العـدد (٢٢)، (عـدد خـاص)، الجـزء الثـانـي، نوفمبر ٢٠٢٣، ص ص ٤٤ – ٥٦

أنظمة مُقترحة لكشف التسلل بالاستناد على خوارزمية التعلم الآلي Rao-SVM

إعـداد

الأستاذ / منصور محمد هجاج العجمي

مدرب متخصص (ب)

قسم الحاسب الآلي

المعهد العالي للخدمات الإدارية

أنظمة مُقترحة لكشف التسلل بالاستناد على خوارزمية التعلم الآلي Rao-SVM

الأستاذ / منصور محمد هجاج العجمي (*)

ملخـــــــص

هدفت الدراسة تطوير بعض أنظمة كشف التسلل بناءً على خوارزميات التحسين للزيادة في ميزات بيانات التدقيق ؛ حيث ينخفض أداء خوارزميات التحسين أيضًا لـ IDS للأساليب القائمة على الإنسان من حيث وقت التدريب ودقة التصنيف، كما يهدف البحث تطوير طريقة محسنة للكشف عن التسلل للتصنيف الثنائي في IDS المقترحة ، ودمج خوارزمية Rao Optimization ، و Support Vector Machine (SVM) ، و Extreme Learning Machine (ELM) ، و Logistic Regression (LR) (اختيار الميزة والوزن) مع خوارزمية NTLBO مع تقنيات ML الخاضعة للإشراف لاختيار مجموعة الميزات الفرعية (FSS). نظرًا لأن اختيار المجموعة الفرعية للميزة وتعتبر مشكلة تحسين متعددة الأهداف، كما اقترحت الدراسة مفهوم RSS أقل. وتم استخدام مجموعة بيانات التعلم الآلي البارزة ، UNSW-NB15 ، للتجارب وأظهرت النتائج أن Rao-SVM وصلت دقة ٩٢.٥٪ على مجموعة بيانات UNSW-NB15 .

الكلمات المفتاحية: أنظمة – كشف التسلل – خوارزمية التعلم الآلي – Rao-SVM.

(*) مدرب متخصص (ب) - قسم الحاسب الآلي - المعهد العالي للخدمات الإدارية

Proposed intrusion detection systems based on machine learning algorithms Rao-SVM

Mansour Mohammed Hajaj alajmi

Abstract

The study aimed to develop some intrusion detection systems based on optimization algorithms to increase audit data features; The performance of optimization algorithms also decreases for IDS for human-based methods in terms of training time and classification accuracy. Machine (ELM), and Logistic Regression (LR) (feature selection and weight) with NTLBO algorithm with ML supervised feature subset selection (FSS) techniques. Since feature subset selection is considered a multi-objective optimization problem, the study also suggested a less RSS concept. A prominent machine learning dataset, UNSW-NB15, was used for the experiments and the results showed that Rao-SVM reached 92.5% accuracy on the UNSW-NB15 dataset..

Keywords: systems - intrusion detection - machine learning algorithm - Rao-SVM..

المقـــــدمـــــة:

تستدعي قضايا السرية والخصوصية والأمان الخطيرة عند استخدام الإنترنت في الآونة الأخيرة بسبب العمليات المتضمنة في تحويل البيانات ونقلها عبر الإنترنت، ويتطلب ذلك بذل الكثير من الجُهد نحو تحسين خصوصية وأمن أنظمة الكمبيوتر ؛ ومع ذلك ، لم يتم التعامل مع هذه المشكلات بشكل صحيح حيث لا يوجد حاليًا نظام آمن تمامًا في العالم. علاوة على ذلك ، هناك أنواع عديدة من هجمات الشبكات (Boumaaraf, 2023 : 102) .

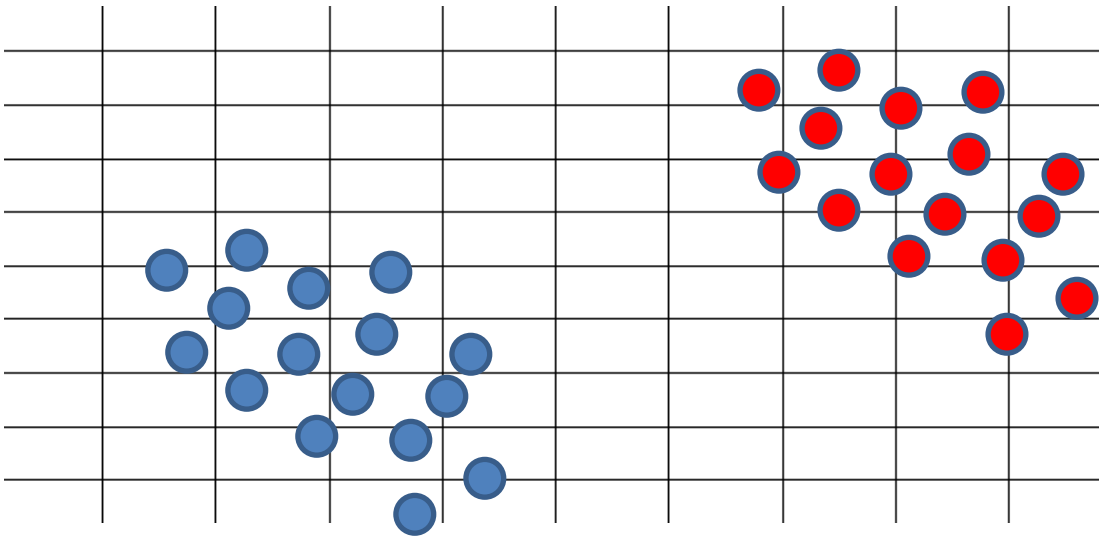
والتي تتطور عند إضافة توقيعات هجوم جديدة إلى قاعدة بيانات التوقيع، وأدى ظهور توقعات هجوم جديدة إلى الرغبة في تطوير أنظمة جديدة للكشف عن مثل هذه الهجمات فور ظهورها، ويُعد نظام كشف التسلل أحد الأدوات المُستخدمة لاكتشاف هذه الهجمات الجديدة حيث يمكنه مراقبة واكتشاف مجموعة من أنظمة الشبكات وأنظمة المعلومات وأنظمة الحوسبة السحابية، (Al-Haimoudi, 2023 : 177) ويتمثل عمل Intrusion Detection System [IDS] في مراقبة النظام والكشف عن وجود هجمات تهدف إلى مُهاجمة توافر النظام وسلامته وسريته، ويستعرض هذا البحث الأساليب والتقنيات في IDS ، والقسم الثاني يعطي نظرة عامة حول IDS ، (Negm, 2023 : 89) ثم يليه وصف موجز حول الأنواع الرئيسية لـ IDS والتقنيات المُستخدمة ، والتوصل إلى التحديات القائمة في IDS ، وفي القسم الرابع تتم مراجعة خوارزميات Machine Learning (ML) الأكثر استخدامًا في IDS بالتفصيل نقاط الضعف والقوة الرئيسية لكل خوارزمية في القسم IV.F. ويشرح القسم الخامس نوعين من معلمات خوارزميات التحسين التي تحتوي على معلمات أقل من الخوارزميات، وأخيرًا استنتاج بشأن ما تم إنجازه في القسم السادس .(Nassif, 2023 : 27)

الإطار النظري :

خوارزمية التعلم الآلي Rao-SVM

تُعد مسائل التصنيف Classification من المسائل الهامة في تعلم الآلة Machine Learning وذلك لأهميتها في العديد من المجالات المُتعددة من صناعية ومكتبية، وزراعية، وطبية، ..... إلخ ، ومن هذا المُنطلق اهتم الباحثين بهذا النوع من حيث قاموا على إعداد العديد من الخوارزميات التي أسهمت في مُعالجة العديد من المُشكلات الهامة ، ومن هذه الخوارزميات خوارزمية Support Vector Machine والتي يرمز لها بالرمز SVM ، وهذه الخوارزمية من خوارزميات التعلم المراقب Supervised Learning والتي من ضمن استخداماتها مسائل التصنيف ، والانحدار ، وأثبتت هذه الخوارزمية فعاليتها بدقة ممتازة في أغلب البيانات المستخدمة.

وهي تقوم على نوعين التصنيف الثنائي Binary Classification وتقوم بالتصنيف وفق مجموعتين Multi-class Classification حيث تقوم بالتصنيف لأكثر من مجموعة، وهذا طبقاً للشكل التالي (Negm, 2023 : 89) :



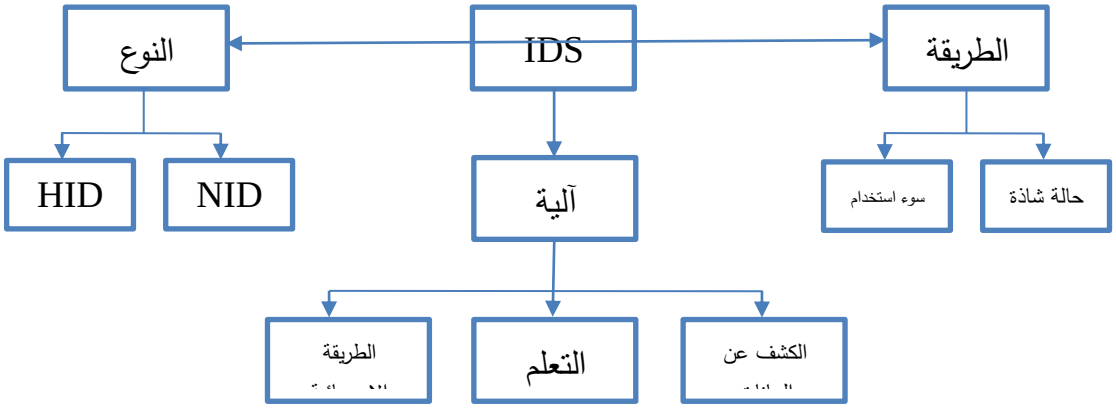
شكل (١) خوارزمية SVM

يتضح من الشكل السابقة أن الفكرة الرئيسة التي يقوم عليها عمل خوارزمية SVM تقوم على فكرة أيجاد أفضل مستوى Separating Hyperplane الذي يفصل بين المجاميع Classes ، وذلك عن طريق محاولة جعل حاشية Margin أكبر ما يمكن والحاشية Margin هي تمثل المسافة بين بيانات التدريب Training Data وبين المستوى الفاصل Separating Hyperplane والتي تعتبر محور عملية التدريب في بيئة خوارزمية SVM.

الدراسات المرتبطة :

يتم نشر أنظمة كشف التسلل على أنظمة الشبكة لمراقبة مثل هذه الأنظمة لمصادر التطفل المختلفة، وتنقسم أنظمة IDS الحالية إلى فئتين – أنظمة IDS القائمة على المضيف والقائمة على الشبكة ، وبالنسبة لـ IDS المستندة إلى الشبكة (NIDSs) ، يمكنهم تحديد اختراق الشبكة من خلال تحليل أنماط الشبكة المحددة، ولكن بالنسبة لـ IDS القائم على المضيف (HIDS) ، فإن عملهم بشكل أساسي هو اكتشاف المتسللين في المضيفين الفرديين (Ismail, et al., 2007 : 4) ، ويتم نشر NIDS أو تبديل الشبكة ؛ وهو برنامج يقرأ الحزم الأولية لقطاع الشبكة المحلية، والميزة لـ NIDSs هو أنه تستطيع اكتشاف وتحديد الهجمات التي قد تفوتها HIDS لأن HIDSs وليست مُصممة لرؤية رؤوس الحزم ؛ وبالتالي لا يمكنهم اكتشاف بعض أنواع هجمات الشبكة فعلى سبيل المثال يُمكن لـ NIDS اكتشاف العديد من هجمات DoS المستندة إلى IP لأنها تستطيع رؤية رؤوس الحزمة أثناء انتقالها عبر الشبكة المراقبة، ومن ناحية أخرى تتطلب HIDS أنظمة تشغيل مختلفة لتعمل بشكل صحيح على عكس NIDS التي لا تتطلب أي نظام تشغيل للمضيف كمصدر لتحديد الهجوم، ويمكن أيضًا تصنيف IODS إلى أنظمة الكشف عن سوء الاستخدام وأنظمة الكشف عن العيوب (Agustín, et al., 2023).

وتم تطوير أنظمة تحديد الهوية الهجينة كمزيج من كل من HIDS و NIDS. آلية الكشف المستخدمة في IDS هي ثلاثة أنواع رئيسية وهي: الطريقة الإحصائية ، والتعلم الآلي ، وطرق التنقيب في البيانات ، ويلخص الشكل التالي IDS :



شكل (١) نظرة عامة على نظام كشف التسلل IDS

١- كشف سوء الخطأ :

يكتشف سوء الخطأ في عمليات التطفل من خلال البحث عن الأنماط المعروفة للهجمات، ويتم استخدام هذه الاستراتيجية من قبل NIDSs التجارية الحالية؛ والجانب السلبي لمنع إساءة الاستخدام هو أنه لا يمكنه اكتشاف الهجمات غير المعروفة، وتم استخدام أساليب مُحددة مثل: شبكات الخبراء، وتحليل التوقيع، وتحليل انتقال الحالة، والتنقيب في البيانات، لتحديد الانتهاكات، ولوصف التدخلات يستخدم النظام الخبير مجموعة من القواعد .(Krishna, 2003)

ويتم تحويل أحداث المراجعة إلى حقائق في إطار عمل الخبراء والتي تحمل أهميتها الدلالية، وبعد ذلك باستخدام قواعد وأدلة معينة، حيث يقوم محرك الاستدلال باستخلاص النتائج، ويهدف تحليل انتقال الحالة إلى تحديد الهجمات بناءً على مجموعة من الأهداف والانتقالات باستخدام تحويل الحالة، ويستخدم تحليل التوقيع لوصف هجوم جديد بناءً على التوقيعات المضمنة بالفعل في مسار التدقيق.

وتُعتبر أنماط الشبكة التي تطابق التوقيعات في قاعدة البيانات تدخلاً، وتتوفر العديد من الأعمال حول استخدام طريقة استخراج البيانات القائمة على المُعرف، والتنقيب عن البيانات هو تقنية لاستخراج أنماط مهمة لم يلاحظها أحد من قبل من مجموعات البيانات الكبيرة ؛

ويمكن تعريف هذه الأنماط على أنها سلاسل قرار أو قواعد أو شبكات عصبية أو مثيلات قائمة على المثيل. وتتضمن خوارزميات DM الشائعة لاكتشاف إساءة الاستخدام بيانات تدقيق التعدين للنماذج التلقائية لكشف التسلل (معرف MADAM) .

وكشف التسلل باستخدام تقنيات التنقيب عن البيانات (IDDM) ، ومراجعة البيانات ومعالجتها (ADAM)، وتستند هذه النماذج على خوارزمية قواعد الرابطة ، كما تم تحسين أداء IDS باستخدام خوارزمية الشبكة العصبية.

٢- البحث عن الخلل :

لا يمكن تحديد التهديدات غير المعروفة واكتشاف الخلل في الاستخدام، ويتم استخدام اكتشاف التشوهات لمواجهة هذا القصور، وتم اقتراح وتنفيذ طرق مُختلفة لاكتشاف التهديدات كم خلال التجميع والتصنيف وما إلى ذلك ، ويستفيد الكشف الخاضع للإشراف عن الحالات الشاذة من بيانات التدريب الخالية من الهجمات لإنشاء أنماط حركة المرور العادية ؛ والكشف عن أي انحراف عن نمط المرور العادي المتعارف عليه باعتباره تدخلاً، ويُنشئ نمط حركة المرور العادي من مجموعة بيانات تدريب خالية من الهجمات ويصف الملف الشخصي كمجموعة من قواعد الارتباط، ويقوم بالكشف في الوقت الحقيقي عن الاتصالات المشبوهة بناءً على الملف الشخصي، وطرق أخرى خاضعة للإشراف، مثل : الخوارزميات الجينية ، والتنقيب عن البيانات الضبابية ، وSVM والشبكات العصبية، وتستخدم أيضًا لاكتشاف البيانات الضبابية، وغالبًا ما يتطلب الكشف عن الشذوذ الخاضع للإشراف استخدام الهياكل المتخصصة والتقنيات الرياضية، ويتم إنشاء ملف تعريف المستخدم باستخدام طرق إحصائية تستند إلى عدة حالات من السلوك العادي، ثم تتم مقارنة السلوكيات الجديدة مع الأشكال العادية، ويتم الكشف عن الانحرافات كتطفل، وبالنسبة للأنظمة الخبيرة فإنها تصف سلوك المستخدم العادي باستخدام مجموعة من القواعد ، ويتم تطبيق هذه القواعد لاكتشاف التدخلات.

المنهج المقترح :

اقترحت هذه الدراسة تنفيذ خوارزمية RAO في مرحلة FSS ؛ وتمت تهيئة RAO بواسطة مجموعة أولية تم إنشاؤها عشوائيًا ؛ وتكونت عينة الدراسة من المعلم ومجموعة من الطلاب الذين يعتبرون مجموعة من الحلول المحتملة. وتم تمثيل ميزات RAO من خلال دمج مشغلي التقاطع والطفرة في GA ؛ ويتيح تمثيل ميزات RAO ككروموسومات. حيث يتم استخدام عامل التقاطع لتحديث الكروموسوم، ويعتبر كل حل في كل جيل فردًا أو كروموسومًا كما هو موضح في الشكل (١)، ويتم تمييز الميزات المختارة للكروموسوم بالرقم (١) بينما يتم تمييز الميزات غير المحددة بعلامة، وتظهر تفاصيل الطريقة الجديدة في الخوارزمية (١) أثناء تدفق الطريقة موضحة في الخطوة (٤).

الخوارزمية ١: يعرض تفاصيل خوارزمية RAO-SVM

- الخطوة [١] تهيئة العينة بشكل عشوائي مع وجود مجموعة مختلفة من الميزات لكل مجموعة.
- الخطوة [٢] بناءً على دقة التصنيف لكل مجموعة من الميزات ، وتحديد أفضل وأسوأ مجموعة (العينة).
- الخطوة [٣] تعديل الحلول بناءً على أفضل الحلول وأسوأها والتفاعلات العشوائية على أساس *New_set= ran- dom_set crossover with (best_set crossover with worst_set)*
- الخطوة [٤] إذا كانت مجموعة الميزات الجديدة أفضل من أفضل مجموعة قديمة (من حيث دقة التصنيف)، فيتم الاحتفاظ بالمجموعة الجديدة وإلا احتفظ بالمجموعة القديمة.
- الخطوة [٥] هل تم استيفاء معايير الإنهاء أم لا ، إذا أبلغت بنعم عن أفضل مجموعة من الميزات ، فانتقل إلى الخطوة [٣].

الإعداد التجريبي :

تم عرض السيناريو التجريبي وحالات المشكلة ونتائج التجارب في هذه الدراسة ، وتم إجراء التجارب على مجموعة بيانات التطفل المسماة UNSW-NB15 والتي تم تقليلها بسبب التركيز على التصنيف الثنائي لاستيعاب فئتين فقط (عادي وتطفل) لضمان التحقق من صحة أفضل ، وتم التحقق من صحة K-fold ، حيث تم تعيين قيمة K على ١٠ .

إعـــداد البيانات :

تم إنشاء مجموعة بيانات UNSW-NB15 على الشبكة العنكبوتية للأمن السيبراني (ACCS) لاختبار أداء IDSs الجديدة ، وتتكون مجموعة البيانات من ١٠٠ جيجابايت من البيانات الأولية التي تلتقطها باستخدام أداة IXIA Perfect Storm وأداة tcpdump ؛ وتُمثل البيانات الأولية حركة مرور الشبكة المحاكاة للهجوم المعاصر والسلوكيات العادية الحديثة، وتم التقاط البيانات الأولية خلال فترتين محاكاة لمدة ١٦ ساعة و ١٥ ساعة، وبلغ الحجم الإجمالي لمجموعة البيانات ٢.٥ مليون سجل، حيث تم إنشاء ما مجموعه ٤٩ ميزة باستخدام أدوات Argus و Bro-IDS و ١٢ خوارزمية أخرى، وتم تصنيف الميزات الـ ٤٩ إلى ٥ فئات (ميزات التدفق ، وميزات المحتوى ، والميزات الأساسية ، وميزات الوقت ، والميزات الإضافية التي تم إنشاؤها) بينما تعمل ميزتان كتسمية (هجوم–قط يشير إلى فئة الهجوم والحالة الطبيعية، بالإضافة إلى التسمية التي تأخذ القيمة ١ للهجوم و ٠ للقيمة العادية)، وتحتوي مجموعة البيانات على ٩ فئات للهجوم وهي Fuzzers و DoS و Exploits والتحليل والباب الخلفي و Shellcode والديدان والعامة والاستطلاع.

النتائج :

يعرض الجدول أدناه نتائج الدقة لكلتا مجموعتي البيانات، ويتم عرض نتيجة الدقة لمجموعة بيانات UNSW-NB15 في الجدول (١) :

جدول (١) نتيجة دقة مجموعة بيانات UNSW-NB15

التصنيف	Rao	
	المزايا	الدقة
LR	17	0.921
	18	0.923
SVM	16	0.922
	19	0.925
ELM	19	0.92
	20	0.9215

يُظهر الجدول السابق كل من RAO-SVM و RAO-SVM نفس وقت التنفيذ لكل تقنية ML. لكل ML، وتم حساب عدد الميزات والدقة ووقت التنفيذ ، وتشير الأرقام إلى أفضل النتائج لكل من RAO-SVM وRAO-SVM، وقدم RAO-SVM باستمرار دقة أفضل مقارنةً بـ RAO-SVM باستخدام تقنيات ML الثلاثة، كما قدمت دقة وقت أفضل باستخدام تقنيات LR و SVM ML. ومع ذلك ، قدم RAO-SVM وقت تنفيذ أفضل مع ELM مقارنةً بـ RAO-SVM

خلاصة النتائج :

اقترحت الدراسة (RAO-SVM) جديدًا لمشكلات اختيار مجموعة الميزات في اكتشاف التسلل الشبكي، وتم إثبات أن أداء الخوارزمية الجديدة متفوق على العديد من الخوارزميات الأخرى في مشاكل FSS على مجموعتي بيانات اقتحام كبيرتين. حيث قدمت RAO-SVM المقترحة باستمرار دقة أفضل في وقت التنفيذ في الاختبارات الإحصائية (مصفوفة الارتباك) المطبقة على معدل اكتشاف RAO-SVM ومعدل الخطأ المستخرج من مصفوفة الارتباك ، وأظهر RAO-SVM معدل اكتشاف أعلى لمجموعة بيانات UNSW-NB15. وأظهر معدل خطأ منخفض لمجموعتي البيانات كتوصية ، ويجب تطبيق RAO-SVM المقترح على مشاكل التصنيف متعدد الفئات ، ويمكن استخدام المزيد من تقنيات غسل الأموال لتقييم أدائها.

المراجـــــع

Agustín Lara A , & Vicente Mayor A, & Rafael Estepa A , & Antonio Estepa A , & Jesús E. Díaz-Verdejo, (2023) : Smart Home Anomaly-Based Ids: Architecture Proposal And Case Study, Internet Of Things Journal, Vol. (22), No. (4).

Al-Haimoudi, Badr (2023): Cybersecurity And The Protection Of Information Systems, Volume (1), Number (2), North African Journal Of Scientific Publishing, African Academy For Advanced Studies, Libya.

Boumaaraf, Manal (2023): Sirian Bullying, Its Causes, Methods, The Most Important International Programs To Confront It, Volume (7), Special Issue, Rawafed Journal For Studies And Scientific Research In The Social And Human Sciences, Belhaj Bouchaib University Center, Ain Temouchent, Algeria.

Ismail A. Ghaffar & Mohamed Aborizka, & Khaled A. Fatah (2007) : An Architecture For Distributed Intrusion Detection System Using Autonomous Agents, Proceeding Of The 12-Th Asat Conference, 29-31 May 2007.

Krishna Kavi, David C. Kung, Hitesh Et. Al., (2003) : ” Extending Uml For Modeling And Design Of Multiagent Systems”, 2nd Intl Workshop On Software Engineering For Large-Scale

Multi-Agent Systems (Selmas2003), Portland, Or, May 3-10, 2003.

Nassif, Ahmed Mostafa (2023): Merging Cybersecurity Into The National Security System: Cybersecurity And National Security, Issue (645), Money And Trade Magazine, Al-Tijara Club, Egypt.

Negm, Al-Sayed (2023): Vogue Cyberspace Wars, Issue (418), Journal Of The National Progressive Unionist Gathering Party, Egypt.