DOI: http://dx.doi.org/10.21123/bsj.2016.13.2.0278

تطوير خوارزمية الانتقاء السلبي المناعية لكشف التطفل في مجموعة بيانات NSL-KDD

مفاز محسن خليل العنزي * علاء حازم جار الله**

*قسم علوم الحاسوب، كلية علوم الحاسوب والرياضيات، جامعة الموصل.
** كلية الادارة والاقتصاد، جامعة الموصل.

استلام البحث 2015/9/28
قبول النشر 2015 /11/5

الخلاصة :

مع تطور تقانات الاتصالات من الاجهزة النقالة والاتصالات الالكترونية، وتوجه العالم الى الحكومات الإلكترونية، والتجارة الإلكترونية، والصيرفة الإلكترونية. لذا أصبح من الضروري مراقبة هذه النشاطات ومنعها من التعرض للتطفل أو إساءة الإستعمال وتوفير الحماية لها، لذا فمن المهم تصميم انظمة قوية وكفوءة تقوم بهذا الغرض.

استعمل في هذا البحث عدة اصناف من خوارزمية الانتقاء السلبي المناعية (ذات القيم الحقيقية، خوارزمية الانتقاء السلبي مع كاشفات ذات نصف قطر ثابت، وخوارزمية الانتقاء السلبي مع كاشفات متغيرة الحجم) لكشف التطفل الشبكي من نوع اساءة الاستعمال، حيث تقوم الخوارزمية بتوليد مجموعة من الكاشفات لتمييز عينات الذات.

أثبتت التجارب العملية تحقيق نسبة كشف عالية في النظام المصمم باستعمال بيانات NSL-KDD ذات 12 حقلاً من دون التأثر بتغيير نصف قطر الكاشف أو تغيير عدد الكاشفات إذ تم الحصول على نسبة كشف ما بين (0.984, 0.998, 0.999) و نسبة انذار كاذب ما بين (0.003, 0.002, 0.001). على عكس نتائج التجارب العملية التي أُجريت على بيانات NSL-KDD ذات 41 حقل أتصال، التي تأثرت فيها نسبة الكشف بتغيير نصف قطر وعدد الكاشف اذ تم الحصول على نسبة كشف ما بين (0.44, 0.824, 0.992) ونسبة انذار كاذب ما بين (0.5, 0.175, 0.003).

الكلمات المفتاحية: NSL-KDD ، نظرية تمييز الذات وغير الذات،خوارزمية الانتقاء السلبي ذات القيم الحقيقية، خوارزمية الانتقاء السلبي ذات القيم الحقيقية العشوائية.

المقدمة :

الانتقاء السلبي هو احد آليات نظام المناعة الطبيعي وهو مصدر الهام لمعظم أنظمة المناعة الاصطناعية الحالية. تتم معالجة الخلايا التائية الناضجة في الغدة السعترية قبل أن تنتشر في جهاز المناعة. وتقوم خوارزمية الانتقاء السلبي بتوليد مجموعة من الكاشفات من خلال معاملة أي كاشف مرشح لا يتطابق مع مجموعة عينات الذات. ان أساسيات خوارزميات الانتقاء السلبي استعملت في مختلف المساحات التطبيقية مثل كشف الشذوذ الذي قدم خوارزمية الانتقاء السلبي بوصفها فكرة اساسية للخوارزمية لأجل توليد مجموعة من الكاشفات المرشحة[1،2].

البحوث ذات الصلة:

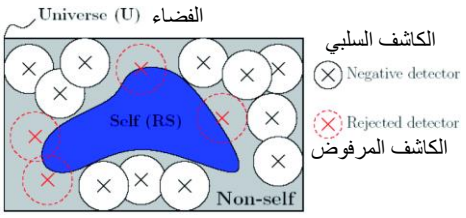
في عام 2010 قدم الباحث Shane Edward Dixon رسالة ماجستير عَرَف فيها النظام المناعي الاصطناعي AIS على أنه حقل نشّأت فيه بحوث الذكاء الحسابي المستوحاة من جهاز المناعة البيولوجية، واستعمل خوارزمية الانتقاء السلبي ذات القيم الحقيقية RNSA بوصفه أنموذجاً حسابيا للتمييز ما بين الذات وغير الذات التي تقوم بها الخلايا التائية في النظام المناعي الطبيعي، حصل نظام المناعي الاصطناعي على نسبة كشف 95.4 [3].

وفي عام 2012 قام الباحثان (Rumsha R.) و (Farrukh A.) باستعمال خوارزمية الانتقاء النسيلي(CLONALG), وخوارزمية الانتقاء السلبي (NSA) في كشف الشذوذ على الشبكة. النتائج التي

حُصِل عليها باستعمال(NSA)، وخوارزمية الانتقاء النسيلي لمجموعة البيانات نفسها، والتي تمت مقارنتها، واظهرت نتيجة المقارنة أن (NSA) أعطت نتائج أفضل من استعمال خوارزمية الانتقاء النسيلي [4].

نظرية تمييز الذات وغير الذات:

التمييز مابين الذات وغير الذات، قد تكون إحدى المهام المهمة في نظام المناعة التكييفة في أثناء العمليات المسببة للأمراض. ومن وجهة نظر اللمفاويات هناك عناصر جيدة وسيئة في الجسم، ويشار إلى العناصر الجيدة بالذات، كما يشار إلى العناصر غير الجيدة بغير الذات [1]. وتكون الحاجة للخلايا اللمفاوية، لتعليم الجسم على ماهية السلوك الصحيح أوغير الصحيح لهذه الأنماط، كما يدرس الطفل من احد الوالدين. وفي الوقت نفسه تحتاج الخلايا اللمفاوية، لتعليم الجسم وبشكل صحيح للرد على الهجمات التي يتعرض لها، ويتم تعليم الخلايا اللمفاوية في الغدة السعترية وفي نخاع العظم في جسم الإنسان، إذ يتعرضون إلى الببتيدات نفسها عن طريق مراحل الاختيار السلبي (NS) والاختيار الايجابي (PS) كما في الشكل (1) [5].



شكل (1) تمييز الذات وغير الذات[5]

أنواع خوارزمية الانتقاء السلبي:

الانتقاء السلبي يستعمل من قبل النظام المناعي للإنسان. ويعد المعالجة الأولية لمضادات الأجسام غير مكتملة النمو. تطلق الغدة السعترية والعظام مضادات أجسام جديدة ويحرص الانتقاء السلبي على إبقاء تلك التي لا ترتبط مع أي من خلايا الذات وتوزع بعد ذلك إلى باقي أجزاء الجسم، لمراقبة الخلايا الحية. لهذا السبب فان حالة الانتقاء السلبي للنظام المناعي للإنسان ضرورية لتؤكد أن مضادات الأجسام المولدة لا تهاجم خلايا الذات [6]. هناك ثلاثة أنواع من خوارزمية الانتقاء السلبي:

1. خوارزمية الانتقاء السلبي ذات القيم الحقيقة

أقترحت خوارزمية الانتقاء السلبي ذات القيم الحقيقية عام 2002. عدة عوامل مهمة تحدد التمييز والكفاءة في خوارزمية الانتقاء السلبي ذات القيم الحقيقة بوساطة تعريف البيانات والكاشفات، التي تمثل بوساطة بيانات القيم الحقيقية. الاختيار المناسب، لقياس المسافة او التشابه بشكل تام يؤدي إلى خوارزمية الانتقاء السلبي ذات القيم الحقيقية. الأول والاهم في خوارزمية الانتقاء السلبي ذات القيم الحقيقية هي مسافة المصفوفة التي تحدد شكل الكاشف في فضاء N ثنائية البعد. ولعل وجود عدة باراميترات سيطرة، يعدل من تأثير أداء واجهة التوليد تجاه المصفوفة وهي آلية مركزية، لتؤدي الخوارزمية وظيفتها. وتعطي مسافة المصفوفة المنفذة حصيلة ثانوية لعدد الكاشفات المولدة وتقدير تغطية الكاشفات [6].

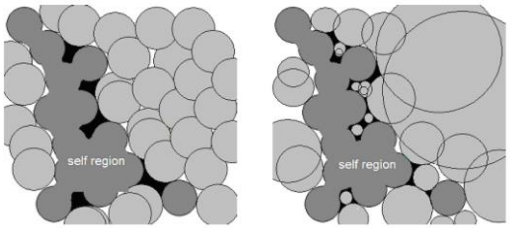
2. خوارزمية الانتقاء السلبي مع نصف قطر ثابت

هي أول خوارزمية انتقاء سلبي نفذت واختبرت وهي الأساس في معالجة التقنيات من قبل Gonzalez and Dasgupta 6. ان مفهوم البيانات ذات القيم الحقيقة يستعمل لتمييز فضاء الذات وغير الذات وتطوير مجموعة كاشفات تعد من مكملات المسافة الجزئية التي تغطي فضاء غير الذات، ومدخلات الخوارزمية هي مجموعة عينات الذات التي تمثل بوساطة نقطة N ثنائية الأبعاد. بعد ذلك تحاول الخوارزمية أن تطور مجموعة أخرى من النقاط تغطي فضاء غير الذات والتي تسمى (كاشفات) وتعد خوارزمية جيدة من خلال المعالجة التي تقوم بتحديث مواقع الكاشفات، وتحقق هدفين أساسيين هما الكاشفات التي يجب أن تترك مجموعة مسافات العتبة، لتغطية الذات، يجب أن تترك مسافات فاصلة بين الكاشفات، لتستطيع ان تغطي جميع نقاط غير الذات. وتنفذ واجهة تنفيذ خوارزمية الانتقاء السلبي ذات القيم الحقيقية مع الكاشفات ذات نصف قطر ثابت، يبدأ بوساطة تعيين قيم لعدة باراميترات سيطرة وهي كما ذكر سابقا مثل عتبة الكاشف التي تمثل قيماً حقيقية وهي العملية التي تمييز ما بين الذات وغير الذات [3].

3. خوارزمية الانتقاء السلبي مع كاشفات متغيرة الحجم

أول تطبيق لخوارزمية الانتقاء السلبي ذات القيم الحقيقية هي توليد كاشفات تحتوي على مسافة عتبة ذات نصف قطر ثابت لمجموعة من الكاشفات في كل المراحل. ومع ذلك فان ميزات الكاشفات يمكن أن تتجاوز هذا الشرط اقترح Zhou and Dasgupta [6] نظاماً جديداً لتوليد الكاشفات ومطابقة آليات خوارزمية الانتقاء السلبي التي قدمت كاشفات مع خصائص متغيرة. تتضمن الخوارزمية المقترحة باراميترات متغيرة جديدة، هي نصف القطر لكل كاشف، والعتبة المستعملة من قبل المسافة المطابقة للقاعدة تحدد نصف قطر الكاشف وهو خيار واضح لجعل النظر في منطقة غير الذات التي تمت تغطيتها بوساطة الكاشفات ومن المرجح ان تكون متغيرة الحجم، الشكل(2) يوضح المرونة التي تقدمها دائرة ذات نصف قطر متغير[3].

يوضح الشكل(2) العديد من المزايا الأساسية لطريقة تنفيذ الكاشفات متغيرة الحجم. أول ميزة واضحة أن مساحة غير الذات يمكن تغطيتها بأقل عدد من الكاشفات. اما المشكلة الأساسية في خوارزمية الانتقاء السلبي، فهي المسافة مابين الكاشفات ونقاط الذات التي لايمكن الاحاطة بها من قبل كاشفات ذات حجم ثابت كما هو موضح باللون الأسود في الشكل ب(2)، ومع ذلك باستعمال الكاشفات بإحجام متغيرة كما في الشكل أ(2) يمكن توليد كاشفات صغيرة، لتغطي مساحات صغيرة بينما الكاشفات الكبيرة تغطي اكبر مساحة من غير الذات. هناك ميزة أخرى في أسلوب الكاشفات متغيرة الحجم لا تظهر بالشكل (2) إذ لها القدرة على التغطية بدلا من زيادة عدد الكاشفات، ويمكن أستعمالها بوصفها باراميترات سيطرة.
ان الخوارزمية لها القدرة على التغطية وبصورة آلية ولها القدرة على توفير معيار التوقف. خوارزمية الانتقاء السلبي للكاشفات ذات الأحجام المختلفة، لها وظائف مماثلة لخوارزمية الانتقاء السلبي مع نصف قطر ثابت مما نوقش سابقا.



أ. كاشفات متغيرة الحجم ب. كاشفات ثابتة الحجم

شكل (2) يوضح المقارنة مابين الكاشفات متغيرة وثابتة الحجم[3]

خوارزمية الانتقاء السلبي ذات القيم الحقيقية العشوائية (RRNS):

الهدف من هذه الخوارزمية هو توليد مجموعة من الكشافات الكروية الضخمة التي تغطي فضاء غير الذات (non-self) وتتكون الخوارزمية من مرحلتين:

أولا: تقوم بإنشاء مجموعة أولية من الكاشفات.

ثانيا: تحسين توزيع هذه المجموعة لتحقيق أقصى قدر من التغطية الواسعة لغير الذات. مدخلات هذه الخوارزمية هي مجموعة عينات من مجموعة الذات S^1 ، تسمح للتغير في مجموعة الذات r_{self}، نصف قطر الكاشف r_{ab} ، ومجموعة البراميترات II ، كما في الخطوات الاساسية الاتية لخوارزمية RRNS [7].

RR-Negative-SELACTION(S^1 ,r_{self}, r_{ab} , II)

S^1 : set of self samples
r_{self} :self variability threshold
r_{ab} : detector radius
II : additional parameters
1:D ⟵ CALCULATE-INIT-DETECTOR-SET(S^1 ,r_{self} , r_{ab})
2: D^1 ⟵ OPTIMIZE-DETECTOR-DISTRIBUTION (D^1 ,r_{ab} ,S^1,r_{self})
3: Return D^1

ووفقا لذلك تحوي الخوارزمية على وظيفتين الاولى : (CALCULATE-INIT- DETECTOR) هي حساب تقديرات لقيمة فضاء (non-self) لانتاج مجموعة جيدة من الكشافات و-(OPTIMIZE-DETECTOR-DISTRBUTION) الثانية هي توزيع الكشافات بشكل متساوي في فضاء (non-self) لتظهر بشكل منظم.

تحديد عدد الكشافات:

نفرض أن V_d هي قيمة لكل كاشف على حدة ونفرض أن $V_{non\text{-}self}$ قيمة فضاء (non-self) يمكن أن يعطى التقريب غير الدقيق لعدد الكاشفات بوساطة:

$$num_{ab} = \frac{V_{non-self}}{V_d} \quad \ldots\ldots (1)$$

نلاحظ أن هذا هو عبارة عن تقريب مفيد جدا وذلك، لأنه يأخذ بالحسبان أن من المستحيل تغطية حجم معلوم من الكاشفات الكروية دون السماح ببعض التداخل، وإذا سمح بالتداخل على نحو فعال فإنه يغطي حجم hyper sphere والتي تعرف بالكاشف ولكن بقيمة أصغر، والذي يمكن تعريفه على انه الحجم المسجل في المكعب الزائدي الذي يغطي الكاشف. ويعمل السبب الرئيس لاختيار هذا التعريف وجود طريقة مباشرة لتغطية (n) من الأبعاد باستعمال مكعبات ذات بعد رابع من دون فتحات. وفقا للمناقشة السابقة فان الحجم الفعال الذي يغطيه الكاشف d مع دائرة نصف قطر r تعرف على النحو الأتي:

$$V_d = \left(\frac{2r}{\sqrt{n}}\right)^n \quad \ldots\ldots (2)$$

باستعمال المعادلتين (1)و(2) فمن الممكن حساب تقريب جيد لعدد الكواشف مع حساب نصف قطر دائرة اللازمة لتغطية مساحة غير الذات ومعرفة مساحة غير الذات [7].

حساب حجم مجموعة الذات وغير الذات:

في معظم الأحيان الإدخال إلى خوارزمية الانتقاء السلبي هي مجموعة فرعية من مجموعة الذات، وعموما فان وحدة التخزين بأكملها في فضاء الذات غير معروفة، وعلى فرض أن نبني انموذج من مجموعة الذات Š ، التي تعرف بانها مجموعة من عينات الذاتS^1، الافتراض الأساسي في هذا التعريف هو احد العناصر التي تكون قريبة بما يكفي لعينات الذات والتي تعرف على نحو الذات. وتم تحديد التقارب من قبل عتبة، rself التي تعرف الحد الأدنى

للتقريب ما بين عينات الذات والعنصر x، يمكن أن يكون x جزءاً من مجموعة الذات وهو انموذج لمجموعة الذات Š ، التي تعرف بالشكل الاتي:

$$\check{S}:=\{x \in U\ \exists_s \in S^1, \|s-x\| \leq r_{self}\} \ldots\ldots\ldots (3)$$

وكما هو معروف فان التقدير الجيد لمتوسط متغير عشوائي (القيمة المتوقع تحصيلها) هي متوسط مجموعة العينات لذا سوف نستخدم المعدل لـ $X\hat{s}(xi)\{i=1\ldots m\}$ لقيمة الذات $V\hat{s}$ وفي التقدير

$$\check{V}_{\hat{s}}=V\hat{s} \approx \frac{\sum_{i=1}^{m} X\hat{s}(xi)}{m} \ldots\ldots\ldots\ldots (4)$$

تعريف التقدير يتجزأ عن طريق حساب متوسط مجموعة من عينات عشوائية كما في تكامل مونتي كارلو الميزة الأساسية في هذا الأسلوب يتعارض مع الأساليب غير الاحتمالية الأخرى إذ أنه من الممكن حساب المدة الفاصلة من الثقة في التقدير الأساسي مستعملة (central limit theorem) اذ من الممكن حساب مدة للثقة باستعمال نظرية النهاية المركزية ومن الممكن حساب هذا الفاصل الزمني [7] :

$$Pr\left(\left|V\hat{s}-\check{V}\hat{s}\right| < \sqrt[3]{\frac{V\hat{s}-\check{V}\hat{s}2}{m}}\right) \approx 0.998 \ldots (5)$$

وفيما يأتي خوارزمية الانتقاء السلبي ذات القيم الحقيقية العشوائية (RRNS)

CALCULATE-INIT-DETECTOR-SET(S^1 , r_{self} , ϵ_{max} ,init-iter)
S^1:set of self samples
r_{self} :self variability threshold
ϵ_{max} : : maximum allowed error
m_{min} : initial number of iteration
r_{ab}: detector radius
n : dimension of the self/non-self space
1: num_hits ← 0
2: m ← 0
3: Repeat
4: m ← m + 1
5: x ←uniformly distributed random sample from $[1, 0]^n$
6: y ←NEAREST-NEIGHBOR(S^1, x)
7: If $\|x-y\| \leq r_{self}$
8: Then num_hits ← num_hits + 1
9: EndIf
10: $\hat{W}_{\hat{s}} \leftarrow \frac{num-hist}{m}$
11: $\epsilon \leftarrow 3\sqrt{\frac{\hat{W}\hat{s}-\hat{W}\hat{s}2}{m}}$
12: Until $m \geq m_{min}$ and $\epsilon \leq \epsilon_{max}$
13: $num_{ab}\left\lfloor\frac{1-\hat{W}\hat{s}}{\left(\frac{2rab}{\sqrt{n}}\right)^n}\right\rfloor$
14: r_{ab} ←
15: D ←^{0}
16: Repeat
17: x ←uniformly distributed random sample from $[1, 0]^n$
18: y ←NEAREST-NEIGHBOR(S^1,x)
19: If $\|x-y\| \geq r_{self}$
20: Then D ← D ∪ {x}
21: EndIf
22: Until |D| = num_{ab}
23: Return D

محاسن خوارزمية الانتقاء السلبي:

يرجع سبب استعمال خوارزمية الانتقاء السلبي لعدة محاسن منها[6] :

1) الكشف عن الحالات الغريبة

إحدى الميزات الهائلة هي أن هذا النهج المستحدث لا يعرف الحالات الشاذة المحددة، ليتم الكشف عنها، ومن ثم فإنه لا يتطلب معرفة مسبقة بالحالات الشاذة .هذه الميزة تسمح لها أن تكون قادرة على الكشف عن الحالات الغريبة المشبوهة التي لم يسبق لها مثيل.

2) القدرة العالية على التكيف

بما أن كل نسخة من الكاشفات فريدة من نوعها ومستقلة لذا فانه من الممكن لكل مضيف مطابقة نسختهم الخاصة من الكاشفات وفقا لاحتياجاتهم الخاصة و بيئة التشغيل.

3) الجمع ما بين الكاشفات الموزعة و المحلية

فضلا عن ذلك، يكون الكاشف موزعا و محليا. و بعبارة أخرى فان الكاشف الواحد يحتوي فقط على مجموعة فرعية من الأنماط اللازمة لوصف جميع الحالات المشبوهة، وتراقب أجزاء صغيرة من النظام فقط، لذلك فان كل كاشف يميز الحالات المشبوهة من قسم صغير من النظام الذي يراقبه فقط، ويتم تشخيص الحالات غير الطبيعية بأكملها من قبل مجموعة من نتائج الكشف المستقلة. فضلا عن ذلك، فأن هذا الكشف الذي وزعته الكاشفات المحلية يوفر متانة داخل النظام.

مساوئ خوارزمية الانتقاء السلبي:

1) وقت حسابي زائد

Excessive Computing time

المشكلة الأكثر أهمية هي الأوقات الحسابية الزائدة الناجمة عن نهج الجيل العشوائي لبناء كاشفات صالحة. هذه النتائج في نمو أسي من الجهد الحسابي مع حجم أنماط ذاتي.

2) عدد الكاشفات يصعب تحديدها مسبقا

فضلا عن ذلك، فإنه من الصعب جدا أن نعرف ما إذا كان عدد من الكاشفات ولدت بحجم كبير كاف يمكنه من تغطية فضاء الكشف.

مجموعة بيانات NSL-KDD:

اقترحت بيانات NSL-KDD من قبل (Tavallaee et al.). وتعد نسخة مختزلة من بيانات KDD الأصلية وتتألف من مميزات البيانات نفسها التي تحتوي في كل سجل اتصال TCP KDD99

على إحدى وأربعين ميزة مع عنوان يوضح هل هذا الاتصال هو اتصال أعتيادي، أو نوع من أنواع الهجمات، وهناك ثمان وثلاثون ميزة رقمية وثلاث ميزات رمزية [8،9].

في مشكلة كشف التطفل يمكن أن يكون محاكاة الهجوم ضمن الأصناف الأربعة الآتية: التجسس (probe)، منع الخدمة (DoS) Denial of Service، الوصول غير مخول إلى مستوى جذر حاسبة الضحية User to Root (U2R)، الوصول غير المخول عن بعد Remote to Local (R2L) [8،9].

معمارية النظام المقترح:

يتكون النظام المقترح شكل (3) من عدة اجزاء وهي:

1. المعالجة الأولية

تمت عملية التحويل بعد دراسة موسعة لبيانات التدريب البالغ عددها 25191 سجل اتصال وكان الهدف منها تحويل القيم غير الرقمية إلى أرقام حقيقية فكانت قيم الصفات (Protocol Type ,Server) قد حُوّلت إلى قيم حقيقية. في بداية الأمر ثم توحيد البيانات لكل حقل اتصال أذ حولت من صيغها المختلفة إلى صيغة الأعداد الحقيقية، لكي تلائم الإدخالات.

تم توحيد التعامل مع هذه القيم بالطريقة الاتية: مثلا يتكون (Protocol Type) من ثلاث قيم هي (TCP,UDP,ICMP) تمثيلهم بالقيم (1.0,2.0,3.0)على التوالي.

أما صفة الخدمة Service فتحوي على 66 قيمة أكثرها تكراراً هي الـ private واقل قيمها تكراراً هي الـ Time_i. أما باقي القيم فرتبت أيضا من الأقل تكراراً إلى الأكثر تكراراً وأعطيت قيم حقيقية بدلا من الرمزية، فمنحت القيم الأقل تكراراً بقيمة (1.0) والأكثر تكرارا (66.0).

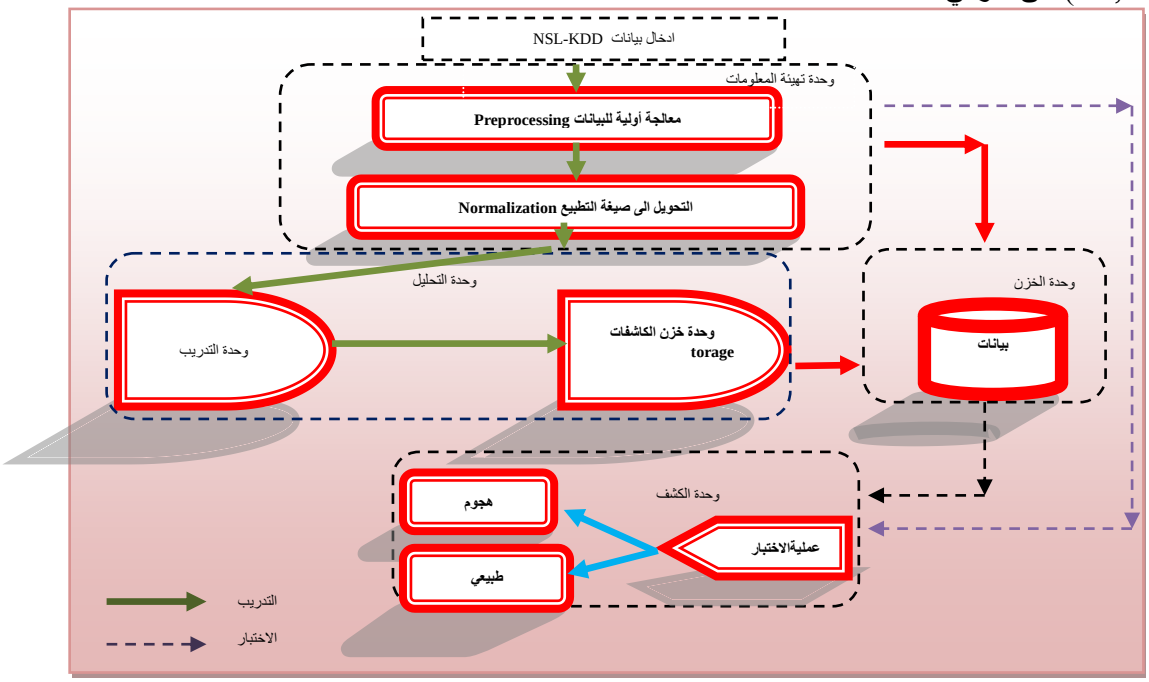
2. البيانات المختارة كإدخال

تتكون مجموعة بيانات NSL-KDD من سجلات حركة مرور شبكة الاتصال الأولية التي تتكون من 41 حقلاً ولكل حقل سمات خاصة تختلف عن الحقل الأخر ويمكن تصنيف هذه الحقول إلى أربع مجاميع.

المجموعة الأولى: حقول 0, 4, 5, 7 هذه الحقول هي الخصائص الأساسية للاتصال، فهي اوقات نقل بحجم ثماني بتات. حيث نتجاهل الحقول , 2 , 3 , 6 1 في مجالات الاتصال.

المجموعة الثانية: الحقول 10-19 هذه الحقول هي في الواقع ليست ميزات حركة المرور، هذه القيم لا يمكن الحصول عليها من خلال مراجعة سجلات المرور لوحدها، فهناك حاجة لتسجيل البيانات بمساعدة المضيف.

المجموعة الثالثة: هي الحقول 22-30 ، هذه الحقول هي الوقت الأساسي لحركة المرور اعتمادا على ميزات الإحصاءات الموجودة على حركة مرور في ثانيتين حيث تستند إلى حساب عنوان IP للمصدر.



شكل (3) معمارية النظام المقترح لكشف التطفل باعتماد النظام المناعي.

المجموعة الرابعة: هي الحقول 31-40 وتشبه المجموعة الثالثة ماعدا انه يتم حساب عنوان IP الموجه إلى الهدف [6].

من خلال المجاميع الأربع يتم أيجاد معدل الكشف، وبصورة خاصة عند استعمال حقول المجموعة الثالثة أو المجموعة الرابعة، حيث يزداد معدل

الكشف. في تجاربنا أختيرت حقول المجموعة الثالثة فضلا عن اختيار الحقول 1 , 2 , 3 [6]. فضلا عن تطبيق البيانات الكاملة بوضعه إدخالا أخر لنظام الكشف المقترح اي 41 حقل اتصال.

3. طرائق التطبيع

بعد إن قمنا بعملية تصحيح قيم سجلات الاتصال في بيانات NSL-KDD إلى قيم حقيقية يجب القيام بتطبيق نظريات التطبيع الـ Normalization على البيانات، لكي تكون جاهزة لمرحلة تكوين الكشافات. هناك عدة طرائق مطبقة في هذا المجال، مثل MIN-MAX، Scaling، Logarithmic وقد استعملت في عدة مجالات من الباحثين وعلى مختلف أنواع البيانات [10]

التطبيع اللوغاريتمي

Logarithmic Normalization

هذه الطريقة تعطي صيغة التطبيع بقيم محصورة بين الصفر والواحد والهدف منها تصغير القيم الى اصغر شيء نتيجة لتطبيق هذه النظرية، وصيغة الحساب لهذه القيمة هي:

القيمة الجديدة بصيغة التطبيع=

$$\frac{\text{لوغاريتم}(\text{للصفةالحاليةالقيمة} \boldsymbol{1+minX-X})}{\text{لوغاريتم}(\boldsymbol{maxX-minX+1})} \quad \ldots\ldots(6)$$

أثبتت التجارب السابقة أن التطبيع اللوغاريتمي يعطي نتائج أفضل في التطبيق من الطرائق الأخرى للتطبيع، لذا تم اعتمادها في هذا البحث.

تصميم نظام كشف التطفل باستعمال خوارزمية الانتقاء السلبي:

يعمل نظام الكشف المقترح بثلاثة أطوار هي:

الطور الاول: تحديد بيانات الذات التي تناظر الأنموذج الطبيعي في بيانات التدريب والاختبار NSL-KDD. وتحديد بيانات غير الذات التي تناظر سجلات الهجوم في بيانات NSL-KDD.

الطور الثاني: توليد كاشفات عشوائية مقارنة إلى كل شكل من أشكال الذات الذي يعرف في الواجهة الأولى. إذ ان أي شكل مولد عشوائيا يطابق شكل الذات سوف يفشل في أن يكون كاشفا وهذا الشكل سوف يُزاح. فيما عدا ذلك سوف يصبح كاشفاً.

الطور الثالث: يقوم النظام باستعمال الكاشفات المولدة من عملية التدريب، لتحديد الشذوذ الظاهر في سجلات الاتصال وإصدار رسائل تحذير عند اكتشافه.

مراحل خوارزمية الانتقاء السلبي لكشف التطفل:

الغرض من خوارزمية الانتقاء السلبي توليد مجموعة من الكشافات اعتماداً على مجموعة التدريب التي تميز نماذج الاتصالات الاعتيادية من الاتصالات الشاذة.

1. مرحلة التدريب Training Stage

تتكون مرحلة التدريب (كما في الشكل 4) من عدة خطوات وعمليات:

1. إدخال نماذج التدريب

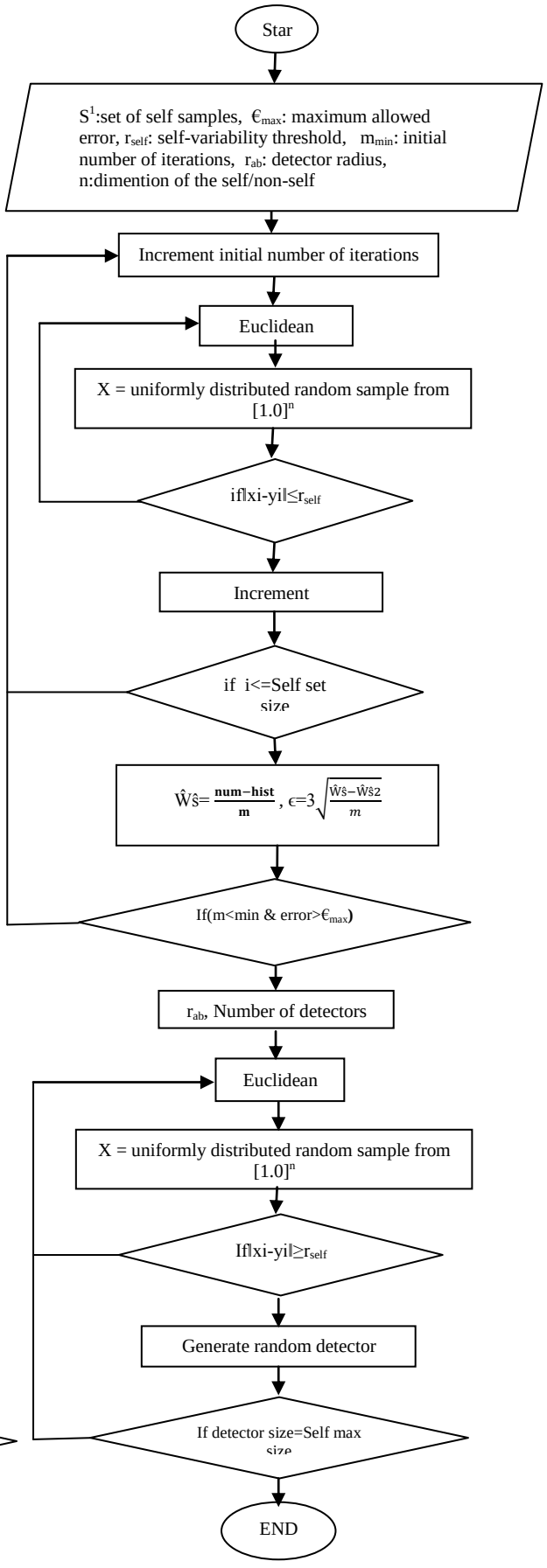
أن مجموعة نماذج التدريب الداخلة إلى النظام تحتاج إلى القيام بعملية معالجة، وتطبيع البيانات. اذ كل أنموذج تدريب يتكون من (41) ميزة يحوي أسم الهجوم في نهايته. هناك مجموعة من المدخلات وهي: S^1مجموعة عينات الذات،R_{self} العتبة،R_{ab} نصف قطر الكاشفات،$€_{max}$ مقدار الخطاء، وn فضاء الذات وغير الذات.

2. فرز مجموعة الذات وغير الذات.
3. توليد الكاشفات ضمن المدى المطلوب.
4. نطبق المعادلة الاقليدية.
5. المقارنة مع شرط العتبة.
6. إيجاد عدد الكاشفات المطلوبة.
7. ايجاد نصف قطر الكاشف.
8. خزن الكاشفات النهائية.

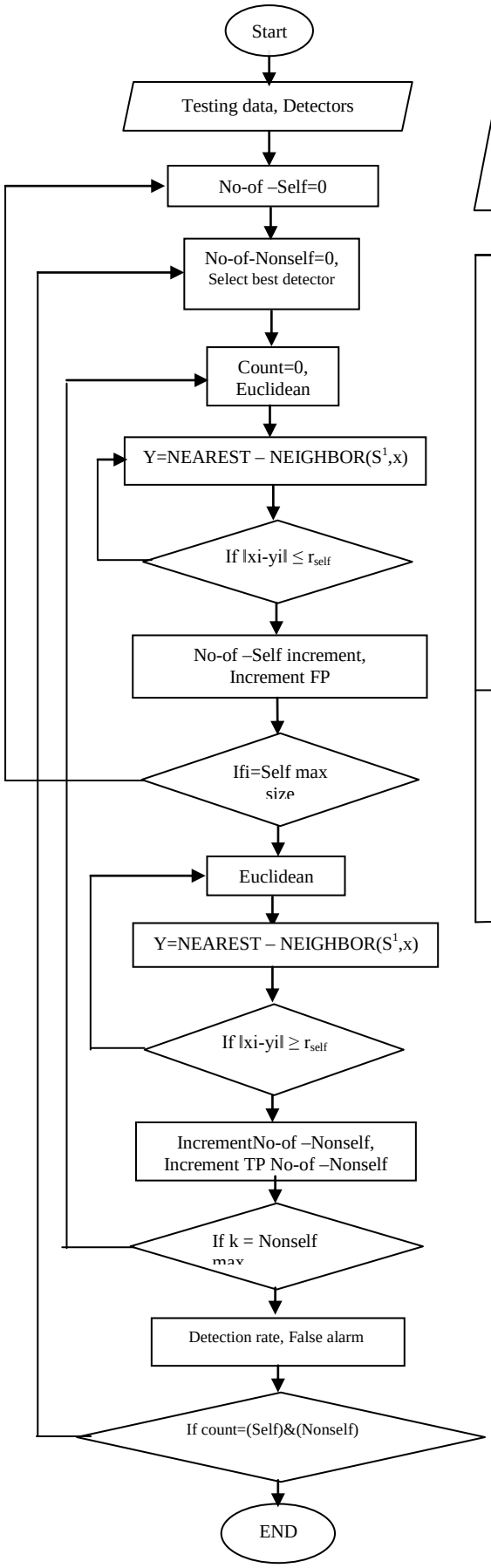
2. مرحلة الاختبار Testing Stage

بعد الانتهاء من مرحلة التدريب والحصول على الكاشفات تبدأ مرحلة الاختبار (شكل 5) كالاتي:

1. إدخال مجموعة نماذج الاختبار NSL-KDD والبالغ عددها 22544 حقل اتصال.
2. عمل تطابق مابين الكاشفات ومجموعة الذات.
3. تطبيق المعادلة الاقليدية، ايجاد عدد الذات وغير الذات.



شكل (4): خوارزمية مرحلة التدريب.



شكل (5): خوارزمية مرحلة الاختبار.

التجارب المنجزة :

طُبق ونُفذ النظام باستعمال لغة فجوال سي شارب (Visual C# 2010) على حاسبة تعمل تحت بيئة التشغيل (Microsoft Windows 7) وتحتوي على ذاكرة حجمها 4.00GB. ومعالج (Intel(R) Core TM) i5-3210M CPU@2.5GHz) .

1. التجربة الأولى

طبقت عمليا لتدريب وأختبار أنظمة كشف التطفل المصممة على بيانات NSL-KDD باستعمال 41 حقل اتصال. في مرحلة التدريب نقوم بإدخال بيانات NSL-KDD للتدريب باستعمال 41 حقل اتصال، نعطي قيمة ثابتة للعتبة (threshold)، 10 قيم متغيرة لنصف قطر الكاشف(detector radius)، وقيمة ثابته لعدد الكاشفات (numab)، وكما موضح بالجدول(1) قيم المدخلات في كل مرحلة.

وتبين نتائج التجربة الأولى أن نسبة الكشف (DR) عندما تكون قيمة نصف قطر الكاشفات (rab) (1) تكون عالية، وقلت هذه النسبة كلما زادت قيمة (rab) عن (1)، وزادت هذه النسبة كلما قلت قيمة (rab) عن (1) ماعدا قيمة (rab) (0.5) التي اصبح فيها انخفاض كبير بنسبة الكشف، وزادت هذه النسبة عندما قلت قيمة (rab) عن (0.5) وصولا الى (0.3) كما ورد في الجدول رقم (1) وموضح بالاشكال (6، 7، 8، 9).

2. التجربة الثانية

في التجربة الثانية تم إجراء عملية التدريب عن طريق إدخال بيانات NSL-KDD للتدريب باستعمال 41 حقل اتصال. وبعدها يتم تثبيت قيمة العتبة وإعطاء قيمة ثابتة لنصف قطر الكاشف(detector radius) في كل مرحلة وبإعطاء قيم متغيرة لعدد الكاشفات (numab)، وكما موضح بالجدول (2) قيم المدخلات في كل مرحلة تدريب ونتائج الاختبار.

تبين من نتائج التجربة الثانية أن نسبة معدل الكشف (DR) زادت بشكل ملحوظ مع تناقص عدد الكاشفات (numab)، وصولا الى عدد الكاشفات (500) وأنخفضت هذه النسبة بشكل ملحوظ، عندما أصبحت عدد الكاشفات (600)، كما ورد في الجدول رقم (2) وموضح بالاشكال (10، 11، 12، 13).

3. التجربة الثالثة

أُجريت هذه التجربة العملية لتدريب واختبار أنظمة كشف التطفل المصممة على بيانات NSL-KDD باستعمال12 حقل اتصال، في مرحلة التدريب يتم إدخال بيانات NSL-KDD للتدريب التي تتضمن 12 حقل اتصال، تعطى قيمة ثابتة للعتبة (threshold)، (10) قيم متغيرة لنصف قطر الكاشف(detector radius)، وقيمة ثابتة لعدد الكاشفات (numab)، كما موضح بالجدول (3) قيم المدخلات في كل مرحلة تدريب ونتائج الاختبار.

وقد تبين من نتائج التجربة الثالثة أن نسبة الكشف (DR) تكون عالية بشكل ملحوظ على الرغم من أن قيمة نصف قطر الكاشفات (rab) متغيرة كما ورد في الجدول رقم (3)، وموضح بالاشكال 14، 15، 16، 17.

4 التجربة الرابعة

أجريت في التجربة الرابعة عملية التدريب عن طريق إدخال بيانات NSL-KDD للتدريب باستعمال12 حقل اتصال. وبعدها يتم تثبيت قيمة العتبة (threshold) وإعطاء قيمة ثابتة لنصف قطر الكاشف(detector radius) في كل مرحلة وإعطاء قيم متغيرة لعدد الكاشفات (numab)، وكما موضح بالجدول (4) قيم المدخلات في كل مرحلة تدريب باستعمال12 حقل اتصال.

تبين من نتائج التجربة الرابعة أن نسبة الكشف (DR) تكون عالية بشكل ملحوظ على الرغم من أن عدد الكاشفات (numab) متغيرة كما ورد في الجدول رقم (4)، وموضح في الاشكال 18، 19، 20، 21.

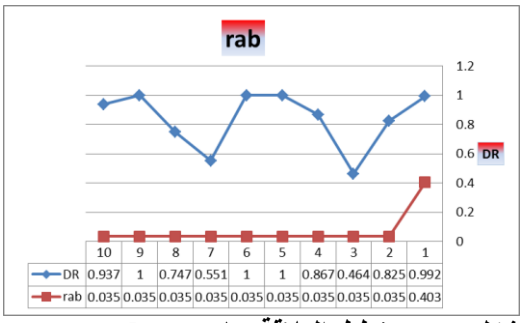
الاستنتاجات:

من خلال ما تقدم من التجارب السابقة، بإدخال بيانات NSL-KDD للتدريب التي تتضمن 41 حقل اتصال، وإدخال بيانات NSL-KDD للتدريب التي تتضمن 12 حقل أتصال. لاحظنا ما يأتي:

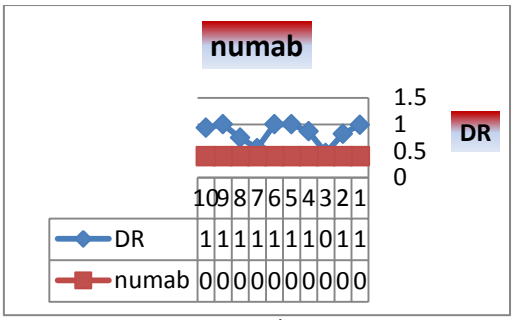
ان استعمال بيانات NSL-KDD للتدريب التي تتضمن 12 حقل أتصال، نحصل على نسبة كشف (DR) عالية جدا ونسبة إنذار كاذب (FAR) قليلة جدا، على الرغم من استعمالقيم متغيرة لنصف قطر الكاشف(detector radius)، واستعمالعدد الكاشفات (numab) متغيرة. بينما باستخدام بيانات NSL-KDD للتدريب التي تتضمن 41 حقل أتصال نحصل على قيم كشف (DR) ونسبة إنذار كاذب (FAR) متغيرة بتغير نصف قطر الكاشف(detector radius) وعدد الكاشفات (numab)، وبهذا نستنتج أن استعمال بيانات NSL-KDD للتدريب التي تتضمن 12 حقل أتصال أفضل من استعمالبيانات NSL-KDD للتدريب التي تتضمن 41 حقل أتصال.

جدول (1) يوضح قيم المدخلات في كل مرحلة تدريب ونتائج الاختبار للتجربة الاولى باستعمال 41 حقل اتصال.

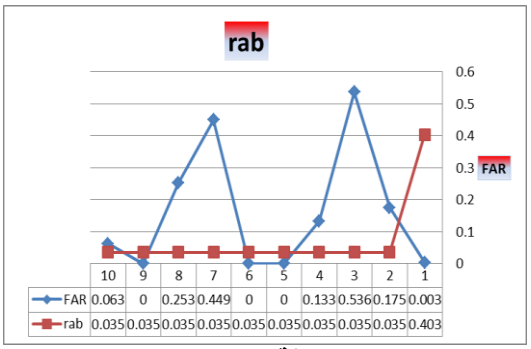
#Exp.	Threshold(rself)	Detector radius		number of directors		DR	FAR
		(rab) in	(rab) result	(numab)in	(numab)resut		
1	0.02	1	0.40311	400	400	0.992145	0.003198
2	0.02	2	0.035355339	400	399	0.824561	0.17543
3	0.02	2.5	0.035355339	400	399	0.46365	0.53634
4	0.02	0.9	0.035355339	400	399	0.86716	0.1328
5	0.02	0.8	0.035355339	400	399	1	0
6	0.02	0.7	0.035355339	400	399	0.9999	0.0001
7	0.02	0.6	0.035355339	400	399	0.55137	0.44862
8	0.02	0.5	0.035355339	400	399	0.74686	0.25313
9	0.02	0.4	0.035355339	400	399	1	0
10	0.02	0.3	0.035355339	400	399	0.937343	0.062656



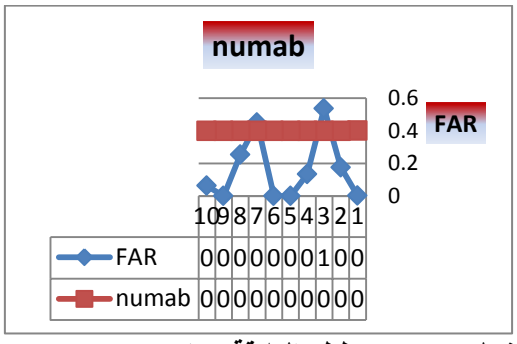
شكل (6) مخطط العلاقة ما بين (rab) و(DR) باستعمال 41 حقل اتصال.



شكل (8) مخطط العلاقة ما بين (numab) و(DR) باستعمال 41 حقل اتصال.



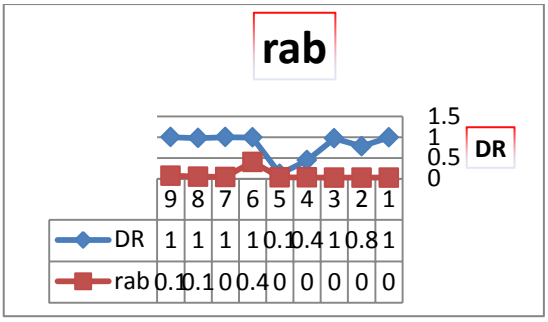
شكل (7) مخطط العلاقة ما بين (rab) و(FAR) باستعمال 41 حقل اتصال.



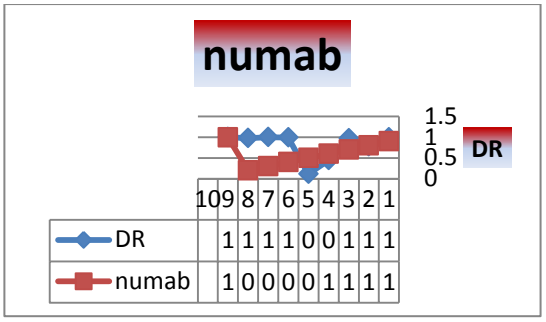
شكل (9) مخطط العلاقة ما بين (numab) و(FAR) باستعمال 41 حقل اتصال.

جدول (2) يوضح قيم المدخلات في كل مرحلة تدريب ونتائج الاختبار للتجربة الثانية باستعمال 41 حقل اتصال

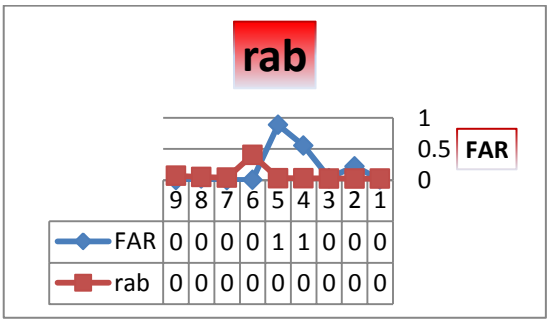
#Exp.	Threshold(rself)	Detector radius		number of directors		DR	FAR
		(rab) in	(rab) result	(numab)in	(numab) result		
1	0.02	1	0.02357022	900	900	0.991729	0.00786
2	0.02	1	0.025	800	800	0.779171	0.22082
3	0.02	1	0.02672612	700	700	0.96714	0.03285
4	0.02	1	0.0288675	600	597	0.4463	0.5536
5	0.02	1	0.03162277	500	499	0.11422	0.88577
6	0.02	1	0.40311	400	400	0.992145	0.003198
7	0.02	1	0.0408248	300	300	0.9999	0.0001
8	0.02	1	0.05	200	200	0.975	0.025
9	0.02	1	0.0707106	100	99	0.9999	0.0001



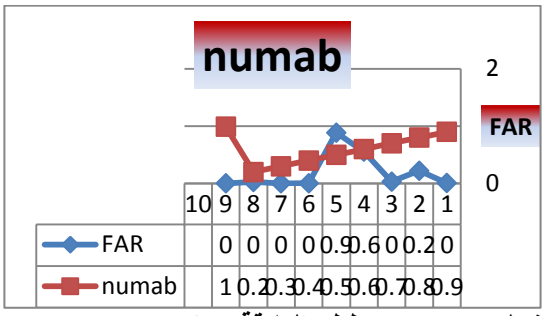
شكل (10) مخطط العلاقة ما بين (rab) و(DR) باستعمال 41 حقل اتصال.



شكل (12) مخطط العلاقة ما بين (numab) و(DR) باستعمال 41 حقل اتصال.



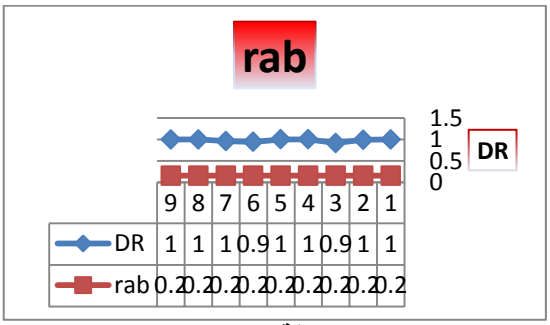
شكل (11) مخطط العلاقة ما بين (rab) و(FAR) باستعمال 41 حقل اتصال.



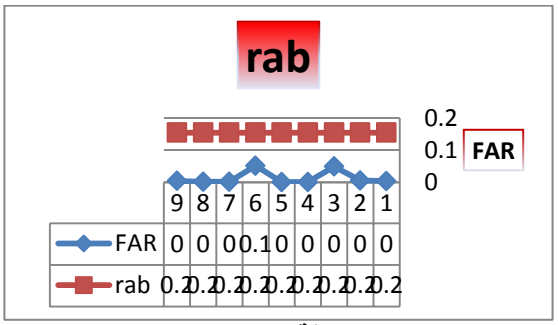
شكل (13) مخطط العلاقة مابين (numab) و(FAR) باستعمال 41 حقل اتصال.

جدول (3) يوضح قيم المدخلات في كل مرحلة تدريب ونتائج الاختبار للتجربة الثالثة باستعمال 12 حقل اتصال

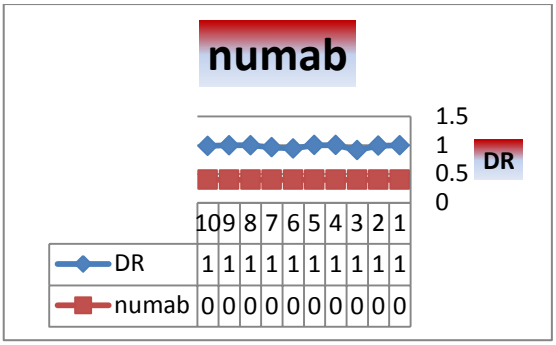
#Exp.	Threshold(rself)	Detector radius		number of directors		DR	FAR
		(rab) in	(rab) result	(numab)in	(numab) result		
1	0.0005	1	0.15411035	400	400	0.99552	0.0034
2	0.0005	2	0.15411036	400	400	0.98791	0.00554
3	0.0005	3	0.15411037	400	400	0.9139	0.0486
4	0.0005	4	0.15411038	400	400	0.99941	0.00036
5	0.0005	0.9	0.15411039	400	400	0.99941	0.00036
6	0.0005	0.8	0.15411040	400	400	0.94002	0.05144
7	0.0005	0.7	0.15411041	400	400	0.95889	0.00175
8	0.0005	0.6	0.15411042	400	400	0.99652	0.00167
9	0.0005	0.5	0.15411043	400	400	0.996	0.004
10	0.0005	0.4	0.15411044	400	400	0.9846	0.0031



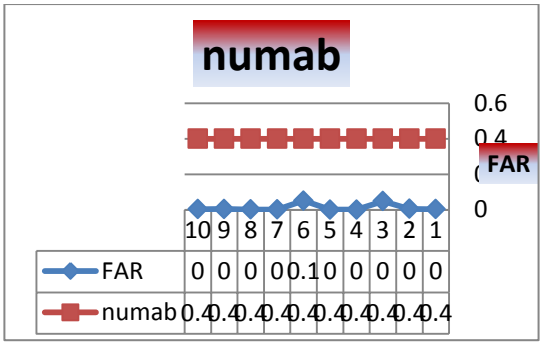
شكل (14) مخطط العلاقة ما بين (rab) و(DR) باستعمال12 حقل اتصال.



شكل (15) مخطط العلاقة ما بين (rab) و(FAR) باستعمال12 حقل اتصال.



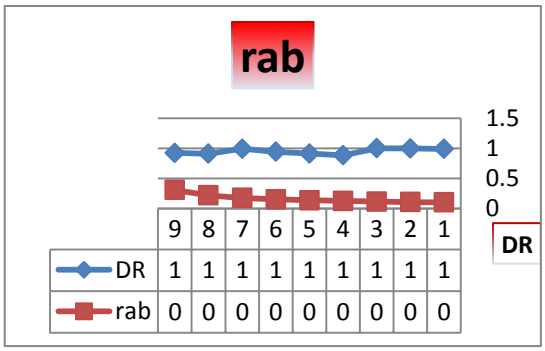
شكل (16) مخطط العلاقة ما بين (numab) و(DR) باستعمال12 حقل اتصال.



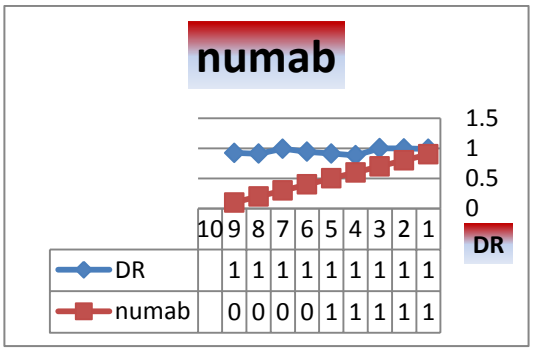
شكل (17) مخطط العلاقة ما بين (numab) و(FAR) باستعمال12 حقل اتصال.

جدول (4) يوضح قيم المدخلات في كل مرحلة تدريب ونتائج الاختبار للتجربة الرابعة باستعمال12 حقل اتصال

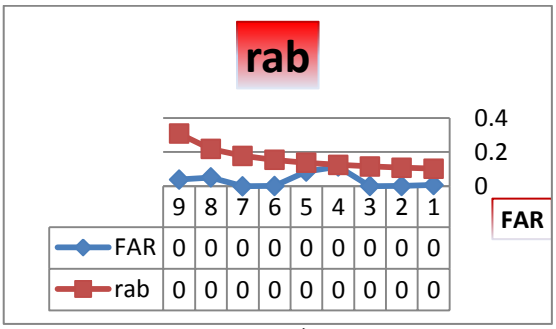
#Exp.	Threshold(rself)	Detector radius		number of directors		DR	FAR
		(rab)in	(rab)result	(numab)in	(numab) result		
1	0.0005	0.4	0.102740	900	900	0.9896	0.0074
2	0.0005	0.4	0.108972	800	800	0.998	0.002
3	0.0005	0.4	0.116496	700	700	0.999	0.001
4	0.0005	0.4	0.125830	600	600	0.88302	0.1150
5	0.0005	0.4	0.137840	500	500	0.91305	0.08630
6	0.0005	0.4	0.154110	400	400	0.94277	0.00231
7	0.0005	0.4	0.177951	300	300	0.9919	0.00025
8	0.0005	0.4	0.217944	200	200	0.91111	0.05112
9	0.0005	0.4	0.308220	100	100	0.9220	0.0395



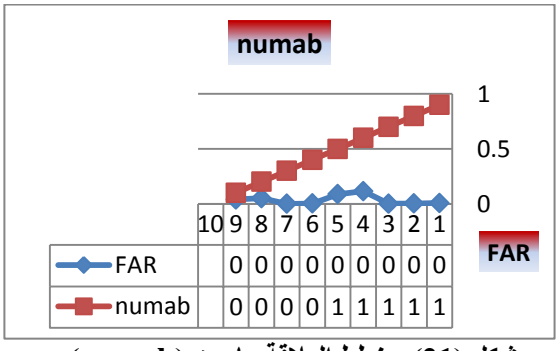
شكل (18) مخطط العلاقة ما بين (rab) و(DR) باستعمال12 حقل اتصال.



شكل (20) مخطط العلاقة ما بين (numab) و(DR) باستعمال12 حقل اتصال.



شكل (19) مخطط العلاقة ما بين (rab) و(FAR) باستعمال12 حقل اتصال.



شكل (21) مخطط العلاقة ما بين (numab) و(FAR) باستعمال12 حقل اتصال.

المصادر :

[1]Aldabagh N. and Ali I., 2011. Design and implementation of artificial immune system for detecting flooding attacks, High Performance Computing and Simulation (HPCS), International Conference on 08/2011.

[2]Al-Enezi J. R.; Abbod M. F. and Alsharhan S., 2010. Artificial Immune System –Models ,Algorithm and Application, Electric and Computer Engineering Department: School of Engineering and Design, Bronal University ,UK, Computer Science Department ,Golf University for Science and Technology Hawalli ,32093,Kuate.

[3]Dixon Shane E., 2010. Studies on Real-Valued Negative Selection Algorithm for Self-Non Self Discrimination, In Partial Fulfillment of The Requirements for The Degree Master of Science in Electrical Engineering, California Polytechnic State University, San Luis Obispo.

[4] Rizwan, R. and Khan A. F., 2012. Artificial Immune System for Anomaly Detection in Wireless Sensor Network, ICARIS Conference Program, IVSL.

[5]Al-Anezi M. and Aldabagh N., 2011. An Immune Inspired Multilayer IDS, InternationalJournal ofComputer Science and Information Security, Vol. 9 No. 10, Paper 30091144, pp. 30-39.

[6]Forrest S.; Allen L.; Perelson A. and Cherukuri R., A Change-Detection Algorithm Inspired by the Immune System, Submitted to IEEE Transactions on Software Engineering, 1995.

[7]Gonzalez F., Dasgupta D. and Nifio L. F., 2003. A Randomize Real-Valued Negative Selection Algorithm, Depto de ing. De Sistemas, Universidad Nacional de Colombia, Bogota, Colombia, Division of Computer Science , The University of Memphis , Memphis TN 38152, IVSL.

[8]Ibraheem, N.; Jawhar, M. and Osman, H., 2013. Principle Components Analysis and Multi Layer Perceptron Based Intrusion Detection System, AL-Rafidain Journal of Computer Sciences and Mathematics, Volume: 10 Issue: 1 Pages:127-135.

[9]Dhanabal L. and Shantharajah S., 2015. A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 6, Pages 446-452.

[10] Zhang T. and You X., 2015. Improvement of the training and normalization method of artificial neural network in the prediction of indoor environment, Elsevier Procedia Engineering 121, Pages: 1245 – 1251.

Developing an Immune Negative Selection Algorithm for Intrusion Detection in NSL-KDD data Set

*Mafaz Mohsin Khalil Alanezi** *Alaa' Hazim Jar Allah***

*Computer Sciences Department, College of Computer Sciences & Mathematics, University of Mosul.
**College of Administration & Economics, University of Mosul.

Received 28/9/ 2015
Accepted 5/11/ 2015

Abstract:

With the development of communication technologies for mobile devices and electronic communications, and went to the world of e-government, e-commerce and e-banking. It became necessary to control these activities from exposure to intrusion or misuse and to provide protection to them, so it's important to design powerful and efficient systems do this purpose.

It this paper it has been used several varieties of algorithm selection passive immune algorithm selection passive with real values, algorithm selection with passive detectors with a radius fixed, algorithm selection with passive detectors, variable-sized intrusion detection network type misuse where the algorithm generates a set of detectors to distinguish the self-samples.

Practical Experiments showed the process to achieve a high rate of detection in the system designer using data NSL-KDD with 12 field without vulnerability to change the radius of the detector or change the number of reagents were obtained as the ratio between detection (0.984, 0.998, 0.999) and the ratio between a false alarm (0.003, 0.002, 0.001). Contrary to the results of experiments conducted on data NSL-KDD with 41 field contact, which affected the rate of detection by changing the radius and the number of the detector as it has been to get the proportion of uncovered between (0.44, 0.824, 0.992) and the percentage of false alarm between (0.5, 0.175, 0.003).

Key words: NSL-KDD, Self-NonSelf Theory, RNS, RRNS.