Journal of Education and Science
Vol. 27, No.2, 2018, pp. 146-168
ISSN 1812-125X
http://edusj.mosuljournals.com

استخدام خوارزمية النمل في كشف وتصنيف التطفل في شبكات الحاسوب

نجلاء بديع إبراهيم الدباغ **محمود صبحي محمود قصاب باشي**

قسم علوم الحاسوب / كلية علوم الحاسوب والرياضيات والإحصاء / جامعة الموصل

تأريخ الاستلام 28/06/ 2012 تأريخ القبول12/11/ 2012

Abstract

The widespread development of using networks and relying on it and the emergence of many challenges faced by so many institutes like ministry of defense, intelligence agencies , banks, universitie...etc which depend heavily on service provided by those networks, resulted in the emergence of a new security concept towards building new security systems known as " Intrusion Detection Systems".

The current study aims at detecting the known and unknown attacks as well as classifying the known ones.

A network intrusion detection system is developed using Ant-Miner algorithm, which is one of the Swarm Intelligence which assimilate , the behavior of natural environment of ants. It is used to investigate the training data and getting the basic classification which is helpful in classifying the attacks in the tests data. The KDD Cup 1999 dataset concerned in constructing and evaluating intrusion systems which includes Training Dataset of 494021 in number as well as the testing dataset of 311029 in number.

Experiments are conducted on the system using data sets and the complete tests. The results show that the detection rate using the SOM is: 92.37% and false alarm is: 4.67% and classification rate using Ant-Miner algorithm is : DOS(Denial of Service) 96.15 %, U2R(User to Root) 97.13%, Normal 94.50%, probe 72.90%, R2L(Remote to Local) 13.88%. These results show the effectiveness of SOM and Ant-Miner algorithms in the field of network security.

الخلاصة

إن النمو الواسع في استخدام الشبكات الحاسوبية والاعتماد عليها ، وظهور التحديات التي لا يستهان بها والتي واجهتها المؤسسات مثل (وزارة الدفاع و الاستخبارات والمصارف و المحال التجارية والجامعات وغيرها) المبنية على الخدمات التي توفرها تلك الشبكات الحاسوبية ، أدى إلى ظهور مفهوم أمني جديد وهو بناء أنظمة أمنية معروفة بأنظمة كشف التطفل (Intrusion Detection Systems) .

تم في هذا البحث تطوير نظام كشف وتصنيف التطفل الشبكي (Network Intrusion Detection System) بالاعتماد على خوارزمية النمل (Ant Miner) ، أذ أن هذه الخوارزمية هي إحدى خوارزميات النمل التي تقع ضمن حقل ذكاء السرب (Swarm Intelligence) التي تحاكي سلوك العيش لبيئة النمل الطبيعي التي من خلالها يتم تنقيب بيانات التدريب والحصول على قواعد التصنيف المفيدة في تصنيف الهجمات الموجودة في بيانات الاختبار. استخدمت مجموعة بيانات (KDD Cup 1999) الخاصة ببناء وتقييم أنظمة كشف التطفل والتي تتضمن مجموعة بيانات التدريب (Training Dataset) البالغ عددها (494021) سجل اتصال وكذلك مجموعة بيانات الاختبار (Testing Dataset) البالغ عددها (311029) سجل اتصال .

أجريت التجارب العملية على النظام باستخدام مجاميع التدريب و الاختبار كاملة وكانت نتائج التصنيف (Classification Rate) لمجموعة الاختبار باستخدام خوارزمية Ant-Miner ، DOS(Denial of Service) 96.15 % ، U2R(User to Root) 97.13 % ، (Normal 94. 50 %) ، (probe 72.90 %) ، (R2L(Remote to Local 13.88 %) .مما يظهر جدوى خوارزمية (Ant-Miner) في حقل أمنية الشبكات .

المقدمة

إن تعقيد أنظمة الحواسيب الموزعة وأهميتها وموارد المعلومات المتوفرة فيها نمت بسرعة كبيرة ، استنادا لهذه الحقيقة فقد أصبحت الحواسيب وشبكاتها هدفاً لجرائم الحاسوب التي ازدادت أكثر فأكثر [1] . إن مستخدم الانترنيت سوف يشارك في العديد من التطبيقات التي تحتاج إلى التنقلات الأمينة ضمن الشبكة الحاسوبية عندما ينقل المعلومات والاعتمادات المالية ، إذ تعمل المصارف والأسواق التجارية المعتمدة على الانترنيت في كل يوم على تحقيق الوصول إلى مواقع كلمات المرور المحمية. عندما تزداد تلك التنقلات فإن النظام سوف يصبح معرضاً للهجمات التي تعمل على خزن العمليات والبيانات المهمة [2,3].

استخدمت بعض التقنيات لحماية البيانات المهمة مثل الجدار الناري والتشفير وألخ . يعمل الجدار الناري كمدافع لحماية البيانات الحساسة ولكنه يقلل الكشف فضلا عن المراقبة وتحديد نقاط الضعف في أنظمة التشغيل . كما أن الرسالة المشفرة ممكن فك شفرتها وفضلاً عن ذلك فإن التشفير يضيف عباً أضافياً على المستخدمين والتطبيقات . إن أي تقنية أمنية جديدة تحوي في تصميمها على بعض العيوب مما يجعلها هدفا للهجمات ، لذلك فإن من المهم أن نملك أنظمة كشف التطفل لحماية البيانات المهمة . لهذا السبب أصبحت أنظمة كشف التطفل في الشبكة الحاسوبية ذات اهتمام كبير من قبل العديد من الباحثين في السنوات الأخيرة [4].

نظام كشف التطفل Intrusion Detection System (IDS) هو مكون مهم في إطار أمنية الحاسبة والمعلومات وإن هدفه الرئيس هو تمييز بين الفعاليات الاعتيادية للنظام و السلوكيات التي ممكن أن نصفها على أنها تطفل ، إن إي نظام لكشف التطفل لا يهمل استخدام التقنيات الأمنية الأخرى ولكن يعمل كخط دفاع أخير في النظام [4]. استخدم الباحثون العديد من التقنيات في بناء أنظمة كشف التطفل ، حيث هنالك طرائق إحصائية (Statistical Approaches) [5] و تقنيات ذكائية اصطناعية (Artificial Intelligence Techniques) [6] .

في هذا البحث استخدمت التقنيات الذكائية الاصطناعية في عملية كشف التطفل وتصنيفه ، حيث استخدمت خوارزمية تنقيب النمل (Ant-Miner) التي هي إحدى خوارزميات مستعمرات النمل Ant Colony Optimization (ACO) التي تقع ضمن حقل ذكاء السرب (Swarm Intelligence (SI،والتي تعمل على تصنيف اتصالات الشبكة إلى أحد الأصناف الخمسة (DOS , U2R , R2L , Probe , NORMAL) عن طريق توليد مجموعة من قواعد التصنيف التي تستخدم في التصنيف .

إن الدافع من استخدام خوارزميات مستعمرات النمل (ACO) في اكتشاف قواعد التصنيف والمهام المرتبطة بتنقيب البيانات هو أن هناك العديد من المشاريع التي تعمل في حقل تنقيب البيانات اقترحت خوارزميات إنتاج القواعد المحددة (Deterministic Rules Induction) ، هذه الخوارزميات تكون غير مرنة وحساسة في إيجاد قواعد التصنيف المحلية (فضلا عن قواعد التصنيف العامة). إن الاختلاف هو أن خوارزميات (ACO) تحاول أن تخفف هذا العائق باستخدام الدمج بين فكرتين أساسيتين. الأولى إن هذه الخوارزميات تملك مظهر العشوائية (Stochastic aspect) التي تساعد على الكشف في مساحات كبيرة من فضاء البحث ، الثانية هو استخدام إجراء التكيف المكرر (Iterative Adaptation Procedure) المعتمد على التغذية العكسية الايجابية (الزيادة التدريجية للمادة العطرية المرتبطة مع مكونات الحل الأفضل) لكي تحسن قواعد التصنيف المتنبأ باستمرار. عموماً فأن

دمج هاتين الفكرتين الأساسيتين ، يعني بأن خوارزميات (ACO) تؤدي البحث بشكل عام في فضاء بحث قواعد التصنيف المتنبأ بصورة أفضل من خوارزميات قواعد الإنتاج المحددة التي تجعلها البديل الأهم المعتمد عليه في قواعد الإنتاج (Rules Induction) [7].

الدراسات السابقة

دُرست مشكلة كشف التطفل وتصنيفه بشكل واسع ضمن حقل أمنية شبكة الحاسوب ، وهنالك عدد من المفاهيم التي تم العمل عليها لتحديد المتطفلين وكشفهم في شبكة الحاسوب ، إذ أقترح في السنوات السابقة عدد من المفاهيم ، يعرض السرد الآتي الدراسات السابقة لمعالجة مشكلة البحث المختارة وميزة كل دراسة ، والعوائق التي واجهها الباحثون في أنظمتهم إن وجدت ، ضمن كل دراسة :

1- إن خوارزمية النمل (ذكاء السرب) وشبكة خرائط التنظيم الذاتي (نوع من الشبكات العصبية الاصطناعية التي لا تحتاج إلى مشرف) أصبحت مهمة وذات قوة تصنيف حدسية عالية في مجال علوم الحاسوب والذكاء الاصطناعي . في عام 2004 قدم (Michael L. Gargano et al.) [8] أول اقتراح ، لدمج خوارزمية النمل المقترحة مع خوارزمية (SOM) لتصميم نظام كشف التطفل ، جاءت فكرة نظرية فقط دون أي تطبيق .

2- في عام 2004 استخدم الباحثان (Mehdi Moradi , Mohammed Zulkernine) [9] ، الشبكات العصبية في بناء نظام كشف التطفل ، إذ استخدمت الشبكة (Multi-layer Perceptron) في بناء النظام والمعتمد على طرائق التحليل (Offline) . في هذا البحث حُلت مشكلة تعددية الأصناف (Multi Class) ، إذ إن صنف الهجوم سوف يحدد عن طريق الشبكة العصبية . أوضحت النتائج أن النظام المصمم له القدرة على التصنيف بدقة 91% في حالة استخدام طبقتين خفيتين ودقة 87% في حالة استخدام طبقة خفية واحدة . استخدمت مجموعة جزئية من بيانات KDD في اختبار النظام .

3- في عام 2007 قدم الباحثان (Yacine Bouzida , Frderic Cuppens) [10] ، التحسينات على تقنية الشبكات العصبية و تقنية أشجار القرار التي تعمل على كشف الشذوذ الجديدة . أوضحت النتائج أن التقنيات بعد التحسين تكشف الهجمات المعروفة والجديدة واستخدم مجموعة بيانات (KDD) إذ استخدمت في الحالة الأولى مجموعة نماذج التدريب في مرحلة التدريب وفي الحالة الثانية استخدمت مجموعة نماذج الاختبار

في مرحلة التدريب ، واختبر النظام حسب بيانات التي استخدمت في مرحلة التدريب ، حيث كانت نسب اكتشاف الأصناف (R2L , U2R) قليلة جداً في حالة استخدام التقنيات القياسية وتم تحسين خوارزمية (C4.5) عن طريق إضافة صنف جديد بدلا من الصنف الافتراضي فأصبحت نسبة كشف الصنف (U2R) عالية ونسبة الصنف (R2L) قليلة جداً .

4- في عام 2009 استخدمت الباحثتان (Wafa S. Al-Sharafat , Reyadh Sh. Naoum) [11] ، خوارزمية Genetic Based Machine Learning (GBML) التي تعطي القدرة الجيدة على التكيف مع البيئات وتملك القوة في التعامل مع البيانات التي تحوي على الضوضاء وكذلك لها القدرة على تعريف الهجمات غير المعروفة . إذ دمجت تقنيات مختلفة في نظام التصنيف لكي يكشف ويصنف التطفل عن حزم سير البيانات الاعتيادية ومن هذه الخوارزميات استخدم Steady State Genetic Based Machine Learning (SSGBML) التي تستخدم لكشف التطفل ، إذ استخدمت مجموعة نماذج التدريب KDD وهي 10% KDD 99 في مرحلة التدريب ، فحص النظام على (نماذج التدريب نفسها) .

5- في عام 2009 قدم الباحثان (Mansour Sheikan , Davood Gharavian) [12] ، خوارزمية جديدة تدمج (Classification Based Aassociation Rules) مع (Neural Network Elman) التي تحسن من أمنية الشبكات ، إذ قدم في الماضي طريقة أطلق عليها Associative Classification التي تدمج التصنيف وقواعد الارتباط في تنقيب البيانات . إن دمج الطرائق المقدم يعمل على تحسين نتائج التصنيف ، فإن استخدام تحليل الخصائص ذات العلاقة سوف يعمل على تقليل من عدد الخصائص في مرحلة Associative Rules . وأوضحت النتائج أن النظام المقترح يحسن من نتائج التصنيف وبشكل خاص للأصناف R2L و U2R . ويملك نسبة كشف كفوءة وكذلك تكون نسبة الإنذار الكاذب قليلة . استخدم 15000 نموذج اختبار من مجموعة نماذج الاختبار (KDD) في التجارب .

6- في عام 2009 قدم الباحث (عادل صبري عيسى) [4] ، نظام كشف التطفل باستخدام ثلاثة تقنيات (تنقيب البيانات باستخدام قواعد الارتباط Data Mining Associative Rules(DM ARs) ، وأشجار القرار Decision Trees(DTs) والشبكات العصبية الاصطناعية Artificial Neural Network(ANN)) ثم المقارنة بينهم لمعرفة

أفضل تقنية تفيد في بناء نظام كشف التطفل. استخدمت في هذا البحث بيانات KDD Cup 99 المعتمدة في تقييم نظام كشف التطفل ، تم استخدام كل بيانات التدريب التي تسمى KDD Cup 99 % 10 وكذلك بيانات الاختبار. أظهرت النتائج أن التقنيات المقترحة حصلت على نتائج أفضل مقارنة مع التقنيات القياسية وكذلك تبين أن الشبكات العصبية الاصطناعية وأشجار القرار هي أفضل من قواعد الارتباط . ومن جهة أخرى فإن زمن التعليم في الشبكات العصبية الاصطناعية تأخذ أكثر وقتا مقارنة مع أشجار القرار .

إن الهدف من البحث هو تطوير نظام يكون له القدرة على كشف الهجمات فضلاً عن تصنيف تلك الهجمات ، إذ إن أنظمة كشف التطفل السابقة كانت تعمل على تصنيف مجموعة بيانات الاختبار إلى اتصالات اعتيادية أو هجوم ، استخدمت في هذا البحث خوارزمية (Ant-Miner) في بناء المصنف الذي يتضمن مجموعة من قواعد التصنيف التي لها القدرة على تصنيف مجموعة نماذج الاختبار إلى الأصناف (Normal , DOS , R2L , U2R , Probe) . وبهذا فإن النظام له القدرة على كشف الهجوم وتصنيفه .

1- مجموعة بيانات KDD 99

في سنة 1999 اشتقت مجموعة البيانات KDD 99 [13](Knowledge Discovery in Databases) من سير البيانات ضمن الشبكة في DARPA 98 (Defense Advanced Research Project Agency) عن طريق برنامج Bro الذي يجمع حزم TCP في اتصالات TCP وهي مجموعة البيانات المستخدمة في منافسة أدوات تنقيب البيانات واكتشاف المعرفة العالمية والمستخدمة بشكل واسع في مجال كشف التطفل [14] .

يتضمن كل اتصال TCP على واحد وأربعين ميزة كما موضح في الملحق (أ) الجداول (2،3،4،5) مع عنوان يوضح هل أن هذا الاتصال هو اتصال اعتيادي أو نوع من أنواع الهجمات [15] ، هناك ثمانية وثلاثون ميزة رقمية وثلاث ميزات رمزية كما في المثال التالي .

```
0,tcp,private,S0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,279,7,1.00,1.00,0.00,0.00,0.03,0
.06,0.00,255,13,0.05,0.07,0.00,0.00,1.00,1.00,0.00,0.00,neptune.
```

وكما موضح في الجدول رقم (1) [16] .

جدول رقم (1) المكونات الأساسية لمجموعة بيانات كشف التطفل (KDD 99)

Data set	DOS	Probe	U2R	R2L	Normal
10 % KDD	391458	4107	52	1126	97277
KDD Corrected	229853	4166	70	16347	60593
Whole KDD	3883370	41102	52	1126	972780

إن مجموعة بيانات التدريب (Whole KDD) تتكون من 4940000 نموذج اتصال التي تغطي سير الحزم في الاتصالات الاعتيادية في الشبكة وتتضمن أربعاً وعشرين هجمة . وإن مجموعة بيانات الاختبار تتكون من 311029 نموذج اتصال التي تتضمن ثماني وثلاثين هجمة وأن أربعة عشر منهم لم تظهر في مجموعة بيانات التدريب . إن (Whole KDD) مجموعة بيانات كبيرة جداً ولهذا استخدم 10% من بيانات التدريب في عملية تقييم أنظمة كشف التطفل بشكل واسع التي تتضمن على اثنين وعشرين هجمة .

2- الأصناف الرئيسة للهجمات

هناك أربعة أصناف رئيسة للهجمات [13,17] وهي :

1- منع الخدمة Denial Of Service (DOS) : يقوم المهاجم بعمل بعض الحسابات أو جعل مصادر الذاكرة مشغولة كثيرا أو إرسال عدد كبير من الطلبات إلى الخادم مثل إرسال عدد كبير من حزم التزامن ذات عنوان مصدر مزيف إلى خادم معين ...الخ . مما يجعل عدم وصول المستخدمين القانونيين إلى تلك الخدمات . مثل (apache attack , smurf attack , Neptune attack) وغيرها .

2- الوصول غير المخول عن بعد Remote to Local (R2L) : يقوم المهاجم بإرسال الحزم إلى حاسبات الضحية عبر الانترنيت . إن المهاجم لا يملك الصلاحية للوصول إلى تلك الحاسبات ولهذا سوف يستغل بعض نقاط الضعف الموجودة في تلك الحاسبات ومستوى الشرفية التي يملكها مستخدمو تلك الحاسبات مثل تخمين كلمات المرور مثل (xlock attack , guest attack , phf attack) وغيرها .

3- الوصول غير المخول إلى مستوى شرفية الجذر لحاسبة الضحية Users to Root (U2R) : يبدأ المهاجم بهذا الهجوم بعد دخوله إلى نظام الضحية ثم يحاول الحصول على مستوى الشرفية للجذر باستخدام نقاط الضعف في ذلك النظام مثل الأنواع المختلفة لهجمات امتلاء المخازن المؤقتة مثل (perl attack , xterm attack) وغيرها .

4- التجسس (Probe) : المراقبة والتجسس مثل مسح المدخلات (Ports) ومعرفة الخدمات وعدد الحاسبات في الشبكة المقصودة مثل (saint , port sweep) وغيرها .

5- ذكاء السرب Swarm Intelligence (SI)

هو عبارة عن تقنية ذكائية اصطناعية تتضمن دراسة السلوك الجماعي للأنظمة اللامركزية ، إنها تشابه تقريباً سلوك العيش عند الحشرات والسرب لكي يبسط تصميم حلول المسائل المعقدة الموزعة . إن هذا السلوك يشار إليه في طرائق الأنظمة المعقدة والأنماط الذي

ينشأ من تعددية التفاعلات البسيطة [18] . في السنين السابقة فإن ذكاء السرب طبق وبشكل ناجح في الأمثلية والإنسان الآلي والتطبيقات العسكرية [14].

إن طرائق (SI) تحل المسائل المعقدة عن طريق عدد من الأفراد (Agents) البسيطة دون سيطرة مركزية أو مساندة من العالم الخارجي . فإن التفاعلات الداخلية بين الأفراد مع بيئتها غالبا ما يولد هذا السلوك لهذه المجاميع . وإن أهم مميزتين في (SI) هما التوزيع العالي للسيطرة وإستراتيجية الانبثاق ولهذا سوف تنتج أنظمة استقلالية ومتكيفة وقابلة للتوسع ومرنة ومتوازية و ذات تنظيم الذاتي وكفاءة الكلفة [19].

بصورة عامة إن نماذج (SI) تعتمد على المجتمع وكل أفراد هذا المجتمع هي حلول ممكنة وفي الخطوات المتكررة فإن العناصر تغير مواقعها ضمن مجال البحث عن طريق الاتصالات المتجهة أو غير المتجهة ، مفضل على عمليات التزاوج والطفرة الوراثية في حساب التخمين التطوري (Evolutionary Computation) . هناك نظريتان شائعتان ناتجة عن طريق التخمينات في الذكاء التخميني (Computational Intelligence) هما (ACO) Ant Colony Optimization و (PSO) Particle Swarm Optimization إذ إن (ACO) تماثل سلوك النمل وتكون ناجحة في حالة تطبيقها في حل مسائل الأمثلية المتقطعة . (Discreet optimization Problems)أما (PSO) تماثل بشكل بسيط الأنظمة الاجتماعية مثل سرب الطيور أو مجاميع الأسماك وتكون مناسبة في حل مسائل الامثلية غير الخطية مع بعض المحددات [14,20].

5-1- خوارزمية مستعمرة النمل Ant Colony Algorithm

تعد خوارزمية النمل من خوارزميات التعليم التطوري(Evolutionary learning) التي تستطيع أن تطبق لحل المسائل الأمثلية المركبة (Combinational Optimization) [21,22,23]. إن خوارزمية (ACO) هي Meta-heuristic وهي نظرية عامة لتعريف الحدس لمدى واسع من مساحات المشكلة . إن Meat-heuristic ACO تستخدم نظريات لتعريف الحدسية المتأثرة بشكل كبير من قبل سلوك النمل الطبيعي . حيث إن النمل يضع المادة العطرية على الأرض كدليل إلى الآخرين لإيجاد الطعام . إن اختيار المسار يعتمد على كثافة المادة العطرية ولهذا فان سلوك النمل يحث على نشوء خوارزمية تتكون من مجموعة من النمل الذكائي (Artificial Ants) كمجموعة من الأفراد(Agents) البسيطة التي تعمل مع بعضها البعض لحل المسائل عن طريق تبادل المعلومات بوساطة المادة العطرية الموضوعة على حافات المخطط . إن المادة العطرية تعمل كذاكرة موزعة لربط النمل مع بعضها البعض . إن هذه الخوارزمية تكون نظام النمل المطبق على عدة مسائل أمثلية مركبة [24,25] .

إن المسألة التي تحل عن طريق ACO تستخدم المادة العطرية مع الدالة الحدسية حيث إن المادة العطرية تعمل كتغذية عكسية وتحدث بالاعتماد على نوعية القاعدة المكتشفة . وإن أهم جزء في الخوارزمية هو اختيار المسار الصحيح وهذا منجز عن طريق توليد الاحتمالية في اختيار المسارات المتاحة [24,25] .

يتضمن الجدول رقم (2) اسلوب المحاكاة بين سلوك النمل الطبيعي مع اسلوب خوارزمية تنقيب النمل .

جدول رقم (2) أسلوب المحاكاة بين اسلوب النمل الطبيعي واسلوب خوارزمية تنقيب النمل

ت	سلوك النمل الطبيعي	سلوك خوارزمية تنقيب النمل
1	النقاط التي تمر بها النمل للوصول الى الهدف	الميزات التي يتكون منها كل سجل اتصال
2	المادة العطرية التي تعمل كذاكرة موزعة	عبارة عن متغير من نوع Double مرتبط مع كل ميزة
3	اختيار افضل اتجاه بالاعتماد على خبرة النمل مع المادة العطرية	اختيار افضل قيمة للميزة بالاعتماد على دالة الاحتمالية بالاضافة الى قيمة متغير المادة العطرية المرتبط بتلك الميزة
4	زيادة او تبخر المادة العطرية	زيادة او نقصان متغير المادة العطرية بالاعتماد على نوعية قاعدة التصنيف

6-1- تطبيق ACO في أنتاج قواعد التصنيف

قدم العالم Monmarch و آخرون في سنة 1999 [26] ACO إلى تنقيب البيانات عن طريق تطوير التجميع (Clustering) لقواعد البيانات العددية باستخدام النمل الذكي . أما العالم Parpinelli et al قدم سنة 2000 [27] خوارزمية ACO لإنتاج قواعد التصنيف عن طريق تطوير خوارزمية يطلق عليها Ant-Miner .

إن خوارزمية تنقيب النمل (Ant-Miner) هي نظام نمل يستخدم لمهمة تصنيف البيانات . في هذه الخوارزمية كل سجل في قاعدة البيانات يعطى لصنف واحد من بين مجموعة من الأصناف اعتمادا على قيم الميزات الموجودة في السجل المعني ولهذا فإن الناتج من الخوارزمية يكون عبارة عن مجموعة من القواعد وكل قاعدة تتكون من جزأين كما يلي :

If $<Term_1 \& Term_2 \& Term_3 \ldots\ldots Term_n>$ **then** <Class> …(1)

إذ إن جزء (IF) يتضمن بعض التعابير التي ترتبط مع بعضها البعض عن طريق العملية المنطقية (&) وأن تركيب كل تعبير يتضمن الميزة مع العملية العلائقية مع القيمة وكما يلي < attribute , operator , Value > ومثال على ذلك <Sex = Male> .

أما جزء (Then) يتضمن الصنف المتنبأ لسجلات التي تملك القيم المطابقة لتعابير جزء (If) . إن نتائج خوارزمية تنقيب النمل في بعض التطبيقات توضح أن خوارزمية النمل لها القدرة الكبيرة لتعمل في حقل مهمة التصنيف في تنقيب البيانات [24] .

6-1- إدخالات النظام System Parameters

إن نظام النمل الاصطناعي يملك أربعة إدخالات معرفة من قبل المستخدم وهي [7,20] :

1- عدد النمل المستخدم في النظام (No_of_Ants) : هو العدد الأكبر لعدد النمل (Agents) التي تكون قواعد التصنيف في كل دورة من النظام وأن أفضل قاعدة سوف تختار وتضاف إلى قائمة القواعد المكتشفة . يمكن أن يلاحظ أنه كلما كان عدد النمل كبيراً سوف يكون النظام أفضل ولكن يعمل بشكل بطيء .

2- أقل عدد من النماذج لكل القاعدة (Min_cases_per_Rule): كل قاعدة مكونة يجب أن تغطي على الأقل هذا العدد من نماذج بيانات التدريب .

3- أكبر عدد من النماذج غير المغطاة في مجموعة بيانات التدريب (Max_uncovered_cases): إن عملية اكتشاف القواعد سوف تستمر بالأداء إلى أن يكون عدد النماذج في مجموعة التدريب التي لم تغط من قبل القواعد المكتشفة أقل من هذا العدد .

4- عدد التقارب بين القواعد المكونة (No_Rules_Converg) : إذا كانت القاعدة المكونة حاليا تشابه عدداً من القواعد المكونة سابقا بقدر هذا العدد فإن النظام وصل إلى حالة التقارب في مسار واحد و الدورة الحالية سوف تتوقف وتبدأ دورة جديدة .

6-2- خوارزمية Ant-Miner

تبدأ الخوارزمية بقائمة قواعد تصنيف فارغة ثم تهيى كمية المادة العطرية لكل الميزات وبشكل متساوي بعد ذلك تكون قواعد التصنيف بشكل منتظم من خلال دخول الخوارزمية في الدارة الداخلية (3-4) وكما ذكرنا سابقا أن كل تعبير في قاعدة التصنيف يكون بالصيغة التالية < attribute = Value >

واختيار قيمة التعبير تؤدى اعتمادا على دالة هجينة التي تدمج كمية المادة العطرية للقيمة مع الدالة الحدسية (Entropy) للقيمة أيضا . بعد الانتهاء من توليد قاعدة التصنيف سوف تستدعى دالة التهذيب وبعدها تحدث كميات المادة العطرية وفي النهاية يفحص هل أن عدد قواعد التصنيف المكونة تساوي العدد المحدد في المتغير (No. Of Ants) أو عدد قواعد التصنيف المتقاربة يصل إلى العدد المحدد في المتغير (No_Rules_converge) [7,24].

وخطوات خوارزمية تنقيب النمل المطورة في هذا البحث هي

الخطوة الأولى : مجموعة نماذج التدريب = كل نماذج التدريب المراد تدريب النظام عليها
الخطوة الثانية : قائمة قواعد التصنيف المكتشفة = قائمة فارغة // بدائيا تكون فارغة .
الخطوة الثالثة : طالما (مجموعة نماذج التدريب أكبر من العدد الأكبر لنماذج غيرالمكتشفة)
3-1 : t=1 // يمثل عدد النمل الحالي
3-2 : j=1 // مؤشر لعدد حالات التقارب بين قواعد التصنيف المكونة
3-3 : التهيئة البدائية لكل قيم الميزات بالكمية نفسها من المادة العطرية .
3-4 : كرر (أ ، ب ، ج ، د) إلى أن تصبح (t) أكبر من عدد النمل المحدد أو تصبح (j) أكبر من عدد حالات التقارب المحددة .
أ – النملة (t) تبدأ بقاعدة فارغة لتكوين قاعدة التصنيف Rt عن طريق إضافة تعبير واحد إلى القاعدة التصنيف Rt في الوقت المحدد .
ب – تهذيب قاعدة التصنيف Rt .
ج - تحديث كمية المادة العطرية عن طريق زيادة كمية المادة العطرية لكل التعابير الموجودة في قاعدة التصنيف Rt ونقصان كمية المادة العطرية للتعابير غير الموجودة في قاعدة التصنيف Rt
د – التأكد من التقارب :
أ – إذا كانت قاعدة التصنيف Rt تشابه قاعدة التصنيف Rt-1 فإن (j=j+1) .
ب- وإلا فإن (j=1) .
3-5 – اختيار أفضل قاعدة تصنيف Rbest من بين قواعد التصنيف المكونة من قبل النمل .
3-6 – إضافة Rbest إلى قائمة قواعد التصنيف المكتشفة .
3-7 – مجموعة نماذج التدريب = مجموعة نماذج التدريب - { مجموعة كل نماذج التدريب التي تغطى من قبل Rbest .
الخطوة الرابعة : خزن قائمة قواعد التصنيف المكتشفة في الملف النصي .

إذا تحقق أحد الشروط فإن الدارة الداخلية سوف تنتهي وبهذا يُختار أفضل قاعدة تصنيف مكونة وتضاف إلى قائمة قواعد التصنيف المكتشفة ، في هذه اللحظة يتم تحديث مجموعة نماذج التدريب وتتكرر الدارة الخارجية لحين وصول عدد نماذج التدريب أقل من العدد المحدد في المتغير (Max_Uncoverd_cases) .

وبعد الانتهاء من عملية البحث فإن القواعد المكتشفة توضع في قائمة متسلسلة حسب تسلسل الاكتشاف التي سوف تستخدم في عملية تصنيف الحالات الجديدة التي لم تظهر في

مجموعة التدريب . إن النظام سوف يضيف قاعدة (Default Rule) في نهاية قائمة القواعد المكتشفة حيث إن هذه القاعدة لا تملك أي تعبير في جزء الشروط من القاعدة ، ويوضع في جزء الصنف التصنيف الذي يغطي معظم الحالات في مجموعة التدريب التي لم تغط من قبل القواعد المكتشفة . تطبق هذه القاعدة (Default Rule) إذا كانت الحالة المراد فحصها لم تطابق أي قاعدة من القواعد المكتشفة وتأخذ التصنيف من جزء الصنف لهذه القاعدة[7,24] .

وفيما يلي شرح لخطوات خوارزمية تنقيب النمل

أولا : بناء القواعد Rule Construction

إن عملية بناء قواعد التصنيف تبدأ بقاعدة فارغة ثم تقوم بإضافة التعابير الواحد تلو الآخر ، حيث التعبير الذي يختار لكي يضاف إلى القاعدة المكونة حاليا ، يجب أن يملك أعلى قيمة لدالة الاحتمالية مع كمية كبيرة من المادة العطرية ويتبع القيود التالية [25,28] :

1- يجب أن يكون التعبير غير موجود في القاعدة مسبقاً ولهذا فان النمل يجب أن يتذكر ما هي التعابير التي أضيفت إلى القاعدة التي تبنيها . إن هذه الذاكرة الصغيرة تعد إحدى الفروقات بين النمل الاصطناعي والنمل الطبيعي .

2- إن التعبير لا يضاف إلى القاعدة الحالية إذا كان يؤدي إلى توسع القاعدة ويجعل القاعدة تغطي عدداً من نماذج مجموعة التدريب أقل من قيمة الإدخال (Min_cases_per_rule).

ثانياً : الدالة الحدسية Heuristic Function

أي تعبير يمكن أن يضاف إلى القاعدة الحالية من قبل النملة الحالية فإن خوارزمية النمل تحسب الدالة الحدسية حسب المعادلتين رقم (2 , 3) لتخمين نوعية التعبير مع القدرة على تحسين دقة التنبؤ لتلك القاعدة [25,28].

$$InfoTij = -\sum_{w=1}^{k}[\frac{FreqTijw}{Tij}]*\log 2[\frac{FreqTijw}{Tij}] \ldots\ldots\ldots\ldots(2)$$

إذ إن :

- w تمثل الصنف المطلوب .
- k تمثل عدد الأصناف في المسألة .
- $FreqT_{ijw}$ تمثل تكرارات T_{ij} في الصنف w .
- $|T_{ij}|$ تمثل تكرارات T_{ij} .

$$nij = \frac{\log 2(k) - \inf oTij}{\sum_{i}^{a}\sum_{j}^{bi}\log 2(k) - \inf oTij} \ldots\ldots\ldots\ldots\ldots(3)$$

إذ إن :

- K تمثل عدد الأصناف في المسألة .
- a تمثل عدد الميزات .

- bi تمثل عدد القيم في مدى الميزة i .

لكي نستخدم الدالة الحدسية هناك احتمالان .

1- إذا كان التعبير لا يظهر في مجموعة التدريب ، في هذه الحالة يعطى أعلى قيمة في الدالة الحدسية وهذا يقابل إعطاء التعبير أقل احتمالية ضمن تلك الميزة .

2- إذا كان التعبير يظهر في كل نماذج مجموعة التدريب تكون دالة الحدسية تساوي صفراً وهذا يقابل إعطاء التعبير أعلى درجة من الاحتمالية .

بعد أن تحسب الدالة الحدسية سوف تحسب احتمالية إضافة ذلك التعبير إلى قاعدة التصنيف حسب المعادلة رقم (4) ، حيث تعتمد على الدالة الحدسية مع كمية المادة العطرية .

$$Pij(t) = \frac{tij(t).nij}{\sum\sum tij(t).nij, \forall i \in I} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)$$

إذ إن :

Tij(t) : تمثل كمية المادة العطرية المرتبطة مع ذلك التعبير في الزمن (t) .

ثالثاَ : تهذيب القاعدة Rule Pruning

إن تهذيب قاعدة التصنيف هي تقنية معتاد عليها في إنتاج القواعد كما مذكور سابقا . فان الهدف الرئيس من تهذيب القاعدة هو حذف التعابير من القاعدة التي ليس لها علاقة بالصنف المطلوب وتكون عديمة الفائدة في تلك القاعدة . وإن تهذيب القاعدة يزيد من دقة التنبؤ لتلك القاعدة وكذلك يزيد من بساطة القاعدة حيث إن القاعدة الصغيرة (تحوي على عدد قليل من التعابير) تكون بصورة عامة أسهل من ناحية الترجمة والفهم من قبل المستخدم من القاعدة الكبيرة وتتم عملية تهذيب القاعدة بعد بنائها .

إن الفكرة الأساسية لعملية التهذيب هي عملية حذف تعبير واحد من القاعدة في كل دورة على حساب تحسين الدقة . ففي الدورة الأولى من العملية سوف تبدأ النملة بقاعدة كاملة ثم تحاول أن تحذف التعبير من القاعدة (حيث كل دورة تحذف تعبير واحد) وتحسب النوعية (أي أعطاء القاعدة إلى صنف آخر) . إذ إن التعبير المحذوف يجب أن يحسن النوعية وتعاد العملية على تعبير آخر من القاعدة وهكذا . إن هذه العملية سوف تستمر إلى أن تصبح القاعدة تحوي على تعبير واحد أو لا يوجد تعبير يتم حذفه ويحسن من نوعية القاعدة [25,28].

رابعاَ : تحديث كميات المادة العطرية Pheromone Updating

إن كل تعبير يرتبط مع موقع في المسارات التي ممكن أن تتبع من قبل النمل . في بداية عمل كل دورة من النظام سوف تهيئ كل التعابير بالكمية نفسها من المادة العطرية . وإن أول نملة سوف تبدأ عملية البحث في كل المسارات التي تملك كمية المادة العطرية نفسها ، إن عملية التهيئة البدائية لكميات المادة العطرية تكون متناسبة عكسيا مع عدد القيم في كل الميزات كما موضح بالمعادلة رقم (5) .

$$tij(t=0)=\frac{1}{\sum_{i=1}^{a} bi} \quad .....(5)$$

إذ إن :

t : يمثل الوقت (بداية العمل) .

- bi : تمثل عدد القيم في مدى الميزة i .

إذ إن هذه الكمية من المادة العطرية تدمج مع الدالة الحدسية لكي تعطي احتمالية اختيار التعبير لكي يضاف إلى القاعدة . في كل وقت تنتهي النملة من بناء القاعدة فإن كميات المادة العطرية في كل المواقع سوف تحدث وهذا التحديث يملك فكرتين أساسيتين هما

1– كمية المادة العطرية المرتبطة مع التعابير الموجودة في القاعدة المكونة سوف تزداد.

إن زيادة كمية المادة العطرية مع التعابير الموجودة في القاعدة المكونة يؤدي إلى زيادة كمية المادة العطرية في المسار بكامله المكون من قبل النملة وأن زيادة كمية المادة العطرية سوف يجعل هذا المسار مختاراَ من قبل النمل اللاحق في المستقبل . ويفيد النملة الحالية من ناحية تحسين نوعية القاعدة .

إن نوعية (Quality) القاعدة المكونة من قبل النملة يشار إليها بالرمز (Q) وتحسب عن طريق المعادلة رقم (6) :

Q= Sensitinty * Specificity

وتعرف كما يلي

$$Q=\left(\frac{TP}{TP+FN}*\frac{TN}{FP+TN}\right) .....(6)$$

إذ إن :

(TP(True Positive : يمثل عدد النماذج المغطاة من قبل قاعدة التصنيف والتي تحمل صنف مشابه لصنف قاعدة التصنيف .

(FP(False Positive : يمثل عدد النماذج المغطاة من قبل قاعدة التصنيف والتي تحمل صنف مختلف عن صنف قاعدة التصنيف .

(FN(False Negative : يمثل عدد النماذج غير المغطاة من قبل قاعدة التصنيف والتي تحمل صنف مشابه لصنف قاعدة التصنيف .

(TN(True Negative : يمثل عدد النماذج غير المغطاة من قبل قاعدة التصنيف والتي تحمل صنف مختلف عن صنف قاعدة التصنيف .

إن قيمة النوعية (Q) يجب أن تقع ضمن المدى [0,1] .

كمية المادة العطرية للتعابير الموجودة في القاعدة المكونة سوف تحدث بالاعتماد على نوعية القاعدة المكونة وتحسب عن طريق المعادلة رقم (7) .

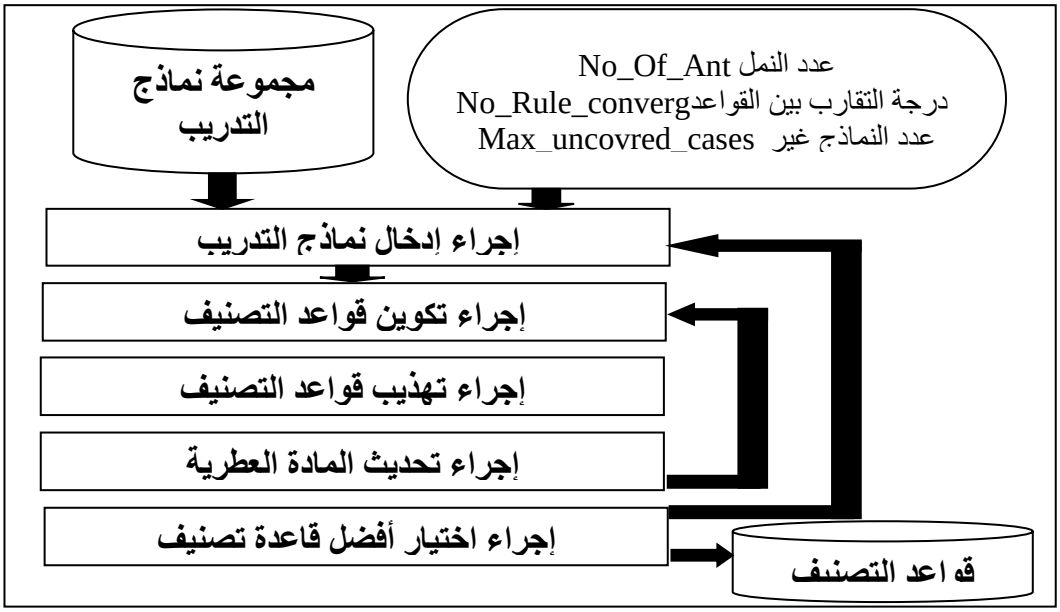
$$Tij(t+1) = Tij(t) * Q, \forall Tij \in R \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (7)$$

إذ إن :

R : يمثل القاعدة الحالية .

2- كمية المادة العطرية المرتبطة مع التعابير غير الموجودة بالقاعدة المكونة سوف تقل ، عملية تقليل كمية المادة العطرية تقابل تبخر المادة العطرية عند النمل الطبيعي [25,28] .

إن عملية اكتشاف قواعد التصنيف سوف تتم حسب المخطط الصندوقي التفصيلي كما في الشكل (2) . تمر خوارزمية Ant-Miner بالمراحل التالية :



الشكل رقم (2) المخطط الصندوقي التفصيلي الخاص بخوارزمية Ant-Miner

7- تجارب نظام التصنيف باستخدام خوارزمية (Ant-Miner)

7-1- مرحلة التدريب

تمت مرحلة تدريب النظام باستخدام حاسبات عدد (5) بشكل متوازي ، حيث كل حاسبة كانت تقوم بإيجاد قواعد التصنيف لصنف معين وخزن القواعد في ملف نصي و استغرقت مرحلة التدريب الكلية فترة زمنية قدرها (ثلاثة أيام) وكانت قيم المدخلات في مرحلة تدريب النظام كما يلي

NO_of_ Ants = 5

Min_Cases_Per_Rule = 10
Max_Uncoverd_cases = 10
No_Rules_converg = 5

بعد انتهاء مرحلة تدريب النظام حصل على مجموعة من قواعد التصنيف كما موضح في الملحق (أ) في الجدول رقم (1) جزء منها ، وكان أعداد تلك القواعد مع زمن اكتشافها حسب التصنيف موضحة بالجدول رقم (3) .

جدول رقم (3) أعداد قواعد التصنيف المكتشفة في مرحلة تدريب النظام

Class	DOS	PROB	R2L	U2R	NORMAL
No. Of Rules	109	61	34	4	105
Training Time Day : hour : minute	2 : 22 : 51	1 : 15 : 39	0 : 22 : 6	0 : 2 : 36	2 : 22 : 15

7-2- مرحلة التصنيف

لقد تم اختبار النظام باستخدام قواعد التصنيف المحصول عليها من مرحلة تدريب النظام باستخدام حاسبة واحدة و كما يلي :

أولا : عملية التصنيف باستخدام مجموعة نماذج التدريب (Training Dataset)

تمت عملية اختبار النظام عن طريق استخدام مجموعة نماذج التدريب (استخدمت البيانات نفسها المستخدمة في مرحلة تدريب النظام) والحصول على نتائج الاختبار الموضحة بالجدول رقم (4) .

جدول رقم (4) نتائج التصنيف بتطبيق خوارزمية النمل على بيانات التدريب

Class	NO of Rules	Detection	Detection Rate	Testing Time m : s
DOS	109	390194	99.66 %	3 : 45
U2R	4	71	94.66 %	1 : 30
Normal	105	92955	95.55 %	3 : 20
Probe	61	4073	99.17 %	2 : 32
R2L	34	1068	99.72 %	2 : 20

ثانياَ : عملية التصنيف باستخدام مجموعة نماذج الاختبار (Testing Dataset)

لقد أختبر النظام عن طريق استخدام مجموعة نماذج الاختبار وكانت النتائج الاختبار موضحة بالجدول رقم (6) .

جدول رقم (6) نتائج التصنيف بتطبيق خوارزمية النمل على بيانات الاختبار

Class	NO of Rules	Detection	Classification Rate	Testing Time m : s
DOS	109	221026	96.15 %	3 : 15
U2R	4	4442	97.13 %	1 : 0

Normal	**105**	**57266**	**94.50 %**	**2 : 50**
Probe	**61**	**3088**	**72.90 %**	**2 : 11**
R2L	**34**	**1641**	**13.88 %**	**2 : 2**

مناقشة النتائج

بعد أن أجريت عملية الاختبار والحصول على النتائج خوارزمية (Ant-Miner) على نسبة التصنيف مع حجم البيانات المستخدمة في الاختبار ، نلاحظ في الجدول رقم (6) أن نسبة الكشف في الصنف (R2L) قليلة مقارنة مع بقية الأصناف وهذا يعود إلى ظهور سبع هجمات جديدة في بيانات الاختبار تعود إلى الصنف (R2L) التي تمثل أكثر من 60 % من بيانات الاختبار التابعة لصنف (R2L) وكذلك هناك عدد كبير من بيانات سجلات الاتصال التي تعود إلى الصنف (R2L) تشابه البيانات في سجلات الاتصال التابعة إلى الصنف (Normal) ونلاحظ أن نسبة التصنيف عالية جداً في الصنف (U2R) مقارنة مع نتائج الباحثين في الجدول رقم (7) .

قورنت النتائج المحصول عليها من التجارب مع نتائج عدد من الباحثين الذين يعملون في المجال نفسه إذ أن الجدول رقم (7) يبين أن خوارزمية النمل حققت نتائج جيدة في تصنيف التطفل مع أنها استخدمت لأول مرة في هذا المجال .

الجدول رقم (7) مقارنة النتائج لخوارزمية (Ant–Miner) مع عدد من الباحثين

Model	Testing Data	DOS %	Probe %	R2L %	U2R %	Normal %	DR %
DM Ars [4]	Testing data	79.08	66.75	9.4	14.03	94.62	
DM NN [4]	Testing Data	97.25	81.23	32.50	0.0	88.04	
DM DT [4]	Testing Data	97.1	83.3	8.2	13.2	99.5	
Pnrule[12]	15000 test data	96.9	73.2	10.7	6.6	99.5	
Multi-classifier[12]	15000 test data	97.3	88.7	9.6	29.8	*	
ESC-IDS[12]	15000 test data	99.5	89.1	31.5	14.1	98.2	
Elman NN [12]	15000 test data	96.0	75.6	11.7	0.0	97.6	
CPAR[12]	15000 test data	96.3	52.2	4.4	9.5	99.2	
Hybrid Elman NN/CPAR[12]	15000 test data	97.0	19.5	31.8	33.3	97.4	
DT(c4.5) [10]	Training Data	99.99	99.78	98.93	90.39	99.94	
SSGBML [11]	Training Data	*	*	*	*	*	97.45
Ant – Miner[29]	Training Data	99.66	99.17	99.72	94.66	95.55	98.85
Ant – Miner[29]	Testing	96.15	72.90	13.88	97.13	94.50	92.42

	Data						

(*)القيمة ضمن هذا الحقل غير محسوبة من الباحث .

الاستنتاجات

من خلال تطبيق النظام في البحث لغرض كشف التطفل الشبكي وتصنيفه وعلى وفق النتائج التي تم الحصول عليها من التجارب ، نستنتج ما يلي :-

1- أثبت البحث أن خوارزمية Ant-Miner تعمل على تصنيف الهجمات أفضل من الخوارزميات العصبية [12,4] من ناحية نسبة التصنيف وكذلك وقت التدريب والاختبار كما موضح بالجدول رقم (7) .

2- أثبت في هذا البحث أن خوارزمية Ant-Miner تعمل على تصنيف الهجمات أفضل من الخوارزميات الجينية [11] من ناحية نسبة التصنيف كما موضح بالجدول رقم (7) .

3- تعامل خوارزمية Ant-Miner مع البيانات العددية والحرفية دون تحويلها إلى صيغة معينة .

4- حذف Default Rule من مجموعة قواعد التصنيف الناتجة يعطي نتائج أفضل من وضعها .

5- خوارزمية Ant-Miner تملك نتائج تصنيف أفضل مقارنة مع خوارزمية أشجار القرار (C4.5) [4] .

6- قواعد التصنيف الناتجة من خوارزمية Ant-Miner تكون سهلة الفهم والوضوح من القواعد الناتجة من خوارزمية أشجار القرار (C4.5) .

7- مرونة تعامل خوارزمية Ant-Miner مع البيانات المضببة وغير الدقيقة بسبب احتوائها على عملية تهذيب قواعد التصنيف .

الملحق (أ)

جدول رقم (1) قواعد التصنيف الخاصة بالصنف (R2L)

1	0,tcp,ftp,SF,,,0,0,0,,0,1,0,0,0,0,0,0,0,0,0,1,1,1,0.00,0.00,0.00,0.00,1.00,0.00,0.00,,,,,,0.00,0.00,0.00,0.00,0.00,R2L CLASS.
2	0,tcp,ftp,SF,,,0,0,0,,0,1,0,0,0,0,0,0,0,0,0,1,1,1,0.00,0.00,0.00,0.00,1.00,0.00,0.00,,,,,,0.00,0.00,0.00,,0.00,R2L CLASS.
3	0,tcp,ftp,SF,,,0,0,0,,0,1,0,0,0,0,0,0,0,0,0,1,1,1,0.00,0.00,0.00,0.00,1.00,0.00,0.00,,,,,,0.00,,0.00,0.00,0.00,R2L CLASS.
4	0,tcp,ftp,SF,,,0,0,0,,0,1,0,0,0,0,0,0,0,0,0,1,1,1,0.00,0.00,0.00,0.00,1.00,0.00,0.00,,,,,,0.00,,0.00,,0.00,R2L CLASS.
5	0,tcp,,SF,,0,0,0,0,,0,1,0,0,0,0,0,0,0,0,0,,1,1,0.00,0.00,0.00,0.00,1.00,0.00,0.00

	,1,1,1.00,0.00,1.00,0.00,0.00,0.00,0.00,0.00,R2L CLASS.
6	0,tcp,,SF,,0,0,0,0,,0,1,0,0,0,0,0,0,0,0,0,,1,1,0.00,0.00,0.00,0.00,1.00,0.00,0.00 ,1,,1.00,0.00,1.00,,0.00,0.00,0.00,0.00,R2L CLASS.
7	0,tcp,,SF,,0,0,0,0,,0,1,0,0,0,0,0,0,0,0,0,,1,1,0.00,0.00,0.00,0.00,1.00,0.00,0.00 ,,,1.00,0.00,1.00,0.00,0.00,0.00,0.00,0.00,R2L CLASS.
8	0,tcp,,SF,,0,0,0,0,,0,1,0,0,0,0,0,0,0,0,0,,1,1,0.00,0.00,0.00,0.00,1.00,0.00,0.00 ,,,1.00,0.00,1.00,,0.00,0.00,0.00,0.00,R2L CLASS.
9	0,tcp,,SF,,0,0,0,0,,0,1,0,0,0,0,0,0,0,0,0,,1,1,0.00,0.00,0.00,0.00,1.00,0.00,0.00 ,,,1.00,0.00,1.00,,0.00,,0.00,0.00,R2L CLASS.
10	0,tcp,,SF,,0,0,0,0,,0,1,0,0,0,0,0,0,0,0,0,,1,1,0.00,0.00,0.00,0.00,1.00,0.00,0.00 ,,,,,1.00,,0.00,0.00,,,R2L CLASS.
11	0,tcp,,SF,,0,0,0,0,,0,1,0,0,0,0,0,0,0,0,0,,1,1,0.00,0.00,0.00,0.00,1.00,0.00,0.00 ,,,,,,0.00,,,0.00,0.00,R2L CLASS.
12	0,tcp,,SF,,0,0,0,0,,0,1,0,0,0,0,0,0,0,0,0,,1,1,0.00,0.00,0.00,0.00,1.00,0.00,0.00 ,,,,,,,,0.00,0.00,R2L CLASS.
13	0,tcp,,SF,,0,0,0,0,,0,1,0,0,0,0,0,0,,0,0,,,,0.00,0.00,0.00,0.00,1.00,0.00,0.00,,,1. 00,0.00,1.00,0.00,0.00,0.00,0.00,0.00,R2L CLASS.
14	0,tcp,,SF,,0,0,0,0,,0,1,0,0,0,0,0,0,,0,0,,,,0.00,0.00,0.00,0.00,1.00,0.00,0.00,,,1. 00,0.00,1.00,,0.00,0.00,0.00,0.00,R2L CLASS.
15	0,tcp,,SF,,0,0,0,0,,0,1,0,0,0,0,,0,,0,0,,,,0.00,0.00,0.00,0.00,1.00,0.00,0.00,,,1.0 0,0.00,1.00,,0.00,,0.00,0.00,R2L CLASS.
16	0,tcp,,SF,,0,0,0,0,,0,1,0,0,0,0,,0,,0,0,,,,0.00,0.00,0.00,0.00,1.00,0.00,0.00,,,,,,0 .00,,,0.00,0.00,R2L CLASS.
17	0,tcp,,SF,,0,0,0,0,,0,1,0,0,0,0,,0,,0,0,,,,0.00,0.00,0.00,0.00,1.00,0.00,0.00,,,,,,, 0.00,0.00,0.00,0.00,R2L CLASS.
18	0,tcp,,SF,,0,0,0,0,,0,1,0,0,0,0,,0,,0,0,,,,0.00,0.00,0.00,0.00,1.00,0.00,,,,,,,,,0.0 0,,R2L CLASS.
19	0,tcp,,SF,,,0,0,0,,0,,0,0,0,0,,0,,0,0,,,,0.00,0.00,0.00,0.00,,,0.00,,,,,,,,,0.00,R2L CLASS.
20	0,tcp,,,,0,0,0,0,,0,,0,0,0,0,0,0,,0,0,,,,,,0.00,,1.00,0.00,,,,1.00,0.00,,,,,0.00,0.00, R2L CLASS.
21	,tcp,ftp,SF,,,0,0,0,,0,1,0,0,0,0,,0,,0,0,1,1,1,0.00,0.00,0.00,0.00,1.00,0.00,0.00, ,,,,,0.00,0.00,0.00,0.00,0.00,R2L CLASS.
22	,tcp,ftp_data,SF,,0,0,0,0,0,0,1,0,0,0,0,0,0,,0,0,0,1,1,0.00,0.00,0.00,0.00,1.00,0 .00,0.00,,,,,,,0.00,0.00,0.00,0.00,R2L CLASS.
23	,tcp,,SF,,0,0,0,0,0,0,1,0,0,0,0,0,0,,0,0,0,1,1,0.00,0.00,0.00,0.00,1.00,0.00,0.00 ,,,,,,,0.00,,,,R2L CLASS.
24	,tcp,,SF,,0,0,0,0,0,0,1,0,0,0,0,0,0,,0,0,0,,,0.00,0.00,0.00,0.00,1.00,0.00,0.00,2, ,,,,,0.00,,0.00,0.00,R2L CLASS.
25	,tcp,,SF,,0,0,0,0,0,0,1,0,0,0,0,0,0,,0,0,0,,,0.00,0.00,0.00,0.00,1.00,0.00,0.00,,,, ,,,,,0.00,0.00,R2L CLASS.
26	,tcp,,SF,,0,0,0,0,,0,1,0,0,0,0,0,0,,0,0,0,1,1,0.00,0.00,0.00,0.00,1.00,0.00,0.00,, ,,,,,0.00,0.00,0.00,0.00,R2L CLASS.
27	,tcp,,SF,,0,0,0,0,,0,1,0,0,0,0,0,0,,0,0,0,1,1,0.00,0.00,0.00,0.00,1.00,0.00,0.00,, ,,,,,,,0.00,0.00,R2L CLASS.
28	,tcp,,SF,,0,0,0,0,,0,1,0,0,0,0,0,0,,0,0,0,,,0.00,0.00,0.00,0.00,,,,,,,,,,,0.00,0.00, R2L CLASS.
29	,tcp,,SF,,,0,0,0,0,0,,0,0,0,,0,0,,0,0,0,1,1,0.00,0.00,0.00,0.00,1.00,0.00,0.00,,,,,, 0.00,0.00,0.00,0.00,0.00,R2L CLASS.

30	,tcp,,SF,,,0,0,0,,0,1,0,0,0,,0,0,,0,0,,1,1,0.00,0.00,0.00,0.00,1.00,0.00,0.00,,,,,,0.00,0.00,0.00,,0.00,R2L CLASS.
31	,tcp,,SF,,,0,0,,,0,1,,0,0,,0,0,,0,0,,1,1,0.00,0.00,0.00,0.00,1.00,0.00,0.00,,,,,,0.00,,0.00,0.00,0.00,R2L CLASS.
32	,tcp,,SF,,,0,0,0,,0,1,,0,0,,0,0,0,0,0,,1,1,0.00,0.00,0.00,0.00,1.00,0.00,0.00,,,,,,0.00,,0.00,,0.00,R2L CLASS.
33	,tcp,,SF,,,0,0,0,,0,1,0,0,0,,,0,0,0,0,,,,,,,0.00,,,,,,,,,,,,,R2L CLASS.
34	,tcp,,,,,0,0,0,,0,,0,,0,0,0,0,,0,0,0,,,,,,,1.00,0.00,,,,,,,,,,,R2L CLASS.

جدول رقم (2) الميزات الأساسية من اتصالات TCP المنفردة

نوع بيانات الميزة	الوصف	أسم الميزة
بيانات مستمرة	فترة الاتصال بالثواني	Duration
بيانات متقطعة	نوعية الاتفاقية مثل (TCP , UDP)	Protocol_type
بيانات متقطعة	نوع خدمة الشبكة في الهدف مثل (http , Telnet)	Service
بيانات مستمرة	عدد بايتات البيانات المرسلة من المصدر إلى الهدف	Src_bytes
بيانات مستمرة	عدد بايتات البيانات المرسلة من الهدف إلى المصدر	dst_bytes
بيانات متقطعة	مؤشر لحالة الاتصال هل هو اتصال اعتيادي أم يحوي على حالات الخطأ	Flag
بيانات متقطعة	تكون 1 إذا كان الاتصال من والى المضيف نفسه أو المدخل نفسه و 0 على خلاف ذلك	Land
بيانات مستمرة	عندما ترسل البيانات من المصدر سوف تقسم إلى مقاطع وتستلم من قبل الهدف فإن عدد المقاطع الخاطئة توضع في هذه الميزة	Wrong fragment
بيانات مستمرة	عدد الحزم المهمة والملحة	urgent

جدول رقم (3) ميزات المحتوى ضمن الاتصالات عن طريق معرفة الميدان (Domain Knowledge)

نوع بيانات الميزة	الوصف	أسم الميزة
بيانات مستمرة	عدد مؤشرات الخطر	hot
بيانات مستمرة	عدد محاولات التسجيل الفاشلة	Num_failed_logins
بيانات متقطعة	تكون 1 إذا تم التسجيل بنجاح و 0 خلاف ذلك	Logged_in
بيانات مستمرة	عدد شروط الخطر أو التعرض للشبهة	Num_compromsed
بيانات متقطعة	تكون 1 إذا حصل على مستوى الشرفية للجذر و 0 على خلاف ذلك	Root_shell
بيانات متقطعة	تكون 1 أذا كان هناك محاولة تنفيذ الإيعاز su root و 0 على خلاف ذلك	Su_attempetd
بيانات مستمرة	عدد مستويات الجذر التي توصل إليها	Num_root
بيانات مستمرة	عدد عمليات تكوين الفايل	Num_file_creations
بيانات مستمرة	عدد مقدمات shell	Num_shells
بيانات مستمرة	عدد العمليات على ملفات سيطرة الوصول	Num_access_files
بيانات مستمرة	عدد الايعازات الخارجة عن حدود أي جلسة ftp	Num_outbound_cmds
	تكون 1 إذا كان الولوج تابع إلى القائمة الساخنة و 0 خلاف ذلك	Is-Hot-Login
بيانات متقطعة	تكون 1 إذا كان التسجيل هو طفيل (مضيف) و 0 على خلاف ذلك	Is_guest_login

جدول رقم (4) ميزات سير الحزم المستندة على الوقت

نوع بيانات الميزة	الوصف	أسم الميزة
بيانات مستمرة	عدد الاتصالات إلى المضيف نفسه التي تشابه الاتصال الحالي في الثانيتين السابقتين	Count
بيانات مستمرة	عدد الاتصالات المتصلة مع المضيف نفسه التي تملك الخدمة نفسها التي تشابه الاتصال الحالي في الثانيتين السابقتين	Srv_count
بيانات مستمرة	النسبة المئوية للاتصالات التي تحوي على أخطاء SYN التي تملك نفس الخدمة	Srv_serror_rate
بيانات مستمرة	النسبة المئوية للاتصالات التي تحوي على أخطاء REJ	rerror_rate
بيانات مستمرة	النسبة المئوية للاتصالات التي تحوي على أخطاء SYN	Serror_rate
بيانات مستمرة	النسبة المئوية للاتصالات التي تحوي على أخطاء REJ التي تملك نفس الخدمة	Srv_rerror_rate
بيانات مستمرة	النسبة المئوية للاتصالات التي تحوي الخدمة نفسها	Same_srv_rate
بيانات مستمرة	النسبة المئوية للاتصالات التي تحوي خدمات مختلفة	diff_srv_rate
بيانات مستمرة	النسبة المئوية للاتصالات التي تملك مضايف مختلفة	Srv_diff_host_rate

جدول رقم (5) ميزات سير الحزم المستندة على المضيف

نوع بيانات الميزة	الوصف	أسم الميزة
بيانات مستمرة	عدد الاتصالات التي تملك المضيف نفسه (هدف)	dst_host_count
بيانات مستمرة	عدد الاتصالات التي تملك المضيف نفسه (هدف) وتستخدم الخدمة نفسها	dst_host_srv_count
بيانات مستمرة	النسبة المئوية للاتصالات التي تملك المضيف نفسه (هدف) وتستخدم الخدمة نفسها	dst_host_same_srv_rate
بيانات مستمرة	النسبة المئوية للخدمات المختلفة على المضيف الحالي	dst_host_diff_srv_rate
بيانات مستمرة	النسبة المئوية للاتصالات القادمة إلى المضيف الحالي وتملك رقم المدخل للمصدر نفسه	dst_host_same_src_port_rate
بيانات مستمرة	النسبة المئوية للاتصالات التي تملك الخدمة نفسها والقادمة من عدة مضايف	dst_host_srv_diff_host_rate
بيانات مستمرة	النسبة المئوية للاتصالات القادمة إلى المضيف الحالي وتملك أخطاء S0	dst_host_serror_rate
بيانات مستمرة	النسبة المئوية للاتصالات القادمة إلى المضيف الحالي والخدمة المحددة وتملك أخطاء S0	dst_host_srv_serror_rate
بيانات مستمرة	النسبة المئوية للاتصالات القادمة إلى المضيف الحالي وتملك أخطاء RST	dst_host_rerror_rate
بيانات مستمرة	النسبة المئوية للاتصالات القادمة إلى المضيف الحالي والخدمة المحددة وتملك أخطاء RST	dst_host_srv_rerror_rate

References

[1] الحمامي علاء حسين ،العاني سعد عبد العزيز ، "تكنولوجيا أمنية المعلومات وأنظمة الحماية " ، جامعة عمان الأهلية (2007).

[2] Vera Marinova-Boncheva, Institute of Information Technologies, supported by The Ministry of Education and Science under Project No. MU-MI-1601 / 2006,Sofia, (2007) .

[3] Theodoros L. and Konstantinos P., Department of Computer Science and Engineering, UC Riverside, Riverside CA 92521(2006).

[4] Adel S. I., M.Sc. Thesis,College of Science, University of Duhok (2009).

[5] Yongro P., Ph.D. Thesis, The Academic Faculty, Georgia Institute of Technology (2005).

[6] Matti M., Helsinki University of Technology (2007), http://www.tml.tkk.fi/Publications/C/25/papers/Manninen_final.pdf .

[7] Alex A.F., University of Kent, Canterbury, UK, Information Resources Management Association(IRMA),pages 6 (2007).

[8] Michael L.G. and Lorraine L.L., Proceedings of Student/Faculty Research Day, CSIS, Pace University, May 7th, (2004).

[9] Mehdi M. and Mohammad Z., School of Computing, Queen University, Canada (2004), http://research.cs.queensu.ca/~moradi/148-04-MM-MZ.pdf.

[10] Yacine B., Mitsubishi Electric ITE-TCL, Rennes, Department RSM GET/ENST Bretagne, France (2007), http://www.diadem-firewall.org/workshop06/papers/monam06-paper-29.pdf.

[11] Wafa' S. A. and Reyadh Sh.N., Al Al-Bayt University, Information Technology College, Jordan, IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.4, April 2009 (2009).

[12] Mansour Sh. and Davood G., World Applied Sciences Journal 7 P.P. 80-86 ISSN 1818-4052 (2009).

[13] 1999 KDD CUP Competition (Available on http://kdd.ice.uci.edu/database /kddcup99/kddcup99.html.

[14] Shall X. Wu and Wolfgang B., Technical Report, Memorial University of Newfoundland, Canada (2008).

[15] The KDD99 Dataset. Retrieved January 26, 2008, from http://kdd. ice.uci.edu/databases/kddcup99/task.html.

[16] Amine B. and Sylvain G., ENST Bretagne 2, rue de la Châtaigneraie F-35576, Cesson-Sévigné, France (2007).

[17] Sophia K. and Nallasamy M., Monash University Australia(2006) .

[18] Wikipedia. Retrieved January 26, 2008, from http://en.wikipedia.org.

[19] Olariu, S. and Zomaya, A. Y. Chapman & Hall/CRC, 2006. ISBN-10:1584884754(2006).

[20] Rafael S.P. and Heitor S. L., CEFET-PR, pp.191-208, Curitiba, Brazil (2002).

[21] Dorigo, M. ; Colorni, A. and Maniezzo, V. , IEEE Trans. Syst. Man Cybern. B, vol. 26, pp. 29–41, Feb (1996).

[22] Dorigo, M. and Caro, G. Di., D. Corne, M. Dorigo, and F.Glover, Eds. New York: McGraw-Hill, pp. 11–32 (1999).

[23] Dorigo, M. ; Caro, G.Di. and Gambardella, L.M. Artif. Life, vol. 5, no. 2, pp. 137–172 (1999).

[24] Xuepeng X. and B.S. , M.Sc. Thesis, College Computer Sciences, Texas Tech University (2004).

[25] Tony W., School of Computer Science, Carleton University, Honours Project （COMP 4905）, (2006) .

[26] Parpinelli, H.S. L. and Freitas, A. A., IEEE Transactions on Evolutionary Computation, Special Issue on Ant Colony algorithms. Vol. 6, No. 4, pp.321-332 (2002).

[27] Parpinelli, H.S.L. and Freitas, A.A., In Data Mining: A Heuristic Approach, H.A. Abbs, R.A. Sarker, C.S. Newton (Eds.), pp. 190-208, Idea Group Publishing, London (2001).

[28] Santhosh S.B.E., M.Sc.Thesis, College of Engineering, Texas Tech University (2006).

[29] Mahmood S.M., M.Sc. Thesis, College of Computer Science (2011). (in Arabic).