه Fuzzy C-Means یک تکنیک مؤثر است ولی زمان بر می‌باشد. اجرای سیستم‌های تشخیص نفوذ می‌تواند با ترکیب ویژگی های تکنیک خوشه‌بندی Fuzzy C-Means با چند تکنیک دیگر پیشرفت کند به طوری که زمان مورد نیاز با Fuzzy C-Means را برای پردازش خوشه‌بندی کاهش داده، میزان تشخیص را افزایش و همچنین میزان خطای واقعی را کم کند. در نتیجه سیستم تشخیص نفوذ را با دقت‌تر و موثرتر سازد.

ترکیب دسته‌بندی K-Means و درخت تصمیم C4.5 بهترین نرخ تشخیص (۹۹/۶) و کمترین نرخ خطای واقعی (۰/۱) را ارائه می‌دهد، اما حملات را به انواع مختلف طبقه‌بندی نمی‌کند. در حالی که آبشار K-Means و دو طبقه‌بندی: KNN و ساده بیز؛ نرخ تشخیص ۹۸/۱۸ و نرخ خطای

واقعی ۰/۸۳۰ را ارائه می‌دهد و همچنین حملات مختلف را به صورت عادی، DOS، U2R، R2L و کاوشگر طبقه‌بندی می‌کند. داده‌کاوی روش مدرنی برای تشخیص نفوذ شبکه است. الگوریتم‌های داده‌کاوی ساخته شده ـ آماده شده، در دسترس هستند. مقدار زیادی از اطلاعات را می‌توان با فن‌آوری داده‌کاوی اداره کرد. این روش هنوز در حال توسعه است و می‌تواند به طور مؤثرتری به سرعت در حال رشد باشد. وظیفه اصلی ما، رسیدن به نرخ صحیح‌تر تشخیص نفوذ برای کاهش نرخ خطای واقعی است. داده‌کاوی هنوز هم در حال توسعه است بنابراین مطالعه و تحقیقات بیشتری باید انجام شود.

۷- منابع مورداستفاده

- Aparna S. Varde “Challenging research issues in data mining, databases and information retrieval” ACM SIGKDD Explorations Newsletter Volume 11 Issue 1, June 2009 Pages 49-52.
- CUP 99 Data Set", Proceedings of the 2009 IEEE Symposium on Computational 1, pp 695-698, 2010.
- D. J. BROWN, B. SUCKOW and T. WANG, "A Survey of Intrusion Detection Systems", and SVM", 2010.
- Data Mining",IEEE Second International Workshop on Education Technology and Detection Systems in Oracle Database 10g", Proceedings of the Fourth International detection systems: The 1998 darpa off-line intrusion detection evaluation,” discex, vol.02.
- Deepthy K Denatious and Anita John, ―Survey on Data Mining Techniques to Enhance Intrusion Detection‖, International Conference on Computer Communication and Informatics, Jan 2012.
- Denatious, D.K., and John, A. 2013. Survey on Data Mining Techniques to Enhance Intrusion Detection.
- Ian Foster, Yong Zhao, Ioan Raicu, Shiyong Lu. "Cloud Computing and Grid Computing 360-Degree", IEEE Conference, Date of Conference: 12-16 Nov. 2008.
- IEEE Computer Society, 2010 Sixth International Conference on Semantics, Knowledge and Grids, Security and Privacy in Cloud Computing: A Survey, 2010.
- INTRUSION DETECTION", National Research Council of Canada, NRC 45842, May 4-7 items in large databases", In Proceedings of the 1993 ACM SIGMOD Conference, New Machine and Decision Tree", International Journal of Computer Applications (0975 – 8887),
- J. HUYSMANS, B. BAESENS, D. MARTENS, K. DENYS and J. VANTHIENEN “New Trends in Data Mining” Tijdschrift voor Economie en Management Vol. L, 4, 2011.
- Jiban K Pal “Usefulness and applications of data mining in extracting information from different perspectives” Annals of Library and Information Studies Vol. 58, March 2011, pp. 7-16.
- Kazi, Aunnurhain, Susun, Vrbsky, Security Attacks and Solutions in Clouds,University of Albama,2009.
- Michael, M. J., A view Of Cloud Computing, Communications of the ACM, April 2013.
- Uma Somani, Kanika Lakhani, Manish Mundra, Implementing Digital Signature with RSA Encryption to Enhance Data Security of Cloud in Cloud Computing, IEEE, 2010.
- Venkatadri. M, Lokanatha C. Reddy “A Review on Data mining from Past to the Future” International Journal of Computer Applications (0975 – 8887) Volume 15– No.7, February 2013-19.
- Ye Qing, Wu Xiaoping and Huang Gaofeng. An Intrusion Detection Approach based on Data Mining‖, 2nd International Conference on Future Computer and Communication, pp. No. 695 – 698, 2012 IEEE.