استفاده از الگوریتم K-means برای افزایش کارایی سیستم های تشخیص نفوذ

شادی لنگری
عضو هیات علمی، گروه کامپیوتر ، موسسه آموزش عالی دانشگاه اشراق بجنورد،ایران،
shadilangari@gmail.com
بهنام اسدی
دانشجوی کارشناسی ارشد نرم افزار ، موسسه آموزش عالی دانشگاه اشراق بجنورد،ایران،
b.asadi_en@yahoo.com
افشین رجبی
دانشجوی کارشناسی ارشد نرم افزار ، موسسه آموزش عالی دانشگاه اشراق بجنورد،ایران،
afshin.ra57@gmail.com
سیاوش کنعانی
دانشجوی کارشناسی ارشد نرم افزار ، موسسه آموزش عالی دانشگاه اشراق بجنورد،ایران،
s.kanani2010@gmail.com

چکیده

مشکل مشترک در IDS های کنونی نرخ بالای تشخیص اشتباه و نرخ شناسایی درست پایین است. یک یادگیری ماشینی بدون نظارت با استفاده از K ابزار برای ارائه ی مدلی برای سیستم های تشخیص نفوذ (IDS) با نرخ بهره وری بالاتر و مثبت های کاذب کمتر و منفی های کاذب مورد استفاده قرار گرفت. مجموعه ی داده ی NSL-KDD که شامل ۵۰۰۰ ورودی با ۱۰ نوع مختلف داده بود مورد استفاده قرار گرفت. نتایج این مطالعه با استفاده از ۱۱، ۲۲، ۱۲، ۱۴ و ۲۰ خوشه به ترتیب نرخ بازده ی ۷۰,۷۵٪، ۸۱,۶۱٪، ۶۵,۴۰٪، ۶۱,۳۰٪ و ۵۵,۴۳٪ را نشان داد. نرخ مثبت کاذب به ترتیب ۰,۷۴٪، ۴,۰۳٪، ۱۵,۵۵٪، ۲۱,۴۷٪ و ۳۱,۹۱٪ و نرخ منفی کاذب ۹۹,۸۲٪، ۹۸,۱۴٪، ۹۷,۷۶٪، ۹۶,۳۲٪ و ۹۵,۷۰٪ بود. جالب است که بهترین نتایج زمانی بدست آمد که تعداد خوشه ها منطبق با تعداد انواع داده در مجموعه ی داده بود.

کلمات کلیدی: داده کاوی ،خوشه بندی ، یادگیری ماشین ، یادگیری بدون نظارت ، K-means

۱– مقدمه

آخرین تحولات در سیستم های کامپیوتری و اینترنت شیوه ی تفکر افراد و طریقه ی انجام کارها را متحول کرده است. فرایندی شبیه ارسال پست های سنتی که به طور معمول ساعت ها و یا حتی روزها طول می کشد حال می تواند با کلیک بر روی موس یا لمس انگشت از طریق پست الکترونیکی و یا ایمیل انجام شود. مردم از طریق چت های رله یکپارچه یا کنفرانس ویدئویی به عنوان یک حالت راحت ارتباطی با یکدیگر ارتباط برقرار می کنند. با این حال، همراه با بسیاری از پیشرفت ها در سیستم های کامپیوتری و زیرساخت های IT ریسک هایی در ارتباط با استفاده از این فن آوری ها وجود دارد. در طول دو دهه گذشته، تهدیدهای رایانه ای و جرایم اینترنتی به ضرر عموم مردم گسترش یافته و تهدیدات جدیدتری هر روز معرفی می شوند که یکپارچگی، اعتبار و محرمانه بودن اطلاعات را به خطر می اندازند. شرکت ها، ملت ها و افراد می توانند قربانی فعالیت های مخرب در اینترنت باشند. به عنوان یک نتیجه از جرایم اینترنتی، میلیون ها دلار صرف استراتژی های کاهش آنها شده است. افرادی که از آسیب پذیری سیستم های اطلاعات سوء استفاده می کنند اغلب در استفاده از تکنیک های برنامه ریزی پیچیده ماهر بوده و از اتصال درونی سیستم ها به شیوه ای استفاده می کنند که حتی نیازی به دسترسی محلی به شبکه ندارند زیرا می توانند حملات خود را از راه دور انجام دهند. فعالیت های مخرب در اینترنت همچنین به عنوان نفوذ شناخته می شوند. نفوذ فعالیتی است که سیاست امنیت شبکه را نقض می کند. سیستم تشخیص نفوذ (IDS) یک نرم افزار و سخت افزار مستقر شده برای اجرای تشخیص استفاده غیر مجاز از، و یا حمله به یک کامپیوتر و یا شبکه مخابراتی است که باید شکاف های بین فایروال و آنتی ویروس ها را پر کند. یک IDS نظارت و تحلیل کاربر و فعالیت سیستم فراهم کرده، می تواند پیکربندی و آسیب پذیری های سیستم را ممیزی کند، یکپارچگی سیستم حیاتی و فایل های را ارزیابی کرده و همچنین تجزیه و تحلیل آماری از الگوهای فعالیت بر اساس تطبیق با حملات شناخته شده فراهم کرده، فعالیت های غیر طبیعی را تحلیل کرده و ممیزی سیستم انجام دهد. یکی از مزایای IDS توانایی مستند سازی نفوذ یا تهدید به سازمان است که به موجب آن اساسی برای اطلاع رسانی به عموم در مورد آخرین الگوهای حمله از طریق لوگ های سیستم فراهم می کند.

انواع حملات کامپیوتری شناسایی شده توسط IDS به سه دسته طبقه بندی می شوند: (۱) حملات اسکن کننده، (۲) حملات انکار خدمات (DOS)، (۳) حملات نفوذی. هر کدام از این سه طبقه بندی حملات کامپیوتری ویژگی ها و رفتارهای متفاوتی دارند که IDS برای تحلیل، شناسایی و هشداردهی در هنگام برخورد با آنها طراحی شده است. هنگامی که یک آلارم تنظیم می شود، مدیران شبکه باید لوگ ها را تحلیل کنند تا تصمیم بگیرند که آیا فعالیت مشکوک در واقع غیر عادی است یا نه. با این حال، در بیشتر IDS ها مثال های زیادی از مثبت کاذب و منفی کاذب وجود دارد که می تواند برای مدیران شبکه دست و پا گیر باشد.

مثبت کاذب یک نمونه است که در آن IDS یک فعالیت خوش خیم را به اشتباه به عنوان مخرب شناسایی می کند، در حالی که منفی کاذب زمانی رخ می دهد که IDS نمی تواند فعالیت مخرب را تشخیص دهد. در حین عملیات عادی، یک IDS می تواند هزاران نادرست در روز ایجاد کند. سیستم های تشخیص نفوذ شبکه – مهم نیست اگر مبتنی بر ناهنجاری باشند یا مبتنی بر امضا– یک مشکل مشترک دارند: تعداد زیاد هشدارهای کاذب یا مثبت های کاذب. تعداد هشدارهای جمع آوری شده توسط IDS می تواند تا ۱۵۰۰۰ هر روز در هر سنسور باشد و تعداد مثبت های کاذب (FP) می توانند هزاران مورد در روز باشد. این مشکلات معمولا منجر می شوند کاربر نهایی، مدیر امنیت، اعتماد خود را به هشدارها از دست بدهد و سطح دفاعی را به منظور کاهش تعداد مثبت های کاذب کاهش داده یا کار بیش از حد برای شناسایی حملات واقعی به دلیل اشتباهات IDS داشته باشد. این مقاله استفاده از یادگیری ماشین و الگوریتم داده کاوی K ابزار را برای توسعه ی یک مدل IDS با رخ بهره وری بالاتر و آلارم کاذب پایین تر پیشنهاد می کند.

۱- بررسی ادبیات مربوط

مقالات پژوهشی زیادی با استفاده از مجموعه داده ی KDD برای توسعه ی مدل هایی برای سیستم های شناسایی نفود انجام گرفته است. اگرچه، بحث های زیادی وجود دارد که آیا مجموعه ی داده در واقع یک سند خوب یا معتبر به عنوان اساس مدل های پیشنهاد برای سیستم های شناسایی نفوذ هست یا نه، این واقعیت که هیچ مجموعه داده ی دیگری برای چنین هدفی وجود ندارد موجب می شود به طور گسترده ای مورد استفاده بوده و برای آزمایش پذیرفته شده باشد.

۲- روش پژوهش

۱) مجموعه ی داده ی NSL-KDD

این مجموعه ی داده ۲۵۱۹۲ ورودی و ۴۳ ویژگی دارد که از بین آنها ۴۱ مورد مشابه KDD’99 بوده، ویژگی ۴۲ برچسب داده و ویژگی ۴۳ سطح دشواری است. ۲۲نوع مختلف از داده وجود دارد: (۱) عادی، (۲) عقبی، (۳) جریان بیش از حد بافر، (۴) حدس پسوورد، (۵) imap، (۶) ipsweep، (۷) multihop (۸) نپتون، (۹)nmap ، (۱۰) phf، (۱۱) pod، (۱۲) portsweep، (۱۳) rootkit، (۱۴) satan، (۱۵) smurf، (۱۶) teardrop، (۱۷) warezclient، (۱۸) warezmaster، (۱۹) ftp_write، (۲۰) load_module، (۲۱) land و (۲۲) spy.

۲) پیش پردازش

پیش پردازش شامل پاک کردن داده ها از تناقضات و/یا نویز و ترکیب یا حذف ورودی های زائد است. پیش پردازش همچنین شامل تبدیل ویژگی های مجموعه ی داده به داده های عددی و ذخیره ی آنها در فرمتی است که قابل خواندن باشد زیرا K ابزار تنها در داده های عددی کار می کند. داده های الفبایی به ارزش های عددی تبدیل شدند که از ۰,۰۰۱، ۰,۰۰۲ و غیره شروع می شد. ارزش های کوچک تر (به جای ۱، ۲ و غیره) برای اطمینان از عدم تاثیرگزاری روی محاسبات مورد استفاده قرار گرفتند.

خوشه بندی K-means

K-means یک تکنیک مبتنی بر مرکز بوده و ساده ترین و اساسی ترین خوشه بندی توسط پارتیشن بندی است که در آن آیتم ها در K پارتیشن ($k \leq n$) قرار می گیرند. K ابزار به طور خاص برای شناسایی دور افتاده ها به کار می رود زیرا زمانی که ارزشی وجود داشته باشد که دورتر از اکثر داده ها باشد، میانگین خوشه به طور قابل توجهی تحریف خواهد شد. این مطالعه از خوشه بندی K ابزار به عنوان روشی برای شناسایی دور افتاده ها استفاده می کند. در این مدل شناسایی دور افتاده، فرض می شود که الگوی رفتار نرمال بسیار مکرر تر از رفتار دور افتاده ها یا غیر عادی ها است.

$$E = \sum_{i=1}^{K} \sum_{p \in c_i} dist(p, c_i)^2$$

۱ - فرمول خوشه بندی K ابزار

که در آن:

E: مجموع خطای مربع تمامی آیتم ها در مجموعه ی داده است.

P: نقطه ای در فضا است که نشان دهنده ی یک آیتم داده شده می باشد.

این الگوریتم برای پارتیشن بندی است که در آن مرکز هر خوشه توسط ارزش میانگین آیتم ها در خوشه نشان داده می شود:

ورودی:

K: تعداد خوشه ها

D:مجموعه داده ی حاوی n آیتم

خروجی: مجموعه ای از k خوشه

روش:

انتخاب دلخواه k آیتم از D به عنوان مراکز خوشه اولیه

تکرار

اختصاص دادن هر آیتم به خوشه ای که بیشترین شباهت را به آن دارد، بر اساس مقدار متوسط آیتم ها در خوشه

به روز رسانی میانگین خوشه، یعنی محاسبه ی میانگین آیتم ها برای هر خوشه

تا زمانی که تغییری رخ ندهد

معیارهای عملکرد

فرمول زیر برای اندازه گیری عملکرد با استفاده از ۴ خوشه ی متفاوت (۲۲، ۴۴، ۶۶، و ۸۸) به کار می رود:

$$DR_{NORMAL} = \frac{number\ of\ true\ normal\ data\ detected}{number\ of\ normal\ data\ derected} \times 100\% \quad (۲)$$

$$DR_{attack} = \frac{number\ of\ true\ normal\ data\ detected}{number\ of\ attack\ data\ derected} \times 100\% \quad (۳)$$

$$FPR = \frac{number\ of\ false\ positivse}{number\ of\ normal\ data} \times 100\% \quad (۴)$$

$$FNR = \frac{number\ of\ false\ negatives}{number\ of\ attack\ data} \times 100\% \quad (۵)$$

$$Efficiemcy\ \ rate = (DR_{normal}) + (DR_{attacik}) \quad (۶)$$

که در آن DR نرخ شناسایی

FPR : نرخ مثبت کاذب (یعنی داده های نرمال طبقه بندی شده به عنوان حمله)

FNR : نرخ منفی کاذب (یعنی حملات طبقه بندی شده به عنوان نرمال)

۴) نتایج و یافته ها

نتایج نشان داد که بسته به خوشه های مورد استفاده (۱۱, ۲۲, ۴۴, ۶۶, ۸۸) نرخ بازده ۸۱٫۶۱٪؛ ۶۵٫۴۰٪؛ ۶۱٫۳۰٪؛ و ۵۵٫۴۳٪ وجود دارد. علاوه بر این، قابل ذکر است که با افزایش تعداد خوشه ها به بیشتر از انواع داده، نرخ تشخیص، نرخ منفی کاذب، و نرخ بهره وری کاهش یافته اما نرخ مثبت کاذب افزایش می یابد. جالب است بدانید که بهترین نتایج زمانی بدست آمد که ۲۲ خوشه مورد استفاده قرار گرفت که منطبق با تعداد داده ها است. این نشان می دهد که عملکرد K ابزار وابسته به تعداد خوشه ها بوده و در نتیجه تعداد خوشه ها باید از قبل تعیین شود. همچنین قابل توجه است که برای تمامی خوشه ها نرخ مثبت کاذب به طور چشمگیری کمتر از نرخ منفی کاذب است. اگرچه عدم توانایی IDS در شناسایی داده های مخرب همچنان یک مشکل بوده و باید بیشتر مورد بررسی قرار بگیرد تا هشدار های کاذب کاهش پیدا کند (یعنی، مثبت کاذب). نرخ مثبت کاذب برای ۲۲ خوشه ۴٫۰۳ درصد و برای ۱۱ خوشه ۰٫۷۴ درصد است.

جدول ۱: نتایج خوشه بندی K ابزار

	تعداد خوشه ها				
	۱۱	۲۲	۴۴	۶۶	۸۸
داده های نرمال واقعی شناسایی شده	۱۳۳۵۰	۱۲۹۰۷	۱۱۳۵۸	۱۰۵۶۲	۹۱۵۷
داده های کلی واقعی شناسایی شده	۲۵۰۷۲	۲۴۴۳۱	۲۲۷۲۰	۲۱۸۷۳	۲۰۳۹۵
حملات واقعی شناسایی شده	۲۱	۲۹۱	۳۸۱	۴۳۲	۵۰۵
کل حملات شناسایی شده	۱۲۰	۷۶۱	۲۴۷۲	۳۳۱۹	۴۷۹۷
مثبت کاذب	۹۹	۵۴۲	۲۰۹۱	۲۸۸۷	۴۲۹۲
منفی کاذب	۱۱۷۲۲	۱۱۵۲۴	۱۱۳۶۲	۱۱۳۱۱	۱۱۲۳۸

۵) نتیجه گیری

نتایج خوشه بندی K ابزار نشان دهنده ی نرخ بازده ی بالاتر هنگام استفاده از تعداد صحیح خوشه ها بوده و همچنین نشان داد که افزایش یا کاهش تعداد خوشه ها فراتر از تعداد انواع داده تنها موجب کاهش بهره وری مدل می شود.

شناسایی تعداد خوشه ها در نتیجه نتایج را به طور چشمگیری تغییر می دهد. فرد باید از ابتدا بداند که چند خوشه برای دستیابی به نتایج خوب مورد انتظار است. در این مدل ۲۲ خوشه بر اساس انواع مختلف داده مورد استفاده قرار گرفت. با این حال، در یک

شبکه پویا، چالش شناسایی تعداد خوشه ها دشوار خواهد بود زیرا هیچ داده ی پایه ای به عنوان اساس تصمیم گیری در مورد تعداد خوشه ها وجود ندارد. در پرتو یافته ها، موارد زیر توصیه می شود:

تکنیک های دیگر داده کاوی (مانند بیزی، سلسله مراتبی، و غیره) برای مقایسه ی ممکن است مورد استفاده قرار بگیرد.

مطالعه ای با استفاده از الگوریتم داده کاوی K ابزار و سپس رویکرد مبتنی بر امضا برای کاهش نرخ منفی کاذب توصیه می شود.

سیستمی برای شناسایی خودکار تعداد خوشه ها ممکن است توسعه پیدا کند.

۶) مراجع

[1].Bischof, H., Leonardis, A., and Selb, A. MDL principle for robust vector quantisation. Pattern Analysis and applications. 2:59-72,1999.

[2].SANS Institute. Understanding Intrusion Detection System. 2001.

[3]. Bace, R., and Mell, P. Intrusion Detection System, NIST Special Publications SP800. November. 2001.

[4].Scarfone, K. and Mell, P. Guide to Intrusion Detection and Prevention System. National Institute of Standards and Technology. Special Publication 800-94. February 2007.

[5].J. Ioannidis, A. Keromytis, and M.Yung (Eds.): "IDS False Alarm Reduction using Continuous and Discontinuous Patterns", ACNS 2005, LNCS 3531, pp. 192–205, 2005. c Springer-Verlag Berlin Heidelberg 2005.

[6].Owen, D., "What is a False Positive and Why are False Positives a problem?", available online at http://www.sans.org/security-resources/idfaq/false_positive.php , last accessed in June 2015.

[7].Han, K., Kamber, M., Pei, J. Data Mining Concepts and Techniques. Third Edition. Morgan Kaufmann, Elsevier Inc. 2012. ISBN 978-0-12-381479-1.

[8].Jiawei, H. and Micheline, K. Data Mining Concepts and techniques, second edition, China Machine Press, pp. 296-303. 2006.

[9].Chapman, S.J. Matlab Programming for Engineers. International Student Edition. Fourth Edition. Thomson Learning, part of Thomson Corporation. ISBN-10:0-495-24451-1. ISBN-13:978-0-495-24451-6. 2008.

[10.Gilat, A. Matlab An Introduction with Applications. Fourth Edition. SI Version. John Wiley and Sons, Inc. 2011.ISBN: 978-0-470-87373-1. Printed in Asia.2011.

[11].Gilat, A. Matlab An Introduction with Applications. Third Edition. SI Version. John Wiley and Sons, Inc. 2011.ISBN: 978-0-470-10877-2. Printed in the United States of America. 2007.

[12].Kayacik, H.G., Zincir-Heywood, A.N. and Heywood, M.L. Selecting Features for Intrusion Detection: A Feature Analysis of KDD 99 Intrusion Detection Datasets.2006.

[13].KDD Cup 1999 Data available at: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html October 2007, last accessed in August 2014.

[14].Lane, T., and Brodley, C.E. Sequence matching and learning in anomaly detection for computer security. In AAAI Workshop: AI Approaches to Fraud Detection and Risk Management pp. 43-49.AAAI Press. July 1997.

[15]. Lee,W., Stolfo, S.J. and Mok,K.W. Data mining approaches for intrusion detection. N Proceedings of the 7th USENIX Security Symposium, March 1999.

[16].Lee, W., Stolfo, S.J. and Mok, K. Data Mining in work flow environments: Experiments in intrusion detection. In Proceedings of the 1999 Conference on Knowledge Discovery and Data Mining.1999.

[17].Lee, W.and Stolfo, S.J. A Framework for Constructing Features and Models for Intrusion Detection Systems. 1999.

[18].Lippmann, R.P., et.al. MIT Lincoln Laboratory Offline Component of DARPA 1998 Intrusion Detection Evaluation. MIT Lincoln Laboratory. PI Meeting. 1998.

[19].Mannila, H., Toivonen, H., and Verkamo, A.I. Discovering frequent episodes in sequences. In Proceedings of the 1st International Conference on Knowledge Discovery in Databases and Data Mining. Montreal, Canada. August 1995.

[20].MIT Lincoln Labs, 1998 DARPA Intrusion Detection Evaluation. Available at: http://www.II.mit.edu/mission/communications/ist/corpora/ideval/index.html last accessed on February 2014.

[21].Mukkama, S., Janoski, G., Sung, A. Intrusion detection using neural networks and support vector machines. Proceedings of IEEE International Joint Conference on Neural Networks, pp. 1702-1707.2002.

[22].Nguyen, H.A., and Choi, D. Application of Data Mining to Network Intrusion Detection: Classifier Selection Model. APNOMS 2008, LNCS 5297, Springer-Verlag Berlin Heidelberg 2008. pp.399-408. 2008.

[23].NSL-KDD Data Set for network-based intrusion detection systems available at: http://nsl.cs.unb.ca/NSL-KDD/, March 2009.

[24].Olusola, A.A., Oladele, A.S., and Abosede, D.O. Analysis of KDD '99 Intrusion Detection Dataset for Selection of Relevance Features. Proceedings of the World Congress on Engineering and Computer Science Vol I WCECS 2010. October 20-22, 2010, San Francisco, USA. 2010.

[25].Patel A.,Sammarvar,S., and Naik, A. DataMining Vs.Statistical Techniques for Classification of NSL-KDD Intrusion Data. International Journal of Computer Science and Information Technologies, Vol 5(4), 2014.ISSN:075-9646.

[26].Quinlan, J. C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo (1993).

[27].Sabhani, M., and Serpen, G. Why Machine Learning Algorithms Fail in Misuse Detection on KDD Intrusion Detection Dataset. Intelligent Data Analysis, vol 6. (Jne 2004).

[28].Stanford-Chen, S. Common intrusion detection framework. Available at: http://seclab.cs.ucdavis.edu/cidf.

[29].Siddiqui, M.K., and Naahid, S.Analysis of KDD CUP 99 Dataset using Clustering based Data Mining. International Journal of Database Theory and Application Vol.6, No. 5. pp.23-24. 2013.

[30].Tavallaee, M., Bagheri, E., Lu, W., and Ghorbani, A. (2009). A Detailed Analysis of the KDD CUP 99 Data Set. In Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Security and Defense Application (CISDA 2009).

[31].The DARPA Intrusion Detection Data Sets. Lincoln Laboratory Massachusetts Institute of Technology. Available at: www.II.mit.edu

[32].Waikato Environment for Knowledge Analysis (WEKA) version3.5.7. Available at: http://www.cs.waikato.ac.nz/ml/weka/, June 2008.

[33].Witten, I. H., Franck, E. Data Mining Practical Machine Learning Tools and Techniques, 2nd edition, Morgan Kaufmann, San Francisco. 2005.

[34].Xu, X. Adaptive Intrusion Detection Based on Machine Learning: Feature Extraction, Classifier Construction and Sequential Pattern Prediction, International Journal of Web Services Practices 2(1-2), 49-58. 2006.