استخدام تقنيات تنقيب البيانات لكشف التطفل في شبكات الحاسوب

أ.د.السماني عبد المطلب أحمد -باحث ثاني
أ.إياد محمد مهيوب غالب البريهي -باحث أول
جامعة النيلين، كلية علوم الحاسوب وتقانة المعلومات، الخرطوم، جمهورية السودان

2

الملخص

أمان وسرية المعلومات في الشبكات تعتبر القضية الرئيسية المؤرقة لكثير من الشركات والمؤسسات التي تستخدم كمية كبيرة من البيانات، وبالمقابل هناك العديد من الطرق المستخدمة لحماية الشبكات في الوقت الراهن مثل التشفير والـ VPN، وجدران الحماية، ولكن كل هذه الطرق تعتبر طرق استاتيكية جداً للحماية الفعالة ضد المهددات وعدد المهددات، وتستخدم تقنيات تنقيب البيانات لهذا الغرض حيث تطبق لكشف التطفل intrusion detection.

تهدف هذه الورقة لإستخدام تقنيات تنقيب البيانات (data mining) لكشف حالات الشذوذ في الشبكات تطبيقاً على عينتين عشوائيتين من مجموعة بيانات NLS-KDD Data Set، استخدمت الورقة تقنية التصنيف مصنف شجرة القرار (decision tree) التي تنفذ خوارزمية C4.5، نتيجة التجربة تعرض أن مصنف C4.5 أظهر نتائج فعالة لكشف التطفل في الشبكات الحاسوبية وأظهرت النتائج أنه كلما كانت كمية البيانات كبيرة تكون نسبة الخطأ أقل ودقة التنبؤ عالية.

الكلمات المفتاحية : تنقيب البيانات، التطفل، التصنيف، شجرة القرار، اكتساب المعلومات، Gain Ratio، gini index.

١- مقـــدمة

أمن أنظمة الحواسيب الخاصة بنا والبيانات في خطر مستمر، ولقد أدى النمو الواسع النطاق للإنترنت وازدياد توفر أدوات وحيل التطفل ومهاجمة الشبكات للتوجه لكشف ومنع التطفل، مما أصبح عنصراً اساسياً وفعالاً في أنظمة الشبكات.

يمكن تعريف التطفل (Intrusion) بانه مجموعة من الأحداث والمهددات التي تهدد السرية والسلامة أوتوافر موارد الشبكة كحساب المستخدم، أنظمة الملفات، نواة النظام ... الخ[١].

يتضمن كشف التطفل تحديد مجموعة من الأحداث الخبيثة التي تهدد سلامة وسرية، وتوافر موارد المعلومات. تستند الطرق التقليدية لكشف التسلل على المعرفة الواسعة لتوقيعات الهجمات المعروفة.

تقنيات كشف التسلل المعتمدة على تنقيب البيانات تستند عموماً على واحدة من فئتين، الكشف عن سوء الاستخدام misuse detection أو الكشف عن الشذوذ anomaly detection. في misuse detection يسمى كل instance في مجموعة البيانات إما 'normal ' أو ' intrusion ' و يتم تدريب خوارزمية التعلم على بيانات التدريب، وهذه التقنيات قادرة على إعادة تدريب نماذج كشف التسلل تلقائياً على بيانات الإدخال المختلفة التي تشمل أنواع جديدة من الهجمات ما دام قد وصفت بشكل مناسب.

الكشف عن الشذوذ anomaly detection ، من ناحية أخرى ، يبني نماذج للسلوك الطبيعي normal behavior، وبشكل تلقائي يتم الكشف عن أي انحراف عن ذلك السلوك.

انظمة كشف التطفل او أنظمة منع التطفل كليهما تعمل على مراقبة حركة مرور الشبكة ومنع الأنشطة الخبيثة، وغالبية أنظمة كشف ومنع التطفل تستخدم إما الكشف المعتمد على التوقيع signature-based detection أو الكشف المعتمد على الشذوذ anomaly-based detection.

أ٠ الكشف المعتمد على التوقيع signature-based detection: هذه الطريقة تستخدم للكشف عن التوقيعات التي هي أنماط هجوم مكونة مسبقاً ومحددة سلفاً من قبل خبراء المجال، ونظام منع الاختراق المستندة إلى التوقيع تراقب حركة مرور الشبكة لمطابقتها مع هذه التوقيعات، واذا تم العثور مرة واحدة على تطابق فإن نظام كشف التسلل سوف يبلغ عن وجود الشذوذ وسيتخذ نظام منع الاختراق إجراءات إضافية مناسبة.

ب٠ الكشف المعتمد على الشذوذ Anomaly-based detection: : هذه الطريقة تبني نماذج لسلوك الشبكات الطبيعية تسمى (تشكيلات profiles) تستخدم بعد ذلك للكشف عن الأنماط التي تحيد كثيراً عن هذه التشكيلات، ومثل هذه الانحرافات قد تمثل الإختراقات الفعلية أو ببساطة السلوكيات الجديدة التي تحتاج أن تضاف إلى هذه التشكيلات. الميزة الرئيسية للكشف المعتمد على الشذوذ قدرته على

كشف الاختراقات غير المألوفة التي لم يتم ملاحظتها، وعادة يجب على المحلل فرز الانحرافات للتأكد أيا منها تمثل الاختراقات الحقيقية.

ستتناول هذه الورقة استخدام تقنية التصنيف كتقنية من تقنيات تنقيب البيانات لتصنيف بيانات الشبكة لعينتين عشوائيتين من مجموعة البيانات NLS- KDD99 data set العينة الأولى تحتوي على (5,290,866 byte) كبيانات تدريب و(946,848 byte) كبيانات اختبار وتحتوي العينة العشوائية الثانية على كمية بيانات (1,058064 byte) كبيانات تدريب و (497,700 byte) كبيانات اختبار يتم تدريب الخوارزمية C4.5 بواسطة أداة الويكا Weka Tools المصممة لتدريب واختبار خوارزميات تنقيب البيانات، يتم تقييم مصنف C4.5 بواسطة حساب دقة المصنف ونسبة الخطأ الناتجة من مصفوفة التضارب Confusion Matrix.

٢ -استخدام تنقيب البيانات في كشف التطفل :

لتعزيز أمن الشبكة هناك العديد من الخوارزميات المستخدمة في كشف التسلل مثل التشفير، و VPN، وجدار الحماية، ...الخ، وهذه الخوارزميات توفر الحماية بشكل ثابت عبر الشبكة، ولكن كشف التسلل عملية ديناميكية ولذلك يجب توفير حماية ديناميكية عبر الشبكة[٤]. يمكن أن تساعد تقنيات تنقيب البيانات في تعزيز أداء أنظمة حماية ومنع التطفل بعدة طرق أهمها :

٢ -١ خوارزميات تنقيب بيانات جديدة لكشف التطفل :

يمكن استخدام خوارزميات تنقيب البيانات للكشف المعتمد عن التوقيع والكشف المعتمد عن الشذوذ معاً.

٢-١-١ في خوارزميات الكشف المعتمد على التوقيع تصنف بيانات التدريب إما "طبيعية normal " او "متطفلة intrusion"، ويمكن بعد ذلك اشتقاق المصنف الذي يكتشف الاختراقات المعروفة في هذا المجال.

٢-١-٢ في خوارزميات الكشف المعتمد على الشذوذ يتم بناء نماذج السلوك الطبيعي normal (behavior) للبيانات وبشكل تلقائي يتم الكشف عن الانحرافات الهامة منها.

٢-٢ تحليل الربط Association، والارتباط correlation ونمط التميز

تساعد هذه الخوارزميات في اختيار وبناء المصنفات التمييزية وتطبق لإيجاد العلاقة بين خصائص النظام system attributes التي تصف بيانات الشبكة، ومثل هذه المعلومات يمكن أن توفر الرؤيا الواضحة فيما يتعلق باختيار سمات مفيدة لكشف التسلل.

٢-٣ تحليل البيانات المجدولة Analysis of stream data:

نظراً لطبيعة تنقل وحركة التطفلات والهجمات الخبيثة لابد من إجراء الكشف عن التطفلات في بيئة البيانات المجدولة، وبالرغم من أن الحدث قد يكون طبيعياً من تلقاء نفسه إلا إنه اعتبر ضاراً اذا نظر إليه كجزء من تسلسل الأحداث، وبالتالي فمن الضروري دراسة ماهي تسلسلات الاحداث المتصادفة معاً بشكل متكرر ولابد من إيجاد الأنماط المتسلسلة وتحديد القيم المتطرفة. طرق تعدين البيانات الأخرى ضرورية أيضاً لإيجاد التجمعات المتطورة وبناء نماذج تصنيف ديناميكية في جداول البيانات في الوقت الحقيقي لكشف التسلل.

٢-٤ طرق تنقيب البيانات الموزعة :

التطفلات يمكن إطلاقها من عدة مواقع مختلفة لتوجه إلى عدة وجهات مختلفة، وفي هذه الحالة يمكن استخدام طرق تنقيب البيانات الموزعة لتحليل بيانات الشبكة من عدة مواقع للشبكة للكشف عن التهديدات الموزعة.[٣]

٣ - انواع مهددات التطفل Type Intrusion Attacks:

ستتناول هذه التجربة في هذه الورقة الكشف عن أربعة مهددات تهاجم الشبكات الحاسوبية، حيث سيتم الكشف عنها بواسطة تقنية التصنيف كتقنية من تقنيات تنقيب البيانات بواسطة تنفيذ خوارزمية C4.5 على قاعدة بيانات NLS-KDD99 Data Set.

٣ -١ مهدد إنكار الخدمة Denial of Service Attack (DoS) :

هو المهدد الذي يجعل مورد الحوسبة او الذاكرة مشغولة جداً او ممتلئة للتعامل مع أي طلبات مشروعة أو حرمان المستخدمين الشرعيين من الوصول إلى الخدمة .

٣ -٢ مهدد المستخدم الرئيسي User to Root Attack (U2R):

هو المهدد الذي يتظاهر بأنه المستخدم الشرعي للنظام دون إذن، ثم يستغل نقاط ضعف النظام للحصول على صلاحية الوصول إلى مصدر التحكم بالنظام على سبيل المثال امتلاك صلاحيات تنزيل البرامج المخفية التي قد تتسبب في فقد النظام أو تنفيذ برامج هجومية كما لو أنها جزء من برامج النظام الأصلية.

٣ -٣ مهدد محلي عن بعد Remote To Local Attack (R2L):

هم المهددين غير المصرح لهم خلال الشبكات ويحصلون على وصول محلي للشبكة كمستخدمي الأجهزة المحلية، ويمكنهم شن الهجمات من أي مكان على الانترنت، وفي حالة ان مستخدماً ما أصبح لديه حق الوصول إلى نظام المعلومات فإنه يمكنه استغلال نقاط الضعف في الجهاز المحلي لسرقة البيانات الهامة أو تدمير نظام المعلومات. أي خدمة تحتاج لكلمة سر للوصول إليها تعتبر هدفاً لهذا

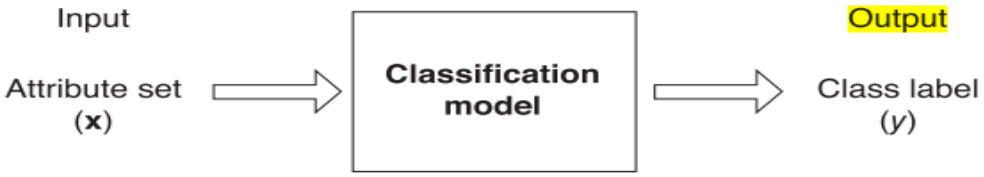
المهدد. هجمات R2L من أصعب الهجمات كشفاً لأنها تنطوي على مستوى الشبكة وميزات مستوى المضيف.

٣ -٤ مهدد التدقيق PROBE ATTACK:

يتم الهجوم من قبل المهاجمين باستخدام برامج فحص تلقائية لعدد كبير من عناوين الـ IP Address الخاص بالشبكات من أجل العثور على نقاط ضعف يمكن استغلالها، فإذا عثر على نقاط ضعف لمرة واحدة فإنه يمكن للمهاجمين استغلالها لكسب الوصول إلى الشبكة والبدء بجمع المعلومات بدون اذن.[٤]

٤ -تقنية التصنيف classification

التصنيف هو مهمة تعليم دالة معينة f لربط مجموعة الخصائص (x) بفئة (class) معرفة مسبقاً تسمى y ، تعرف الدالة f أيضاً بنموذج التصنيف classification model.

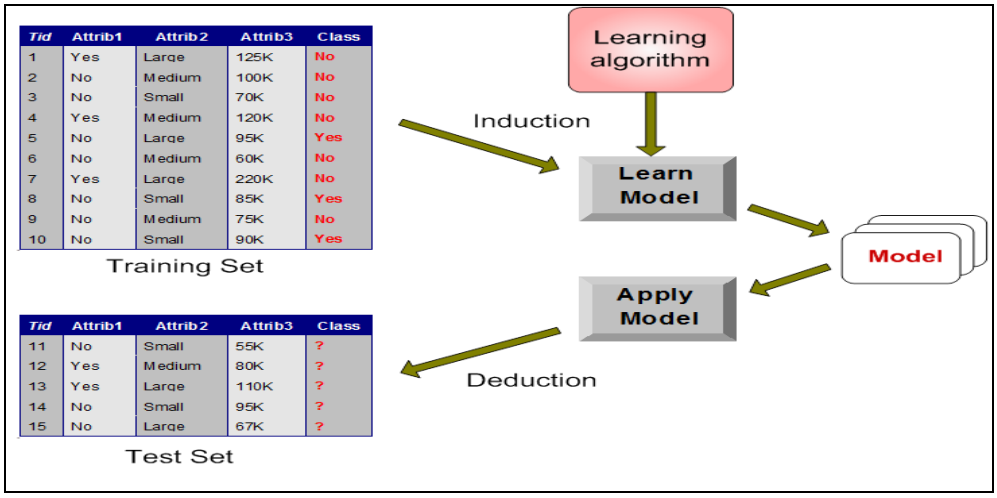


شكل (١) التصنيف

تتم عملية بناء المصنف بخطوتين رئيسيتين هما التعليم (خطوة بناء نموذج التصنيف) والتصنيف (استخدام النموذج ليتنبأ بفئات البيانات غير المعروفة).

أ. في الخطوة الأولى: يتم تدرب خوارزمية التصنيف (التعلم) classifing algorithm على بيانات التدريب training data المحتوية على سجلات معروفة لبناء المصنف الذي سيستخدم لفحص بيانات الفحص data test التي تحتوي على سجلات غير معروفة.

ب. الخطوة الثانية: يقيم أداء المصنف بحساب عدد السجلات المتوقعة المصنفة بشكل صحيح والسجلات المصنفة بشكل خطأ فيما يسمى بمصفوفة التعارض confusion matrix، يتم تقييم أداء نماذج المصنفات بالحصول على خوارزميات تصنيف تسعى للحصول على أعلى دقة و اقل نسبة خطأ عند تطبيقها على بيانات الاختبار data test .[٥]



شكل(٢) خطوات بناء مصنف البيانات

٤ -١ حساب مصفوفة التعارض confusion matrix

يتم جدولة عدد السجلات المصنفة بشكل صحيح وعدد السجلات المصنفة بشكل خطأ على شكل مصفوفة تعرف بمصفوفة التعارض.

		Predicted class	
		Class = 1	Class = 0
Actual class	Class = 1	TP	FN
	Class = 0	FP	TN

شكل(٣) مصفوفة التعارض الثنائية [١] binary confusion matrix

كل مدخل fij في مصفوفة التعارض يشير الى عدد السجلات في الـ class i المتوقع تكون في الـ class j. مثلا fo1 تشير إلى عدد السجلات في الـ class ٠ التي تم توقعها بشكل خطأ في الـ class ١.

اعتماداً على المصفوفة فإن :

مجموع السجلات المتوقعة بشكل صحيح هي (f00 + f11) .

مجموع السجلات المتوقعة بشكل خطأ هي (f01+f10) .

تحسب الدقة Accuracy لتقييم أداء المصنفات اعتماداً على مصفوفة التعارض من المعادلة التالية

$$Accuracy = \frac{Number\ of\ correct\ predictions}{\text{Total number of predictions}} = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

وبشكل مكافئ يحسب معدل الخطأ من المعادلة التالية :

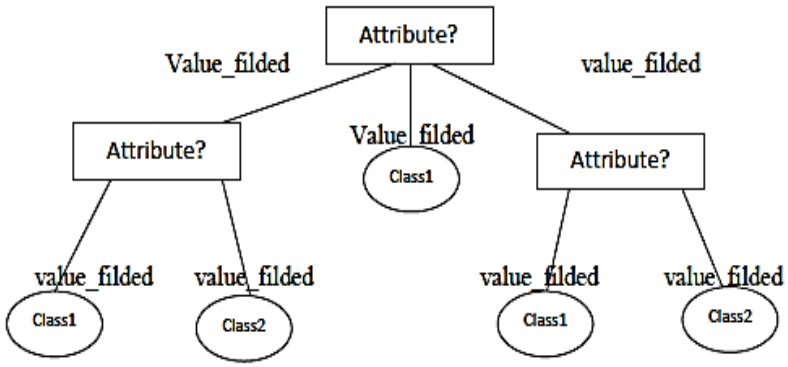
$$Arrore\ rate = \frac{Number\ of\ wrong\ predictions}{\text{Total number of predictions}} = \frac{f_{01} + f_{10}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

٤ -٢ تقنيات التصنيف Classification Techniques

يتضمن التصنيف عدة تقنيات رئيسية ستناقش هذه الورقة أهم هذه التقنيات وهي تقنية شجرة القرار Decision Tree.

٤ -٢ -١ مصنف شجرة القرار Decision Tree classifier

أشجار القرار تصنف الحالات عن طريق فرزها على أساس قيم الصفة attribute. كل عقدة داخلية في شجرة القرار تمثل صفة اختبار test attribute، كل فرع يمثل قيمة العقدة، وكل عقدة طرفية (ورقة) تمثل مخرجات الاختبار (قيمة value)، أعلى عقدة في الشجرة هي جذر الشجرة، وتصنف الحالات انطلاقاً من عقدة الجذر في شجرة القرار، يتم تعيين كل عقدة ورقية كفئة class.



الشكل(١ -١٧) شجرة قرار التصنيف[4]

الخوارزميات ID3، C4.5، CART تستخدم أسلوب فرق تسد من أعلى إلى أسفل (top-down recursive divide-and-conquer manner) لتوليد أشجار القرار التي تبدأ بالتدريب على مجموعة بيانات التدريب والفئات (class labels) المرتبطة بها.

٤ -٢ -١ -١ بناء شجرة القرار How to Build a Decision Tree

من حيث المبدأ، هناك عدد من أشجار القرار يمكن بناؤها من مجموعة معينة من الصفات المتوفرة، بينما كثير من أشجار القرار تكون أكثر دقة من الأخرى، الحصول على الشجرة الأمثل مناسب حسابياً بسبب الكم الهائل من حجم عمليات البحث التي تقوم بها الخوارزمية، ومع ذلك فقد تم تطوير خوارزميات فعالة للحث على الدقة بشكل معقول. واحدة من هذه الخوارزميات خوارزمية هانتHunt's algorithm، التي هي أساس العديد من خوارزميات حث شجرة القرارات القائمة، بما في ذلك ID3، C4.5، CART. يعرض هذا القسم مناقشة رفيعة المستوى لخوارزمية هانت.[٥]

٤ -٢ -١ -٢ خوارزمية هانت Hunt's Algorithm

في خوارزمية هانت يتم توليد شجرة القرار باتباع أسلوب التكرار الذاتي عن طريق تقسيم سجلات التدريب إلى مجموعات فرعية بشكل تتابعي.

ليكن Dt مجموعة سجلات التدريب المرتبطة بالعقدة N و y={ y1,y2,…,yn} عبارة عن أصناف التصنيف class lebeles:

١ - إذا كانت كل السجلات في Dt تنتمي إلى نفس الـ class yt فإن t عقدة ورقية مرتبطة بـ class yt .

٢ - اذا Dt تحتوي سجلات تنتمي إلى أكثر من class ، يتم اختيار شرط اختبار الصفة (attribute test condition) لتقسيم السجلات إلى مجموعات فرعية صغيرة .يتم توليد العقدة الابن (child node) لكل مخرج من مخرجات شرط الاختبار ويتم توزيع سجلات Dt لعقد الأبناء اعتماداً على المخرجات. يتم تطبيق الخوارزمية بشكل تكرار ذاتي لكل عقد الأبناء child node [٥]

٤ -٢ -١ -٣ معايير اختيار الصفة Attribute Selection Measures

لبناء شجرة القرارات، بجانب اختيار خوارزمية البناء يجب أن تؤخذ بعين الاعتبار معايير اختيار الصفة المناسب لاختيار معيار التقسيم الأفضل الذي يقسم بيانات معينة D، من المجموعات التدريبية المرتبطة بصف (class) إلى مجموعات فردية. ستناقش هذه الورقة ثلاثة معايير عامة لإختيار الصفة المناسبة للتقسيم الأفضل Information Gain، Gain Ratio، Gini Index.

٤ -٢ -١ -٣ -١ اكتساب المعلومة Information Gain

يعرف اكتساب المعلومة بالانخفاض المتوقع للمعلومات اللازمة لتصنيف الصفة A. الصفة التي تكون قيمة اكتساب المعلومة لها أعلى من قيمة الصفة الأخرى يتم إختيارها كصفة تقسيم للعقدة N في مجموعة البيانات D [١].

ولحساب اكتساب المعلومة لصفة منفصلة A في مجموعة البيانات D نحسب الأتي :

أ- **المعلومات المتوقعة اللازمة لتصنيف سجل في D أو ما يسمى Entropy(D) من المعادلة التالية :**

$$Info(D) = \sum_{i=1}^{m} p_i \log_2(p_i)_i \quad \ldots(1)$$

أو Entropy(D) = $\sum_{i=1}^{m} p_i \log_2(p_i)_i$

حيث p_i احتمالية انتماء السجل في D إلى الـ C_i class وتحسب من العلاقة الآتية :

$$p_i = \frac{|C_{i,D}|}{|D|}$$

حيث $C_{i,D}$ عدد السجلات في D المنتمية إلى الـ class C ، D : العدد الكلي للسجلات .

تستخدم دالة اللوغاريثم بالاساس ٢ لان المعلومات ترمز بالـ bits .

- Info(D) هي متوسط كمية المعلومات اللازمة لتحديد فئة السجل i في D.

ب - نحسب المعلومات المتوقعة للـ attribute A في D حيث A تحتوي على عدد v من القيم المنفصلة حيث v={a1,a2,...,av} من المعادلة التالية :

$$info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \text{info}(D_j) \ldots \quad \ldots\ldots\ldots (2)$$

أو

$$Entropy_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \text{Entropy}(D_j)$$

حيث : $\frac{|D_j|}{|D|}$ وزن الجزء j بالنسبة لقيم الصفة A، Dj عدد الـ tuples التي تنتمي للـ class j التابعة للـ attribute A في D.

ج- لحساب معادلة Information Gain(A) نحسب الفرق بين المعادلة (١) والمعادلة(٢)

$$\text{Gain(A)} = \text{info}(D) - info_A(D) \quad \ldots. (3)$$

أو

No.	Duration	dst_host_same _src_port_rate	dst_host_srv_ diff_host_rate	Flag	src_bytes	dst_bytes	Class
1	0.0	0.17	0.0	SF	491.0	0.0	Normal
2	0.0	0.88	0.0	SF	146.0	0.0	Normal
3	0.0	0.0	0.0	S0	0.0	0.0	Anomaly
4	0.0	0.03	0.04	SF	232.0	8153.0	Normal
5	0.0	0.0	0.0	SF	199.0	420.0	Normal
6	0.0	0.0	0.0	REJ	0.0	0.0	Anomaly
7	0.0	0.0	0.0	S0	0.0	0.0	Anomaly
8	0.0	0.0	0.0	S0	0.0	0.0	Anomaly
9	0.0	0.0	0.0	S0	0.0	0.0	Anomaly
10	0.0	0.0	0.0	S0	0.0	0.0	Anomaly
11	0.0	0.0	0.0	REJ	0.0	0.0	Anomaly
12	0.0	0.0	0.0	S0	0.0	0.0	Anomaly
13	0.0	0.12	0.03	SF	287.0	2251.0	Normal
14	0.0	1.0	0.2	SF	334.0	0.0	Anomaly
15	0.0	0.0	0.0	S0	0.0	0.0	Anomaly
16	0.0	0.0	0.0	S0	0.0	0.0	Anomaly
17	0.0	0.01	0.02	SF	300.0	13788.0	Normal
18	0.0	1.0	1.0	SF	18.0	0.0	Anomaly
19	0.0	0.02	0.03	SF	233.0	616.0	Normal

$$Gain(A) = Entropy(D) - Entropy_A(D).$$

يعرض الجدول عينة عشوائية من سجلات البيانات D اخذت من مجموعة البيانات الشبكية nls-KDD99.

جدول (١) عينة عشوائية لمجموعة سجلات nls_KDD99 data set

لحساب Gain(flag) اعتمادا على الخاصية class يتم حساب المعادلات الآتية :

١- نحسب Entropy (D) كما يلي :

$$Entropy(D) = -((7/19)\log_2(7/19) + (12/19)\log_2(12/19)) = 0.949 \text{ bits}$$

٢- نحسب EntropyA(D) لكل خاصية في D فمثلا A= flag

S0=8	SF=9	REJ=2	Class
0	7	0	Normal
8	2	2	Anomaly

$$Entropy_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Entropy(D_j)$$

$= -(8/19)((0/8)*\log(0/8) - (8/8)\log(8/8)) + (-9/19((7/9)\log(7/9) - ((2/9)\log(2/9)) + (-(2/19)$
$((0/2)\log(0/2) - (2/2)\log(2/2)))$.
$= -(8/19)(\log_2 1) + (-(9/19)((7/9)*\log_2(7/9) - (2/9)*\log_2(2/9))) + (-(2/19)(-\log_2 1))$
$= 0 + 0.473(0.77*\log_2 0.77) - (0.2*\log_2 0.2) + 0$
$= 0.178 + 0.464$
$= 0.642$ bits

٣- نحسب الـ information gain للخاصية flag :

$$Gain(flag) = Entropy(D) - Entropy_{flag}(D) = 0.949 - 0.642 = 0.307 \text{bits}$$

التقسيم بواسطة معيار information gain يعتمد على اختبار عدد المخرجات أي أن هذا المعيار يفضل اختيار الـ attribute ذات عدد أكبر من القيم. فمثلا الصفات ذات القيم الوحيدة unique attribute (مثلاً product_id) سوف تكون محل اهتمام لهذا المعيار وبالتالي فإن التقسيمات الناتجه عنه سوف تكون كبيرة جداً و وحيدة pure (بعدد قيم الـ unique attribute) ولذلك فإن infounique_attribute(D) =0 ، وبالتالي فإن هذا المعيار غير مناسب لعملية التصنيف لأنه ينتج عدداً كبيراً جداً من التقسيمات.

٤ -٢ -١ -٣ -٢ نسبة الكسب Gain Ratio

نسبة الكسب امتداداً للكسب المعلوماتي يتخلص من نقاط ضعفه ويطبق نوعاً من التطبيع على الكسب المعلوماتي باستخدام معلومات التقسيم "split information" لقيمة معرفة بشكل قياسي كما في المعادلة الآتية[١] :

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right).$$

هذا المعيار يأخذ في الاعتبار حساب عدد المخرجات الناتجة لاختبار الصفة على سبيل المثال، في خوارزمية شجرة القرارات C4.5 يتم استخدام معيار تقسيم Gain Ratio لتحديد الانقسام الأفضل ويعرف هذا المعيار بالمعادلة الآتية: [4] :

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)}.$$

الصفة التي يكون عندها نسبة الكسب أعلى قيمة يتم اختيارها كأفضل صفة تقسيم للمجموعة D .[1]

ولتوضيح ذلك نسبة الكسب للصفة flag من الجدول (1)

$Splitinfo_{flag}(D) = -(8/19)*\log_2(8/19) - (2/19)*\log_2(2/19) - (9/19)*\log_2(9/19)$

$= 0.525+0.342+0.510$

splitinfo(flag) =1.377 bits

Gainratio(flag)= gain(flag)/splitinfo(flag)

=0.307/1.377

=0.2229 bits .

٤ -٢ -١ -٣ -٣ مؤشر جيني Gini Index

معيار مؤشر جيني يحاول تجنب مشكلة الكسب المعلوماتي Information Gain بتقييد شروط الاختبار إلى تقسيمات ثنائية فقط بمعنى يتم حساب مجموع الشوائب الموزونة لكل قسم ، ويستخدم هذا المعيار مع خوارزمية CART .

Gini index يقيس شوائب D و مجموعة سجلات التدريب أو أقسام البيانات حسب المعادلة التالية :

$$Gini(D) = 1 - \sum_{i=1}^{m} P_i^2.$$

حيث $p_i = |C_{i,D}| / |D|$

- اذا الصفةA. تقسم قاعدة البيانات D الى D_1 و D_2 فإن الـ gini index يقسم D الى التقسيم التالي :

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2).$$

D1 تمثل مجموعة من السجلات في D تحقق الشرط A <= spilt_point

D2 تمثل مجموعة من السجلات في D تحقق الشرط A > spilt_point

تخفيض الشوائب المتوقع من الانقسام الثنائي على صفة A منفصلة أو مستمرة القيمة تحسب وفق المعادلة التالية :

$$\Delta Gini(A) = Gini(D) - Gini_A(D).$$

ولتوضيح ذلك لنفترض مجموعة التدريب D كما في الجدول (1) فإن عدد السجلات عندما الـ class =normal تساوي ٧ وعندما الـ class = anomaly تساوي ١٢ ، لإنشاء العقدة N التي تمثل جذر الشجرة لسجلات بيانات التدريب D نحسب الآتي :

اولاً نحسب الـ gini index لـ impurity of D حسب المعادلة التالية :

$Gaini(D) = 1 - (7/19)^2 - (12/19)^2$

$= 1 - 0.1357 - 0.39889 == 0.46541$ bits .

لإيجاد التقسيم المناسب لـ D باستخدام الخاصية flag يتم تقسيم مجموعة التدريب D إلى مجموعتين D1 و D2 كمايلي:

D1={S0,SF} و D2 ={REJ} عدد السجلات في D1 = 17 و عدد السجلات في D2 = 2

$Gini_{flag}(D) = (17/19)Gini(D1) + (2/19)Gini(D2)$

$= (17/19)(1-(7/17)^2 - (9/17)^2) + (2/19)(1-(2/2)^2 - (0/2)^2)$

$= 0.894(1-0.169-0.280)+0.105(1-1-0)$

$= 0.527$ bits .

الصفة التي لها قيمة Gini index أقل من قيمة الصفات الأخرى يتم إختيارها كصفة تقسيم لمجموعة بيانات التدريب D [١].

٤ -٥ استخدام تقنية التصنيف لكشف التطفل في قاعدة بيانات NLS-KDD99 data set

استخدم العديد من الباحثين بيانات KDD Cup 1999 لبناء نظم كشف التطفل وأظهرت الدراسات السابقة وجود بعض المشاكل الكامنة في هذه البيانات، إنّ التحديد المهم لهذه البيانات هو العدد الهائل للسجلات الزائدة بمعنى أن 78% من سجلات التدريب و 75% من سجلات الاختبار متكررة، هذه البيانات تعاني من بعض المشاكل وقد لا تكون مثلى للشبكات الفعلية الموجودة، ويمكن أن يكون محاكاة الهجوم ضمن واحد من الأصناف الأربعة (Dos , R2L , U2R , Probe).

مجاميع البيانات المتولدة KDD Train، KDD Test شملت (125973, 22544) سجلاً على التوالي. اقترحت بيانات NSL-KDD من قبل (Tavallaee et al.) حل مشاكل بيانات KDD المذكورة سابقاً تعتبر NSL-KDD التي تحوي في كل سجل اتصال TCP على ٤١ ميزة مع عنوان يوضح

هل هذا الاتصال هو اتصال اعتيادي أو نوع من أنواع التطفل ، وهناك ٣٨ ميزة رقمية و ٣ ميزة رمزية ، يأتي فوائد NSL-KDD مقارنة بمجموعة بيانات KDD الأصلية .

- لا تشمل سجلات زائدة في مجموعة التدريب ولن تميل المصنفات باتجاه سجلات أكثر حدوثاً.
- عدد السجلات المختارة من كل مجموعة : مستوى الصعوبة يتناسب عكسياً ونسبة السجلات في مجموعة KDD الأصلية، ونتيجة لذلك نسب تصنيف طرائق تعليم الآلة المتميزة تختلف بمدى واسع، مما يجعل 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 للا ختبار العشوائي لنسبة ضئيلة نتيجة لذلك سيكون تقييم نتائج البحوث المختلفة ثابتة ومشابهة.
- ٤ -٦ تدريب خوارزمية c4.5 على NLS-KDD
- استخدمت أداة Weka 3.7.9 لتدريب خوارزمية C4.5 على بيانات التدريب NLS-KDD Train لبناء مصنف البيانات وتم اختباره على بيانات الاختبار NLS-KDD Test ثم تم تقييم أداء المصنف من خلال حساب دقة التصنيف ونسبة الخطأ في التصنيف يمكن قياس دقة المصنفات بحساب الآتي:
- نسبة الضبطية (Precision) : نسبة سجلات الـ class A المصنفة بشكل صحيح إلى السجلات المصنفة في الـ class A.

أ - نسبة الاستدعاء Recall : نسبة سجلات الـ class A المصنفة بشكل صحيح إلى عدد السجلات في الـ class A .

ب - الإيجابية الكاذبة False positive (FP) أو الإنذار الكاذب false alarm : أي يتم تصنيف السجلات على أنها attacks وهي في الحقيقة normal

ت - السلبية الكاذبة False Negative (FN) : تصنيف السجلات على أنها طبيعية وهي في الحقيقة مهددات . وهذه الهجمات هي الهدف من عملية الكشف.

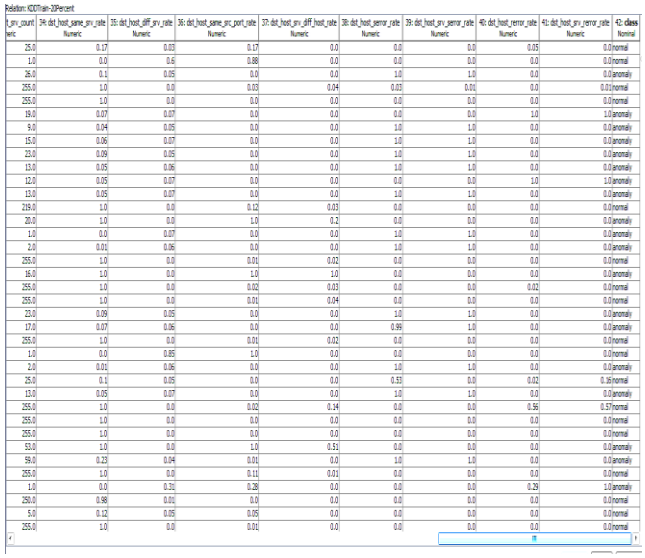
ث - الإيجابية الصادقة True Positive (TP) : تصنيف السجلات على أنها normal وهي في الحقيقة normal .

ج -السلبية الصادقة False Negative (FN) : تصنيف السجلات على انها attacks وهي في الحقيقة attacks.

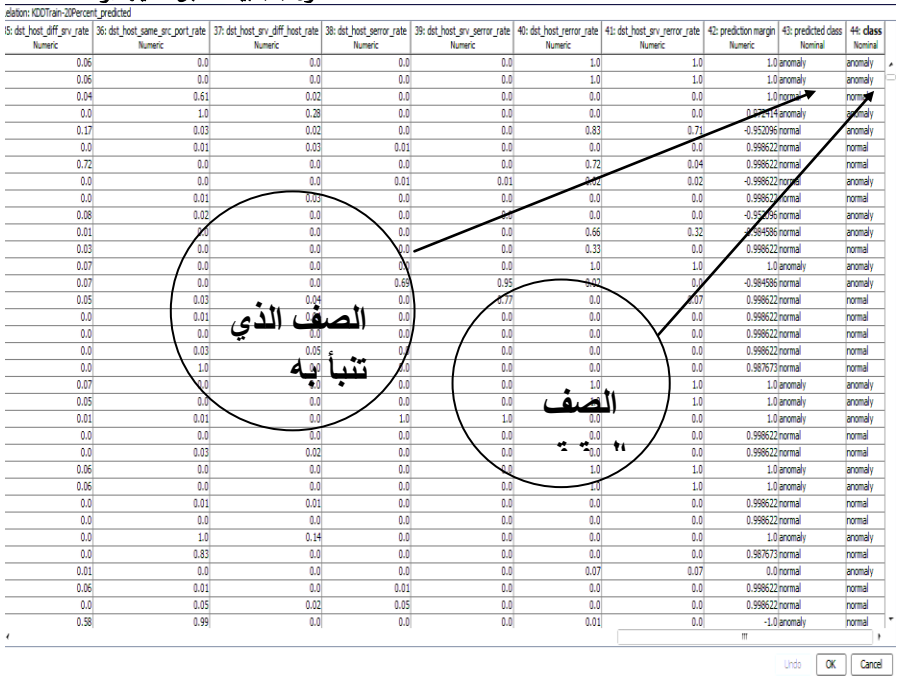
يتم حساب دقة المصنف اعتماداً على معدل الكشف Detected Rate و معدل الإنذار الكاذب False Alarm Rate.

استخدمت هذه الورقة مجموعة عشوائية من بيانات NLS-KDD وكانت نتائج التدريب كما يأتي :

٤ -٥ -١ نتائج تدريب الخوارزمية :



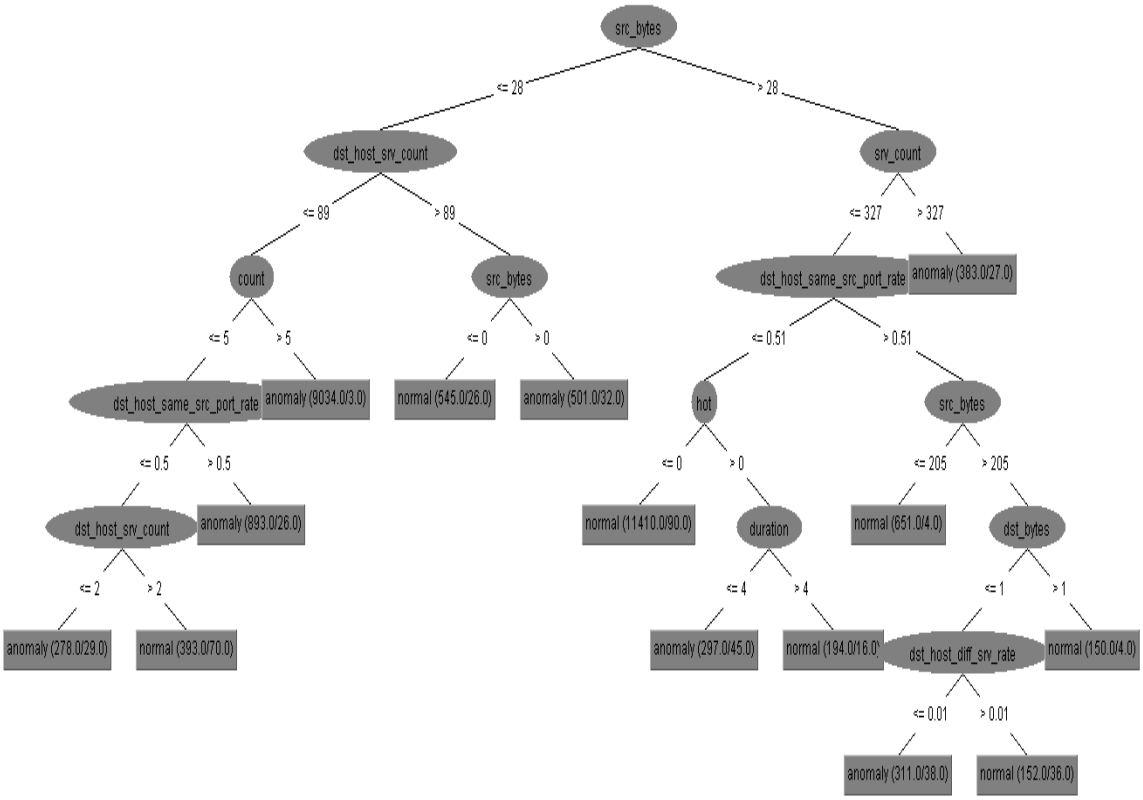
جدول (١)البيانات قبل تصنيفها بواسطة المصنف



جدول رقم (٢) البيانات بعد تصنيفها بواسطة المصنف

يبين الجدول (2) نتيجة اختبار المصنف C4.5 على بيانات الاختبار حيث يوضح الجدول الصف الحقيقي (class nominal) والصف المتوقع بواسطة المصنف (predicted class nominal) الذي يستخدم كصف معتمد لتصنيف هذه السجلات ، كما يبين الشكل أن السجلات التي صنفت بالفعل على أنها طبيعية normal هي في الأصل طبيعية والسجلات التي صنفت على أنها مهددات هي في الأصل مهددات بالفعل، ولكن توجد نسبة خطأ في تصنيف بعض السجلات التي صنفت على أنها مهددات وهي بالحقيقية سجلات طبيعية وكذلك سجلات صنفت على أنها طبيعية وهي بالحقيقة مهددات .

٤ -٥ -٢ مصنف شجرة القرار الناتج :



شكل(٥) مصنف شجرة القرارالناتج لقاعدة بيانات NLS-KDD data set

٤ -٦ تقييم المصنف Evalution Classifier

مصفوفة التضارب الناتجة لعملية اختبار المصنف على ٢٠% Of NLS-KDD Data Set Test

		Predicted class	
		Class = normal	Class = anomaly
Actual class	Class = normal	١٨٧٩	٢٧٣
	Class = anomaly	٣٩٩٦	٥٧٠٢

شكل رقم (٦) مصفوفة التضارب الناتجة لإختبار المصنف C4.5 على ٢٠% Of NLS-KDD Data Set Test

Accurcy = (1879 + 5702) / (1879 + 273 + 3996 + 5702)

= (7581) / (11850)

= 0.639 = 63.9 %

Errore rate = (٣٩٩٦ + ٢٧٣) / (١٨٧٩ +٢٧٣ + ٣٩٩٦ + ٥٧٠٢)

= (4269) / (11850)

= 0.360

= 36%

true positive rate (TPR) or sensitivity= TP/(TP + FN)

= 1879 /(1879 + 273) = 87 %

true negative rate (TNR) or specificity = TN/(TN + FP).

= 5702 / (5702 + 3996) = 59%

false positive rate (FPR) = FP/(TN + FP),

= 3996 / (5702 + 3996) = 41%

false negative rate (FNR) = FN/(TP + FN).

= 273/(1879 + 273) = 13%

Precisoin = tp /(tp + fp)

= 1879 / (1879 + 3996) = 32%

Recall = tp / (tp + fn)

= 1879 / (1879 + 273) = 87%

تلخص نتائج التجربة1 في الجدول التالي :

TPR	FP	TP	FN	Precisoin	Recall	Errore rate	Accurcy
87 %	41%	58%	13%	32%	87%	36%	63.9 %

شكل رقم (٧) نتائج اختبار المصنف C4.5 على بيانات 20% Of NLS-KDD Data Set Test

مصفوفة التضارب الناتجة لعملية اختبار المصنف على Full NLS- KDD 99 Data Set Test

		Predicted class	
		Class = normal	Class = anomaly
Actual class	Class = normal	٩٤٤٨	٢٦٣
	Class = anomaly	٣٩٠٠	٨٩٣٣

شكل رقم (٦) مصفوفة التضارب الناتجة لإختبار المصنف C4.5 على Full NLS-KDD 99 Data Set Test

نتائج التجربة2 :

TP	FP	TN	FN	Precisoin	Recall	Errore rate	Accurcy
97 %	59%	69.6%	2.7%	70.8%	97%	36%	٨١,٥٣٣٩%

شكل رقم (٩) نتائج اختبار المصنف C4.5 على ب على Full NLS-KDD 99 Data Set Test

٥ -الخلاصة CONCLUSION :

إذا أردنا استخدام الشبكة فإن هناك حاجة ملحة لكشف هجوم أنظمة الشبكات، وفي هذه الورقة استخدمت واحدة من أهم تقنيات استخراج البيانات وهي C4.5 للكشف عن الشذوذ في الشبكة. نتيجة التجربة المعروضة سابقاً اظهرت خوارزمية C4.5 نتيجة فعالة في كل من كشف الشذوذ ومعدل الانذار الكاذب في مجموعة البيانات المتوفرة، وأظهرت النتائج أنه كلما كانت كمية البيانات كبيرة تكون نسبة الخطأ أقل ودقة تنبؤ المصنف عالية حيث كانت دقة المصنف % 63.9 عندما كانت كمية بيانات الاختبار (946,848 byte) وعندما زادت كمية بيانات الى(497,700 byte) ارتفت دقة المصنف الى ٨١,٥٣٣٩٪، بينما كانت كمية بيانات التدريب ثابتة، هذه النتائج تدلل قدرة المصنف العالية للتعلم كلما زاد حجم كمية البيانات.

٦- مراجع:

1. Data Mining: Concepts and Techniques Third Edition - Jiawei Han - University of Illinois at Urbana–Champaign - Micheline Kamber Jian Pei Simon Fraser University- Morgan Kaufmann is an imprint of Elsevier.-
2. Data Mining for Network Intrusion Detection Paul Dokas, Levent Ertoz, Vipin Kumar, Aleksandar Lazarevic, Jaideep Srivastava, Pang-Nig Tan ,Computer Science Department, 200 Union Street SE, 4-192, EE/CSC Building ,University of Minnesota, Minneapolis, MN 55455, USA ,aleks@cs.umn.edu ,srivasta@cs.umn.edu ,kumar@cs.umn.edu .
3. Zirkle, L., "What is host-based intrusion detection? "Virginia Tech CNS. SANS -Institute Resources, Intrusion Detection FAQ, Hyperlink ID FAQ, 2000.
4. ISSN (Print): 2279-0047 ISSN (Online): 2279-0055 - International Journal of Emerging -Technologies in Computational and Applied Sciences (IJETCAS) www.iasir.net,Differentiating Network Attacks using C4.5 Algorithm with Multiboosting,V.Balaji1, Varalakshmi.K2 # Department of Computer Science and Engineering, SRM University, SRM Nagar, Kattankulathur – 603203, Kancheepuram District, Tamil Nadu, INDIA.
5. Introduction to Data Mining- by Tan, Steinbach, Kumar – 2004

جامعة الناصر
AL-NASSER UNIVERSITY