Application of the Genetic Algorithm in the Network Intrusion Detection System Using NSL-KDD Data

Naglaa B. Ibrahim **Hana M. Usman**

College of Computer Science and Mathematics
University of Mosul, Mosul, Iraq

Received on: 18/10/2011 **Accepted on: 15/02/2012**

ABSTRACT

With the development of the Internet, technological innovation and the availability of information emerged new computer security threats. The researchers are developing new systems known as Intrusion Detection Systems IDSs for detecting the known and unknown attacks. IDS have two approaches depending on the detecting theories: Misuse Detection and Anomaly Detection.

This paper aims to design and implement a misuse network intrusion detection system based on Genetic Algorithm. The efficiency of using GA for building IDS based on NSL-KDD is verified. For rules generation NSL-KDD Data Set is used which include, KDDTrain and KDDTest, 125973 and 22544 records respectively, each record consists of 41 features and one class attribute for specifying normal and abnormal connection (complete train and test data are used), In order to get rid of redundancy and inappropriate features Principal Component Analysis (PCA) is used for selecting (5) features.

Number of experiments have been done. The experimental results show that the proposed system based on GA and using PCA (for selecting five features) on NSL-KDD able to speed up the process of intrusion detection and to minimize the CPU time cost and reducing time for training and testing, that the detection rate: 91.6% and false alarm is: 0% and classification rate (DoS 93.48 %), (Normal 99.52%) , (Probe 81.16%), (R2L 69.47%), (U2R 32.84%). C# programming language is used for system implementation.

Keywords: Genetic Algorithm, Network Intrusion Detection System, NSL-KDD Data.

تطبيق الخوارزمية الجينية في نظام كشف التطفل الشبكي باستخدام بيانات NSL- KDD

نجلاء بديع إبراهيم **هناء محمد عصمان**

كلية علوم الحاسوب والرياضيات، جامعة الموصل

تاريخ استلام البحث: 2011/10/18 **تاريخ قبول البحث: 2012/02/15**

الملخص

مع تطور الانترنت وابتكار التكنولوجيا وتوفر المعلومات ظهرت تهديدات أمنية جديدة للشبكات. وبغية الكشف عن الهجمات الجديدة والمعروفة طور الباحثون أنظمة أمنية عرفت بأنظمة كشف التطفل IDS (Intrusion Detection Systems). هناك نوعان من أنظمة كشف التطفل اعتمادا على نظريات الكشف, هما كشف إساءة الاستخدام, وكشف الشذوذ.

يهدف البحث إلى تصميم وتنفيذ نظام كشف التطفل الشبكي (Network Intrusion Detection System) (NIDS) بكشف إساءة الاستخدام Misuse Detection واعتمادا على الخوارزمية الجينية. ولتوليد القواعد تم اعتماد مجموعة بيانات NSL-KDD. التي تتضمن بيانات تدريب (125973) سجل اتصال وبيانات اختبار (22544) سجل اتصال, حيث يحتوي كل سجل اتصال على 41 ميزة مع عنوان يحدد نوع الاتصال أهو

طبيعي أم شاذ, (وتم استخدام مجاميع التدريب والاختبار كاملة), وللتخلص من الميزات الفائضة والميزات القليلة الفائدة تم استخدام خوارزمية تحليل المركبات الأساسية PCA لاختيار (5) ميزات.

تم إجراء عدد من التجارب المختلفة, وأظهرت النتائج التجريبية أنّ النظام المقترح للخوارزمية الجينية مع PCA باختيار خمس ميزات على بيانات NSL-KDD قادر على تسريع عملية الكشف عن التطفل وتصنيفها مع تقليل زمن المعالج وتقليل زمن التدريب والاختبار, حيث بلغت نسبة الكشف Detection Rate (91.6%) ونسبة الإنذار الكاذب (False Alarm Rate) بمقدار (0%) ، و كانت نسب التصنيف (Classification Rate) (DoS 93.48 %) ، (Normal 99.52 %)، (Probe 81.61 %), (R2L 69.47 %), (U2R 32.84 %). استخدمت لغة فيجوال سي شارب (Visual C# 2008).

الكلمات المفتاحية: الخوارزمية الجينية، نظام كشف التطفل الشبكي، بيانات NSL- KDD.

1. المقدمة

إنّ الاستخدام المتزايد لشبكات الحاسوب في العديد من جوانب الحياة أدى إلى زيادة حالات التعرض للهجمات, وازدياد عدد الثغرات (Vulnerabilities), واستهلاك مصادر الحاسوب, واختراق السرية (Confidentiality) وتكامل (Integrity) الأنظمة [12].

وقد استخدمت عدة آليات لتوفير الحماية مثل الجدار الناري والمضادات الفيروسية والتشفير وغيرها من الآليات الأمنية, غير أنّ المخاطر مازالت موجودة, فكلما قامت مؤسسة أمنية بمعالجة مشكلة ما، قام المهاجمون بالبحث لإيجاد ثغرات جديدة وابتكار هجمات غير معروفة لمهاجمة النظام, مما أدى إلى بناء أنظمة كشف التطفل لمراقبة سلوك النظام ولتحديد التهديد الداخلي من قبل المتطفلين [4].

قدم أول أنموذج لكشف التطفل من قبل الباحثة Denning في سنة 1980. وكانت محاولات الباحثة متمركزة على كيفية تكوين نماذج كشف التطفل بشكل كفوء و دقيق. وبين نهاية عام 1980 وبداية عام 1990 تم الدمج بين الأنظمة الخبيرة مع الطرائق الإحصائية التي كانت الأكثر شيوعاً. وإن هذه النماذج مشتقة من فضاء المعرفة لخبراء الأمنية وفي عام 1990 حولت المعرفة المكتسبة لسلوك الاتصال الاعتيادي والاتصالات غير الاعتيادية من النظام اليدوي إلى النظام الأوتوماتيكي [10,6]. وأعقب هذه الفترة اهتمام بحثي واسع في تطوير الجيل القادم من نظم كشف التطفل والتي كان لها القدرة على التعلم والتكيف مع البيئة الشبكية للحصول على أفضل أداء. واستخدمت هذه النظم الجديدة والتقنيات الذكائية مثل الشبكات العصبية, والمنطق المضبب (Fuzzy Logic), وأشجار المصنف, والخوارزمية الجينية فضلاً عن تنقيب البيانات (Data Mining), وتقنية الوكيل المتنقل (Mobile Agent) في أنظمة كشف التطفل الموزعة [14].

اتجه الباحثون إلى استخدام المفاهيم الذكائية لحل مشكلات أمنية. استخدم الباحث (Adhitya Chittur) [5] منهجاً فريداً للكشف عن التطفل باستخدام الخوارزمية الجينية واختبار كون هذه الخوارزمية خيارا ممكننا لتوليد الأنموذج في أنظمة كشف التطفل, معتمدا على الذكاء الاصطناعي. وأثبتت النتائج إن الخوارزمية الجينية قادرة بنجاح على توليد نموذج سلوكي تجريبي دقيق من تدريب البيانات والقدرة على تطبيق المعرفة التجريبية بنجاح من بيانات لم يسبق التعامل معها.

وفي عام 2007 اقترح الباحثون (Zorana Bankovic et al.) [24] تقليل الميزات باستخدام تحليل المركبات الأساسية (تم اختيار ثلاث ميزات). تم تطبيق الخوارزمية الجينية على مجموعة مختارة من بيانات التدريب والاختبار(KDD), كما تم تصنيف مجموعة مختارة من الهجمات.

وفي عام 2010 قدم الباحثون (Shilpa lakhina et al.) [16] نظام كشف تطفل الشذوذ باستخدام خوارزمية تحليل المركبات الأساسية والشبكات العصبية الاصطناعية (PCANNA). وقللت الخوارزمية المقترحة وقت التدريب والاختبار إلى (40%), (70%) على التوالي, وقد استخدمت مجموعة مختارة من نماذج التدريب والاختبار NSL− KDD.

وفي عام 2010 اقترح الباحثان (Ritu Ranjani و Prof. Neetesh Gupta) [22] استخدام تقنيات الحوسبة المرنة (Soft Computing) لكشف التطفل, حيث استخدمت الشبكة العصبية الاصطناعية (SOM) Self Organize Map مع K-Means Cluster, تم تدريب النظام على مجموعة فرعية من بيانات NSL-KDD. وقد أوضحت النتائج أن نسبة كشف التطفل لشبكة (SOM) أفضل من خوارزمية K-Means علماً أن نسبة كشف التطفل لشبكة (SOM) بلغت (64%) ونسبة الإنذار الكاذب (0.5%), أما نسبة كشف التطفل لخوارزمية K-Means بلغت (60%) ونسبة الإنذار الكاذب بلغت (1%).

استخدمت الخوارزمية الجينية لتطوير قواعد للوصول إلى الشبكة, وهذه القواعد تستخدم للتمييز بين الاتصالات الطبيعية والشاذة. وتشير الاتصالات الشاذة إلى احتمالية وجود تطفل.

تم استخدام مجاميع NSL-KDD [7] على التطفلات التي أصبحت معياراً فعليا لاختبار أنظمة كشف التطفل. والغاية من استخدام هذه البيانات كونها القاعدة الأساسية المشتركة لأغلب الباحثين العاملين في مجال أنظمة كشف التطفل والتي من خلالها يتم المقارنة بين التقنيات الأخرى.

استخلاص الميزات أحد المواضيع الرئيسية في أنظمة كشف التطفل, فهو يحسن أداء التصنيف بالبحث عن مجموعة فرعية للميزات, وتصنيف بيانات التدريب بصورة أفضل. وفي مسالة فضاء الميزات البعدية يمكن لبعض الميزات أن تكون قليلة الفائدة أو عديمة الفائدة وإزالتها مهمة جدا, مع ذلك قد تتسبب في إرباك أداء المصنفات. وهذا مهم جدا عند تفضيل الكشف الفعلي لأنظمة كشف التطفل. لذلك تم في هذا العمل اختيار أفضل الميزات باستخدام تحليل المركبات الأساسية PCA.

2. مدخل إلى أنظمة كشف التطفل [6].

يعرف كشف التطفل بأنه عملية المراقبة الذكية للأحداث التي تحدث في أنظمة الحاسوب أو الشبكة وتحليل تلك الأحداث للإشارة إلى حالة التطفل إن وجدت. إنّ الهدف الأساسي للنظام هو تحقيق تكامل النظام, والسرية وتوافر مصادر المعلومات .

يمكن تصنيف أنظمة كشف التطفل من ناحية نظريات الكشف :

- **كشف إساءة الاستخدام (Misuse Detection)**

تقوم أنظمة كشفِ الإساءة بشكلٍ أساسي بتعريف ما هو "الخطأ" في النظام تحت المراقبة. وتحتوي هذه الأنظمة على أوصاف للهجمات أَو ما يسمى "التواقيع" (Signatures) وتقوم بمقارنتها بسيل البيانات المتدفقة إلى النظام، والبحث عن دليل وشاهد للهجماتِ المعروفةِ وهو توقيع الهجمة [10].

- **كشف الشذوذ (Anomaly Detection)**

تستخدم هذه التقنية في الكشف عن نماذج للسلوك الطبيعي المطلوب والمرتقب لمستخدمي النظام وتطبيقاته، و من ثم ترجمة الانحرافات مِن هذا السلوكِ بأنها مشكلة حدوث أي حالة شذوذ. إن الافتراض البسيط المبني عليه تقنية كشفِ الشذوذ هو أن سلوك النظام تحت المراقبة سيختلفُ عن السلوك الطبيعي له في حال حدوث حالات للتطفل [1].

هناك نوعان من كشف الشذوذ, هما كشف الشذوذ الثابت (Static Anomaly Detection) وكشف الشذوذ الديناميكي (Dynamic Anomaly Detection) [10].

- **كشف المحددات** (Specification Detection)

تحدد هذه التقنية كون سلسلة من الايعازات اخترقت معرفة خصائص سلوك البرنامج. افترضت هذه النظرية بديلاً واعداً يمزج قوة الكشف عن إساءة الاستخدام وكشف الشذوذ. إنّ كشف المحددات يمتلك إمكانية توفير نسبة خطأ منخفضة, ومن الصعب نمذجة برامج أو أنظمة معقدة وكتابة محددات أمنية لها [13].

تقسم أنظمة كشف التطفل طِبقاً لمصدر معلوماتها إلى ثلاثة أقسام :

نظام كشف تطفل المضيف, ونظام كشف التطفل الشبكي, وأنظمة كشف التطفل الهجيني.

2-1 مجموعة البيانات المستخدمة (Data Set)

خلال العقد الأخير، استخدم العديد من الباحثين بيانات KDD Cup 1999 [8] لبناء نظم كشف التطفل. أظهرت الدراسات السابقة وجود بعض المشكلات الكامنة في هذه البيانات. إنّ التحديد المهم لهذه البيانات هو العدد الهائل للسجلات الزائدة بمعنى إن% 78 من سجلات التدريب و 75% من سجلات الاختبار متكررة ، كما موضح في الجدولين (1) و(2)، مما يتسبب في تحيز خوارزمية التعلم نحو السجلات الأكثر استخداماً، ويمنعها من اكتشاف سجلات الهجوم الخلفية المنطويــة تحت أصناف (U2R) Users to Root, Remote to Local (R2L). في الوقت نفسه، تسبب تحيز نتائج التقييم بطرائق تمتلك نسباً أفضل على السجلات المتكررة. بالرغم من ذلك, هذه البيانات تعاني من بعض المشكلات وقد لا تكون مثلى للشبكات الفعلية الموجودة، ويمكن أن يكون محاكاة الهجوم ضمن أحد الأصناف الأربعة الآتية [9]:

1. التجسس (Probe): التجسس والمراقبة وهذا النوع من الهجوم يجمع معلومات عن النظام قبل البدء بالهجوم, مثل (Satan, ipsweep, nmaps) وغيرها.
2. منع الخدمة Denial of Servic (DoS) : ينتج عن منح طلبات مشروعة لمصادر الشبكة باستهلاك حزمة أو استهلاك مصادر حاسوبية, مثلSmurf, Neptune Teardrop, وغيرها.
3. الوصول غير المخول إلى مستوى شرفية الجذر لحاسبة الضحية User to Root (U2R): يبدأ المهاجم بالوصول إلى حساب مستخدم طبيعي على النظام, ويتمكن من استغلال ضعف النظام للحصول على وصول جذري إلى النظام, مثل loadmodule eject, وغيرها.
4. الوصول غير المخول عن بعد Remote to Local (R2L): في هذه الحالة لا يملك المهاجم حساباً على آلة بعيدة, ويرسل حزمة لتلك الآلة عبر شبكة ويستغل بعض الضعف للحصول على وصول موقعي لمستخدم لتلك الآلة.

الجدول (1). السجلات المتكررة في مرحلة تدريب البيانات [9]

	السجلات الأصلية	السجلات المُتميّزة	
سجلات الهجوم	3, 925, 650	262, 178	93. 32%
السجلات الاعتيادية	972, 781	812, 814	16. 44%
المجموع	4, 898, 431	1, 074, 992	78. 05%

الجدول (2). السجلات المتكررة في مرحلة اختبار البيانات [9]

	السجلات الأصلية	السجلات المُتميّزة	
سجلات الهجوم	250, 436	29, 378	88. 26%
السجلات الاعتيادية	60, 591	47, 911	20. 92%
المجموع	311, 027	77, 289	75. 15%

مجاميع البيانات المتولدة, KDD Train , KDD Test شملت (125973, 22544) سجلاً علـــى التوالي.

اقتُرحت بيانات NSL-KDD من قبل (Tavallaee et al.) لحل مشكلات بيانات KDD المذكورة سابقاً. تعتبر NSL-KDD نسخة مختزلة من بيانات KDD الأصلية وتتألف من نفس ميزات بيانات KDD 99 التي تحوي في كل سجل اتصال TCP على إحدى وأربعين ميزة مع عنوان يوضح هل هذا الاتصال هو اتصال اعتيادي أو نوع من أنواع الهجمات، وهناك ثمانٍ وثلاثون ميزة رقمية وثلاث ميزات رمزية .فيما يأتي فوائد NSL-KDD مقارنة بمجموعة بيانات KDD الأصلية [16]:

- لا تشمل سجلات زائدة في مجموعة التدريب, لذا لن تميل المصنفات باتجاه سجلات أكثر حدوثا.
- عدد السجلات المختارة من كل مجموعة: مستوى الصعوبة يتناسب عكسيا ونسبة السجلات في مجموعة بيانات KDD الأصلية. نتيجة لذلك نسب تصنيف طرائق تعليم الآلة المتميزة تختلف بمدى واسع, مما يجعل من زيادة الفعالية امتلاك تقييم دقيق لتقنيات تعليم مختلفة.
- عدد السجلات في التدريب ومجاميع الاختبار معقول, مما يجعل من المحتمل إجراء تجارب على المجموعة الكاملة دون الحاجة للاختبار العشوائي لنسبة ضئيلة. ونتيجة لذلك سيكون تقييم نتائج البحوث المختلفة ثابتة ومشابهة.

3. الخوارزمية الجينية Genetic Algorithm

تُعد الخوارزميات الجينية تقنيات أمثلية (Optimization) تستخدم عملية تطورية. وحل المشكلة يتمثل كهيكل بيانات يعرف بالكروموسومات. ويتم تقييم جودة الحل بدالة تسمى دالة التقييم (Fitness Function), وتتولد سلسلة من الحلول الأولية (مجتمع عشوائي) من خلال مزيج من العمليات المتشابهة لعملية تطورية، وتتجه العملية نحو تطوير حلول تمتلك جودة أفضل عند حساب دالة التقييم والخوارزميات الجينية هي طرائق للبحث، والامثلية, وتعليم الماكنة المتوخاة بالمبادئ الطبيعية والحياتية [21].

3-1 الخوارزمية الجينية مع أنظمة الكشف عن التطفل

إن النجاح والفهم الأفضل للخوارزمية الجينية أدّيا إلى تطبيقها في مجالات متنوعة من العلوم و الهندسة و الصناعة. حيث استخدمت بنجاح في المسائل المعقدة (Complex Problems) والمسائل غير المحلولة مسبقاً. أما في مجال الكشف عن التطفل فتعد الخوارزميات الجينية خوارزميات جيدة لاكتساب حلول مثلى, ويعد استخدامها

لتحديد مجموعة قواعد تطفلات الشبكة الفعلية والكامنة ذا قيمة وفائدة كبيرة [6] . والخوارزمية الجينية نوع حاسوبي متطور من خوارزميات الذكاء الاصطناعي الأفضل للامثلية, ومسائل اختيار الميزات (Features Selection), والعنقدة (Clustering). والخوارزمية الجينية تستخدم أيضاً في التصنيف (Classification), واستخدام نظام مصنف جيني لتحديد فعل معين يتخذ في حالة أي اختراق أمني. ويمكن استخدام الخوارزمية الجينية لاختيار مجموعة فرعية من خواص الإدخال لمصنفات شجرة القرار, بهدف زيادة نسبة الكشف وتقليل نسبة الإنذارات الكاذبة في شبكة الكشف عن التطفل [2]. كما ازداد استخدام الخوارزميات الجينية في كشف البرامجيات العدائية (Malicious Software) مثل الفايروسات, والديدان (Worms), والأحصنة (Trojans).

4. تحليل المركبات الأساسية Principal Component Analysis (PCA)

تُعد PCA تقنية إحصائية مفيدة لها تطبيقات في مجالات متعددة مثل تمييز الأنماط (التعرف على الوجه بشكل خاص), والتعبير الجيني, وتجميع البيانات, وإحداث تدفق المرور في الكشف عن التطفل, وكبس الصور ..الخ.

وتعتمد هذه التقنية الإحصائية على فكرة وجود مجموعة كبيرة من البيانات وترغب في تحليل هذه المجموعة حسبما يتعلق بالعلاقات بين النقاط المفردة في تلك المجموعة.

ويهدف PCA إلى تقليل أبعاد البيانات والاحتفاظ قدر الإمكان بالاختلاف الموجود في مجموعة البيانات الأصلية وهو طريقة لتحديد الأنماط في البيانات والتعبير عن البيانات بطريقة لإيضاح تشابهها و اختلافها. وهي تقنية شائعة تحوّل عدداً من الميزات المترابطة (Correlated Features) إلى ميزات غير مترابطة (Uncorrelated Features) والتي تدعى المركبات الأساسية (Principal Components) [3,23] .

يتطلب تطبيق تحليل المركبات الأساسية حساب مصفوفة التباين / التباين المشترك (Variance/Covariance matrix) للميزات, وبعدها يتم حساب المركبات الأساسية بطريقة جاكوبي (Jacobi Method).

يتطلب تحليل المركبات الأساسية رياضيا إيجاد مصفوفة التباين / التباين المشترك للميزات المتوفرة. ويتم حساب مصفوفة التباين / التباين المشترك على وفق المعادلة (1):

$$Cov_{ij} = \frac{1}{MN}\sum_{k=1}^{M}\sum_{l=1}^{N}\left(X_i(k,l) - M_i\right)\left(X_j(k,l) - M_j\right) \quad \ldots(1)$$

حيث M,N عدد السجلات الموجودة في نماذج التدريب وعدد الميزات الموجودة في كل سجل على الترتيب و j,i تمثل موقع الميزة لسجل وموقع السجل في النموذج على الترتيب و M_i و M_j هي المتوسط الحسابي للميزات i و j ويمكن حسابها على وفق المعادلة (2):

$$M_i = \frac{1}{MN}\sum_{k=1}^{M}\sum_{l=1}^{N} X_i(k,l) \quad \ldots(2)$$

نلاحظ من المعادلة إن قيم المصفوفة تكون تباينا (Variance) لعناصر القطر الرئيسي والتباين المشترك (Covariance) لبقية العناصر, كذلك إن مصفوفة (Covariance) مصفوفة متناظرة أي إن العناصر فوق القطر الرئيسي تكون مساوية بالقيم للعناصر تحت القطر المناظر لها .

وهكذا فان مصفوفة التباين / التباين المشترك ل n عدد من الميزات هي مصفوفة nxn ويمكن ترتيبها على النحو الآتي:

$$A=\begin{bmatrix} a_{11} & a_{12} & . & . & a_{1n} \\ a_{21} & a_{22} & . & . & a_{2n} \\ . & . & . & . & . \\ . & . & . & . & . \\ a_{n1} & a_{n2} & . & . & a_{nn} \end{bmatrix} \quad \ldots(3)$$

وحيث إن $a_{ij}=a_{ji}$ لكل $i \neq j$ فان المصفوفة A متناظرة .

مصفوفة متجه الايكن T تشتق من مصفوفة A باستخدام طريقة جاكوبي الموضحة في الفقرة اللاحقة والتي نحصل عليها من D , وتمثل بالمعادلة (4) :

$$D=\begin{bmatrix} \lambda_{11} & 0 & . & . & 0 \\ 0 & \lambda_{22} & . & . & 0 \\ . & . & . & . & . \\ . & . & . & . & . \\ 0 & 0 & . & . & \lambda_{nn} \end{bmatrix} \quad \ldots(4)$$

إن العناصر القطرية D تسمى قيمة أيكن (Eigen Value) لمصفوفة التباين / التباين المشترك حيث λ_{ii} لكل $i=1,2,\ldots\ldots,n$ هي التباينات لمحاور المركبات الأساسية إن العناصر خارج القطر (Off-Diagonal) للمصفوفة D هي صفر (أو قريبة من الصفر) مما يدل على إن المركبات الخارجة غير مترابطة أي إنها مستقلة . [11]

4-1 طريقة جاكوبي Jacobi s Method

يتم إيجاد اكبر عنصر في مصفوفة معينة مربعة ولتكن A بحيث لا يكون من عناصر القطر الرئيسي أي أنّ $Max\_element=a_{ik}$ و i, k هي الصف والعمود في مصفوفة A و $i \neq k$.

1. يتم إيجاد الزاوية θ وذلك عن طريق ما يأتي:

$$\theta=\frac{1}{2}\arctan\left(2a_{ik}/\left(a_{ii}-a_{kk}\right)\right) \qquad If \qquad a_{ii}\neq a_{kk} \quad \ldots(5)$$

$$\theta=\begin{cases} \pi/4 & when \quad a_{ik}>0 \\ -\pi/4 & when \quad a_{ik}<0 \end{cases} \qquad If \qquad a_{ii}=a_{kk} \quad \ldots(6)$$

2. إجراء عملية التدوير (Rotation) على المصفوفة C وإرجاع الناتج في مصفوفة وكما يأتي:

$$d_{ii}=\frac{1}{2}\left(a_{ii}+a_{kk}+\sigma R\right) \quad \ldots(7)$$

$$d_{kk}=\frac{1}{2}\left(a_{ii}+a_{kk}-\sigma R\right) \quad \ldots(8)$$

$$d_{ik}=d_{ki}=0 \quad \ldots(9)$$

حيث يتم إيجاد قيمة R على وفق المعادلة الآتية:

$$R=\sqrt{\left(a_{ii}-a_{kk}\right)^2+4a_{ik}^2} \quad \ldots(10)$$

ويتم إيجاد قيمة σ على وفق المعادلة الآتية:

$$\sigma = \begin{Bmatrix} 1 & if & a_{ii} \geq a_{kk} \\ -1 & if & a_{ii} < a_{kk} \end{Bmatrix} \quad \ldots(11)$$

أما بقية عناصر مصفوفة التدوير فنجدها باستخدام المعادلتين الآتيتين:

$$d_{ir} = a_{ir} \cos\theta + a_{kr} \sin\theta \quad \ldots(12)$$

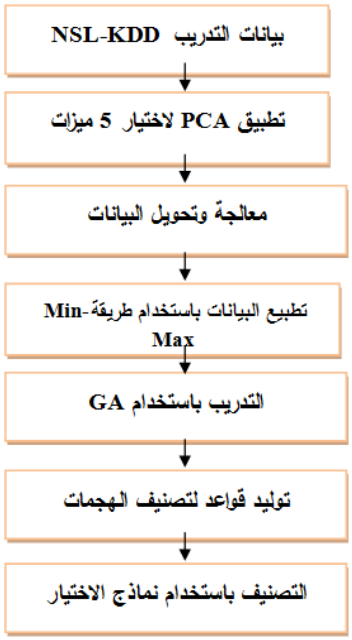
$$d_{kr} = -a_{ir} \sin\theta + a_{kr} \cos\theta \quad \ldots(13)$$

حيث $r \neq i$ و $r \neq k$ و i,k تمثل قيمة الصف والعمود داخل المصفوفة.

3. إعادة الخطوات من 1 إلى 3 على المصفوفة الناتجة إلى أن يتم الحصول على قيم عناصر خارج القطر قريبة من الصفر [11] .

5. النظام المقترح

الشكل (1) يوضح معمارية النظام المقترح.



الشكل (1). معمارية النظام المقترح

5-1 لغرض تصنيف الهجمات في بيانات التدريب تم اعتماد الخطوات الآتية في هذه المرحلة.

أولاً: إدخال نماذج التدريب

حيث كل نموذج تدريب يتكون من (41) ميزة و يحوي اسم الهجوم في نهايته **والشكل (1) يمثل نموذج من نماذج بيانات التدريب.**

```
0,tcp,ftp_data,SF,491,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,2,0.00,0.00,0.00,0.00,1.00,0.00,0.00,15
0,25,0.17,0.03,0.17,0.00,0.00,0.00,0.05,0.00,normal,20
0,tcp,private,REJ,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,1,0.00,0.00,1.00,1.00,0.50,1.00,0.00,255,
1,0.00,0.31,0.28,0.00,0.00,0.00,0.29,1.00,portsweep,20
0,tcp,private,S0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,123,6,1.00,1.00,0.00,0.00,0.05,0.07,0.00,255,
26,0.10,0.05,0.00,0.00,1.00,1.00,0.00,0.00,neptune,19
0,icmp,eco_i,SF,8,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,12,0.00,0.00,0.00,0.00,1.00,0.00,1.00,1,53,
1.00,0.00,1.00,0.51,0.00,0.00,0.00,0.00,ipsweep,17
0,tcp,http,SF,199,420,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,30,32,0.00,0.00,0.00,0.00,1.00,0.00,0.09,2
55,255,1.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,normal,21
```

الشكل (2). نموذج من نماذج بيانات

ثانياً: إدخال معاملات النظام

هنالك مجموعة من المدخلات, هي:

ثالثاً: اختيار الميزات

dst_host_count , Count , dst_bytes , src_bytes , Duration

رابعاً: المعالجة الأولية لبيانات نماذج التدريب

تم إجراء المعالجة الأولية لبيانات نماذج التدريب بتحويل حقل صنف Attack Type الذي يحوي متغيراً حرفياً يمثل أحد هذه الأصناف (R2L, U2R, Normal, Probe, DoS) بأرقام ضمن المدى [1..5], حيث يعطى 1 للصنف Normal, و2 للصنف DoS, و3 للصنف Probe, و4 للصنف R2L, و5 للصنف U2R.

تم تطبيع البيانات بطريقة Min-Max وتقوم هذه الطريقة بإجراء عمليات تحويل خطية على قيم البيانات الأصلية, ثم تطبق المعادلة الخطية على كل قيم الميزة X للحصول على القيمة الجديدة, المعادلة الآتية تستعمل لتطبيع بيانات التدريب والاختبار:

$$X_n = (X - MinX) / (MaxX - MinX) \quad \ldots(14)$$

.بعد انتهاء المعالجة الأولية لبيانات نماذج التدريب يتم إجراء مرحلة التدريب باتباع الخطوات الآتية:

خامساً: إدخال المجتمع الابتدائي

سادساً: حساب دالة التقييم لكل كروموسوم في المجتمع الابتدائي

يتم ذلك بمقارنة كل كروموسوم بسجلات نماذج التدريب, إذا وجد تطابق بينهما يزداد سجل التطابق بمقدار واحد, وحسب المعادلة الآتية:

$$F = WT1 * \log(NA) + WT2 \quad \ldots(15)$$

حيث (NA) هي عدد تطابق الكروموسومات مع سجلات بيانات التدريب, WT2, WT1 قيم ثابتة للدالة. إن قيم دالة التقييم تتراوح بين [0.0 – 1.0].

سابعاً: الاحتفاظ بأفضل الآباء

ثامناً: عملية تكوين الأجيال الجديدة

- استخدام عجلة الروليت لاختيار أفضل الآباء الذين لديهم صلاحية عالية باحتمالية كبرى.
- تتم الطفرة بنسبة (2%), وتستخدم طريقة التبادل بين قيم الكروموسوم باختيار عشوائي للكروموسوم واختيار عشوائي لاثنين من قيم الكروموسوم حيث يتم التبادل بينهما.
- حساب دالة التقييم للكروموسومات الجديدة وترتيبها تصاعدياً بعد إجراء عمليتي التزاوج والطفرة.

تاسعاً: اختبار شرط التوقف وتتكرر عملية تكوين الأجيال الجديدة بالرجوع إلى الخطوة ثامناً إلى أن يتحقق شرط التوقف.

5-2 مرحلة الاختبار Testing Stage

بعد انتهاء مرحلة التدريب والحصول على قائمة قواعد التصنيف المكتشفة تبدأ مرحلة تصنيف نماذج الاختبار بإتباع ما يأتي:

أولا: تبدأ عملية التصنيف بإدخال مجموعة نماذج الاختبار (NSL-KDD).

ثانيا: إدخال قواعد التصنيف المكتشفة وعددها إلى النظام.

ثالثا: اختيار (5) ميزات

رابعاً: المعالجة الأولية لبيانات نماذج الاختبار, تم إجراء هذه المعالجة من تحويل وتطبيع للبيانات, حيث يرمز لحقل Attack Type بالأرقام ضمن المدى [1..5], ويتم تطبيع البيانات باستخدام طريقة Min-Max Normalization.

خامساً: في هذه المرحلة يتم اختبار كل نماذج الاختبار المدخلة إلى النظام, حيث إن كل نموذج اختبار يقارن بقواعد التصنيف, وأول قاعدة تصنيف تتطابق مع نموذج الاختبار سوف يحدد صنف تلك القاعدة,

5-3 خطوات تنفيذ خوارزمية تحليل المركبات الأساسية

أولا: قراءة مجموعة نماذج التدريب NSL-KDD.

ثانيا: معالجة بيانات التدريب, حيث يتم تحويل الميزات الحرفية إلى عددية فالميزات (Protocol type ,Service, Flag) يتم تحويلها من [1 ..عدد القيم ضمن الميزة], وكذلك تحويل حقل Attack Type من (Normal وAnomaly) إلى (1,0) على التوالي.

ثالثا : حساب مصفوفة التباين / التباين المشترك Variance/ Covariance Matrix للميزات الموجودة في كل سجل من نماذج التدريب.

رابعا: إيجاد أكبر عنصر في المصفوفة.

- إيجاد زاوية التدوير.
- إيجاد عناصر مصفوفة التدوير.
- إعادة الخطوات من 1 إلى 3 على المصفوفة الناتجة من الخطوة السابقة حتى يتم الحصول على عناصر خارج القطر قريبة من الصفر.

خامسا: حساب قيمة متجه الايكن من المصفوفة الناتجة ووضعها في مصفوفة الايكن.

سادسا: ترتيب مصفوفة الايكن.

6. النتائج

أجريت التجارب على الخوارزمية الجينية لكشف وتصنيف التطفل باستخدام (5 ميزات) وتم الحصول على مجموعة قواعد التصنيف (276 قاعدة) , والجدول الآتي يوضح ذلك.

الجدول (3) أعداد قواعد التصنيف المكتشفة في مرحلة تدريب النظام

Class	DoS	Probe	R2L	U2R	Normal
No. of Rules	68	49	19	5	135

لقد تم اختبار النظام باستخدام قواعد التصنيف التي تم الحصول عليها من مرحلة تدريب النظام و على النحو الآتي:

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

تمت عملية اختبار النظام عن طريق استخدام مجموعة نماذج التدريب, واستخدمت البيانات نفسها المستخدمة في مرحلة تدريب النظام. ونتائج الاختبار موضحة بالجدول (4) .

الجدول (4) نتائج التصنيف بتطبيق الخوارزمية الجينية على بيانات التدريب

Class	No. of Rules	No. of Detected Rules	Detection Rate	Training Time
DoS	68	44608	97. 06 %	
U2R	5	20	86. 95 %	

Normal	135	63946	94. 95%	H:M:S
Probe	49	9327	80. 06 %	0: 20: 5
R2L	19	794	82. 7 %	

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

لقد اختبر النظام عن طريق استخدام مجموعة نماذج الاختبار ونتائج الاختبار موضحة بالجدول (5).

الجدول (5). نتائج التصنيف بتطبيق الخوارزمية الجينية على بيانات الاختبار

Class	No. of Rules	No. of Detected Rules	Detection Rate	Testing Time M:S
DoS	68	6973	93. 48 %	0: 3
U2R	5	23	32. 84%	
Normal	135	9655	99. 52 %	
Probe	49	1976	81. 61 %	
R2L	19	2007	69. 47 %	

7. المناقشة

بعد أن أجريت عملية الاختبار والحصول على النتائج, تم التركيز على نسبة الكشف, ونسبة الإنذار الكاذب و حجم البيانات المستخدمة بالاختبار في الخوارزمية الجينية لكشف التطفل .

استخدمت تقنيات تقليل الأبعاد تحليلاً للمركبات الأساسية مع بيانات NSL-KDD لتقليل الميزات, واعتمدت على طريقة جاكوبي في إيجاد قيمة الايكن, وأظهرت النتائج انجازاً جيداً لتحليل المركبات الأساسية لاختياره الميزات التي تحمل أعلى قيم في متجه الايكن. تم اختيار (5) ميزات, وبذلك تم تقليل تعقيد العمليات الحسابية. ركزت الخوارزمية الجينية لكشف وتصنيف التطفل على نسبة التصنيف مع حجم البيانات المستخدمة في الاختبار حيث أعطت نسبة كشف تطفل (91.6%) ونسبة الإنذار الكاذب صفراً. بالنسبة لهجومي DoS و Probe كانت نسبة كشف التطفل (81.61%), و(93.48%) على التوالي, وأعطت (R2L) نسبة كشف (69.47%) وهي نسبة جيدة مقارنة بنتائج الباحثين السابقين كما أن عدداً من بيانات سجلات الاتصال التي تعود إلى الصنف (R2L) تشابه البيانات في سجلات الاتصال التابعة إلى الصنف (Normal). أما (U2R) فنسبة كشف التطفل له (32.84%) وسبب ذلك يعود إلى أن عدة أنواع من الهجمات متوفرة في مرحلة التدريب ومفقودة في نماذج الاختبار, فضلاً عن وجود أنواع جديدة من الهجمات في مرحلة الاختبار ,وهذا يفسر نسبة الكشف القليلة له.

قورنت النتائج التي تم الحصول عليها من التجارب مع نتائج عدد من الباحثين الذين يعملون في المجال نفسه إذ إن الجدول (6) يبين أنّ تطبيق الخوارزمية الجينية وخوارزمية تحليل المركبات الأساسية باستخدام (5 ميزات) أعطت نتائج جيدة في كشف وتصنيف التطفل رغم تطبيقها على مجموعة كبيرة من البيانات.

الجدول (6). مقارنة نتائج الخوارزمية الجينية مع عدد من الباحثين

Model	Data Base	DoS %	Probe %	R2L %	U2R %	Normal %	DR %
KDD Winner 2010 [19]	KDD	97. 1	83. 3	8. 4	13. 2	99. 5	
GA – NN [17]	KDD	86. 7	86. 1	81. 2	79. 2	-	
SOM [22]	NSL-KDD	-	-	-	–	-	64
K-Means [22]	NSL-KDD	-	-	-	-	-	60

GA-RBF, 2010 [17]	KDD	86. 7	86. 1	81. 2	79. 2	-	
PCA- C4. 5 2011 [18]	KDD	97. 25	66. 30	2. 30	8. 33	98. 99	
PCA nearest neighbor (NN-rule) 2011 [18]	KDD	97. 14	74. 40	0. 80	7. 91	99, 50	
Ant-Miner 2011 [10]	KDD	96. 15	72. 90	13. 88	97. 13	94. 50	92. 42
PCA-GA Hana Mohammed	NSL-KDD	93. 48	81. 61	69. 47	32. 84	99. 52	91. 6

8. الاستنتاجات

من خلال تصميم النظام المقترح وتطبيقه على بيانات NSL-KDD , وبعد إجراء التجارب المختلفة لقياس كفاءة النظام وأداءه, تم استنتاج الآتي:

1. تم اختبار بيانات NSL-KDD التي حلت مشاكل KDD 99 , وأوضحت التجارب أن هذه البيانات يمكن وضعها لمساعدة الباحثين لمقارنة مختلف نماذج كشف التطفل.
2. أوضحت التجارب ان النظام المقترح قادر على تسريع عمليات التدريب والاختبار لكشف وتصنيف التطفل والتي تعمل على زيادة سرعة تطبيقات الشبكة, كما قلل النظام المقترح وقت التدريب والاختبار.
3. النظام المقترح المعتمد على الخوارزمية الجينية مع PCA أسرع بالتدريب والاختبار من الخوارزمية الجينية بدون PCA . حيث زادت سرعة التدريب 80% باستخدام PCA.
4. الطفرة لم تلعب دورا واضحاً في نتائج البحث.
5. التكرارات الأولية للأجيال يتم الوصول فيها إلى النتائج بسرعة كبيرة مقارنة بالتكرارات المتأخرة.

9. التوصيات والأعمال المستقبلية

1. اعتماد قاعدة البيانات NSL-KDD, بالرغم من إنها لم تعط حلولاً لجميع المشاكل السابقة بصورة كاملة إلا أنها أثبتت فائدتها البحثية في بناء نماذج كشف التطفل على بيئة المحاكاة.
2. استخدام المنطق المضبب أو العصبي المضبب Neuro-Fuzzy مع الخوارزمية الجينية لتحسين النظام.
3. استخدام دوال تقييم مختلفة لأنها تؤدي دوراً مهماً في كشف التطفل.
4. استخدام طرق أخرى لتطبيع البيانات.
5. اختيار ميزات مناسبة لكل نوع من هجمات الشبكة.
6. تطوير النظام لكي يعمل على البيئة الحقيقية (On-line) .
7. استخدام تقنيات أخرى في استخلاص الميزات واختيار الميزات مثل الخوارزمية الجينية , LDA, وIDA.

المصادر

[1] Al Shilany, Ismail Ali, 2010, "Design and Implementation of Artificial Immune System for Detecting SYN-Flood Attack", MSc. Thesis, Computer Science College, University of Mosul, Iraq.

[2] Anand Takale, 2004, "Constructing Predictive Models to Assess the Importance of Variables in Epidemiological Data Using a Genetic Algorithm System Employing Decision Trees", MSc. Thesis, University of Minnesota.

[3] Ansam O. Abdul-Majeed, 2011, " New Steganographic Method for VQ-Compressed Images", MSc. Thesis, Compute Science College, University of Mosul, Iraq.

[4] Christina Lee, 2007 ," An Evaluation of Machine Learning Techniques in Intrusion Detection", MSc. Thesis, Computer Science of the Graduate School of Vanderbilt University.

[5] Chittur A., 2001, "Model Generation for an Intrusion Detection System Using Genetic Algorithms", Ossining High School, NY.

http://www1.cs.columbia.edu/ids/publications/gaids-Thesis1.pdf.

[6] Ciza Thomas, 2009, " Performance Enhancement of Intrusion Detection Systems Using Advances in Sensor Fusion", Ph.D. Thesis, Faculty of Engineering Indian Institute of Science.

[7] G.MeeraGandhi , Kumaravel Appavoo and S.K. Srivatsa , 2010 ,"Effective Network Intrusion Detection Using Classifiers Decision Trees and Decision rules", Int. J. Advanced Networking and Applications Volume: 02, Issue: 03, Pages: 686-692.

[8] KDDCup 1999 Dataset. Available at:

http:// kdd.ics.uci.edu/databases/kddcup1999.html.

[9] M. Tavallaee, E. Bagheri, W. Lu, and A. Ghorbani, 2009, "A Detailed Analysis of the KDD CUP 99 Data Set", Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA).

[10] Mahmood S. Mahmood, 2011, "Using Ant and Self-Organization Maps Algorithms To Detect and Classify Intrusion In Computer Networks", MSc. Thesis, Computer Science College, University of Mosul, Iraq.

[11] Muna J. Alshamdeen, 2011, " The Best Band Selection Using Hybrid Techniques Applied on Remote Sensing Data", MSc. Thesis, Computer Science College, University of Mosul ,Iraq.

[12] Murad Abdo Rassam, 2010, "Anomaly Intrusion Detection System Using Immune Network with Reduced Network Traffic Features, Msc. Thesis, University Technology Malaysia.

[13] Safaa Zaman, 2009, "A Collaborative Architecture for Distributed Intrusion Detection System based on Lightweight Modules" Ph.D. Thesis in Electrical and Computer Engineering, University Waterloo.

[14] Sandeep V. Sabnani, 2008, " A Machine Learning Approach", MSc. Thesis, Information Security at Royal Holloway, University of London.

[15] Shaheen A., 2010, "A comparative Analysis of Intelligent Techniques for detecting Anomalous Internet Traffic", MSc. Thesis, King Fahad University, Saudi Arabia.

[16] Shilpa lakhina, Sini Joseph and Bhupendra Verma, 2010, "Feature Reduction Using Principal Component Analysis for Effective Anomaly–Based Intrusion Detection on NSL-KDD", International Journal of Engineering Science and Technology, Vol. 2(6), 1790-1799.

[17] S. Selvakani and R.S.Rajesh, 2009, "Escalate Intrusion Detection Using GA – NN", Int. J. Open Problems Compt. Math., Vol. 2, No. 2.

[18] S. Selvakani Kandeeban and Rengan S. Rajesh, 2010, " Integrated Intrusion Detection System Using Soft Computing, International Journal of Network Security, Vol.10, No. 2, PP. 87.

[19] Syed Muhammad Aqil Burney, M. Sadiq Ali Khan and Dr.Tahseen A. Jilani, 2010. " Feature Deduction and Ensemble Design of Parallel Neural Networks for Intrusion Detection System, (IJCSE) International Journal of Computer Science and Network Security ,Vol.10 No.10.

[20] Philip p. Winter, 2010, " Inductive Intrusion Detection in Flow-Based Network Data using One-Class Support Vector Machines", MSc. Thesis, Information System, Hagenberg.

[21] Pohlheim, H. , 2006, " Genetic and Evolutionary Algorithms: Principles, Methods and Algorithms. Genetic and Evolutionary Algorithm". http://www.geatbx.com/docu/algindex.html.

[22] Ritu Ranjani Singh, Prof. Neetesh Gupta, 2010, " To Reduce the False Alarm in Intrusion Detection System using self Organizing Map", International journal of Computer Science and its Applications.

[23] T. Jolliffe, 2002, "Principal Component Analysis", Springer-Verlag, (New York), ISBN 0-387-95442-2.

[24] Zorana Bankovic, Dus an Stepanovic, Slobodan Bojanic and Octavio Nieto-Taladriz, 2007, "Improving Network Security Using Genetic Algorithm Approach", Published by Elsevier Ltd. i:10.1016/ j.compeleceng.