Journal of Kerbala University , Vol. 14 No.1 Scientific . 2016

Proposal to Enhance NIDS
مقترح لتحسين نظام الشبكة لكشف التطفل

Asst. Prof. Dr. Soukaena Hassan Hashem
University of Technology/Computer Sciences Department
Soukaena.hassan@yahoo.com

Abstract

Proposed work aim to build a proposed Gain Association Rules -Based Network Intrusion Detection System (GARNIDS). GARNIDS trend to enhance traditional NIDS through using three of data mining algorithms; these are: Gain which is measure the entropy for each feature to detect it is Domination Degree (DD) for each attack, then feeding these features with their DD to a proposed Gain Association Rule (GAR) algorithm that to rank the features according to two parameters (frequency and DD). Finally customize K Nearest Neighbor (KNN) as misuse classifier (detect the intrusions and specify their types) the proposal assume the k equal to 3.

Many experimental works are conducted to evaluate the proposal over the KDD'99 dataset and show the efficiency of KNN through registering 86% of accuracy with all features, 90% of accuracy with 25 top features and the accuracy was 98% with 8 top features. Also the Detection Rate (DR) and False Alarm Rates (FAR) are both measured with those three cases and still KNN with the top 8 features is the higher in DR and lower in FAR. Finally when try the proposal in real-time with tcpdump the third case register higher accuracy (93%).

Keyword: NIDS, AR, KNN, Gain, feature selection, detection rate, accuracy.

الخلاصة

العمل المقترح يهدف لبناء نظام شبكة لكشف التسلل المستند على الكسب للقواعد المترابطة (GARNIDS) . GARNIDS توجه النظام لتعزيز NIDS التقليدية من خلال استخدام ثلاثة من خوارزميات التنقيب عن البيانات؛ وهي: الربح الذي هو قياس الكسب لكل خاصية للكشف عن درجة الهيمنة لها (DD) لكل هجوم، ثم تغذية هذه الخصائص مع DD لخوارزمية القواعد المترابطة المعتمدة على الكسب لترتيب الخصائص وفقا لمعلمتين (التكرار ودرجة الهيمنة). وأخيرا خوارزمية أقرب جار(KNN) ، استخدامت كمصنف من نوع اساءة الاستخدام (كشف الاختراقات وتحديد أنواعها) اقترح النظام عدد الجيران يساوي 3.

تم اجراء العديد من الأعمال التجريبية لتقييم الاقتراح على مجموعة البيانات KDD'99 واظهر كفاءة KNN من خلال تسجيل 86٪ من الدقة مع كافة الخصائص، 90٪ من الدقة مع أهم 25 من الخصائص وكانت الدقة 98٪ مع أهم 8 خصائص. أيضا معدل الاكتشاف (DR) ومعدلات الإنذار الكاذبة (FAR) كلاهما تم قياسه مع الحالات الثلاث، ولا تزال KNN مع أهم 8 خصائص هي أعلى DR وأقل FAR. أخيرا حاول الاقتراح تنفيذ النظام بالوقت الحقيقي بواسطة TCPDUMP وتم ملاحظة ان الحالة الثالثة سجلت أعلى قدر من الدقة 93%.

1. General Introduction

The detection of intrusion (intrusion attempts) operates with records and information supplied by network system. ID is most critical content of environment of security technology. DM-based ID techniques generally fall into two main categories: misuse detection and anomaly detection. In misuse detection systems, use patterns of well-known attacks to match and identify known intrusion. Anomaly detection, on the other hand, builds models of normal behavior, and flags observed activities that deviate significantly from the established normal usage profiles as anomalies, that is, possible intrusions [1, 2, 3, and 4].

The selection of features is the important stage in constructing intrusion detection systems. Through this stage, the collection features can be consider the most essential features to build proper detection mechanism. The main difficulty that most scientists face, is selecting the proper group of attributes, because of whole attributes are not pertinent to the learning algorithm, but in most instances, unrelated and repeated attributes have the ability to generate noisy information that divert the learning algorithm [5, 6, and 7]. Information gain (IG) measures the amount of information in bits about the class

prediction, if the only information available is the presence of a feature and the corresponding class distribution [7, 8, and 9].

Frequent itemset mining leads to the discovery of associations and correlations among items in large transactional or relational datasets. With massive amounts of data continuously being collected and stored, many industries are becoming interested in mining such patterns from their DBs. The discovery of interesting correlation relationships among huge amounts of business transaction records can help in many business decision-making processes, such as catalog design, cross-marketing, and customer shopping behavior analysis [10, 11].

A technique that classifies each record in a dataset based on a combination of the classes of the k record(s), sometimes called the k-nearest neighbor technique. It is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions) [12].

2. Proposal of GARNIDS

Security still the important field in all e-business over the internet, that since web sites have sensitive data and intruder still penetrate them. So a solutions must be taken to overcome these threats, IDS is the best solution to do that by build a learning system trained on previous registered normal and intrusions sessions. The proposal aim to enhance the NIDS; that by suggest GARNIDS which use KNN as a misuse classifier and GAR as a proposed features extraction depending on gain and frequency of features. The GARNIDS algorithm will be explained in the following sequential stages see algorithm (1).

Algorithm (1) GARNIDS

Input: KDD'99 for learning (training and testing) and tcpdump for real-time implementation

Output: Rules for detecting the intrusions

Steps:

1. Use KDD'99 dataset for training and testing; normalize these dataset to range their features' values from 0 to 1 to avoid the bias in learning.
2. Proposing GAR algorithm; which has the following steps, ***see algorithm (2)***;
   - For each feature in dataset calculate the gain to detect the DD for the feature with the all types of intrusions and normal. DD is measured depending on the entropy.
   - AR mining algorithm modified to evaluate the association rules by the frequency and DD, so the generated association rules have two parameters not just the frequency but also the DD. So the confidence will be calculated according to these two parameters.
   - The results of proposed GAR will extract the high domination features in two dimensions (frequency and entropy).
3. Train and test the KNN misuse classifier, ***see algorithm (3)***, for three cases these are; all features, top 25 features and top 8 features.
4. Evaluate these three cases of KNN classifier to depend the most precision one depending on rates of detection, accuracy and false alarms.
5. Validate the three cases in real-time environment using tcpdump to verify the results of testing and measure the time spent for detection.
6. End.

2.1 GAR Algorithm

This section will explain the proposed algorithm which hybrid the gain with association rules to extract the correlated critical feature for dataset.

Algorithm (2) GAR
Input: Normalized KDD'99 dataset
Output: Correlated features extracted by GAR.
Steps:
1. Give all items (features) their DD; the DD of features will be taken by its correlation to the classes of KDD'99; this correlation will be measured by *Gain.*

Gain (Feature i) = Info (KDD'99) – Info (KDD'99 (Feature i)) ……. (1)

$$Info\ (KDD'99) = -\sum_{i=1}^{m} Pi\ Log\ 2\ (Pi) \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2)$$

Where m= number of classes in the KDD'99 dataset, Pi = probability of appearance the specified class' session from all number of sessions in KDD'99.

$$Info\ (Dataset\ (Feature\ i)) = \sum_{j=1}^{n} \frac{|Dataset j|}{|Dataset|} * I\ (D\ j) \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (3)$$

Where n= number of feature' values in the KDD'99 dataset, dataset = KDD'99, dataset j = dataset with the specified feature' value.

2. Detect minimum support =50% and minimum confidence = 100%.
3. Find the frequent value of each feature, start with single features, determine the most frequent among them and specify two parameters for each feature (DD and frequency). Eliminate the features their frequency less than 50%, unless their DD more than 80%.
4. While all single features are registered with the two parameters create feature-sets which consist of multi features and determine the frequent feature-sets and specify two parameters for each feature-set (DD and frequency).
5. Generate association rules for all single features and feature-set with two parameters (DD and frequency). The confidence has two values; first is the Traditional-Confidence calculated from the support of features and the second is DD-Confidence which calculated as in traditional. Finally average Traditional-Confidence and DD-Confidence values to calculate the Final-Confidence.

$$Traditional\text{-}Confidence = (X \rightarrow Y) = \frac{support\ (X \cup Y)}{support\ (X)} \quad \ldots\ldots\ldots\ldots\ldots \quad (4)$$

$$DD\text{-}Confidence = (X \rightarrow Y) = \frac{DD\ (X \cup Y)}{DD\ (X)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (5)$$

$$Final\text{-}Confidence = \frac{(Traditional-Confidence)+(DD-Confidence)}{2} \quad \ldots\ldots\ldots\ldots \quad (6)$$

6. The last step is to select the higher Final-Confidence values to extract the most correlated feature-sets to detect them as critical correlated features.
7. End

2.2 KNN Classifier Algorithm
This section will explain the proposed algorithm of KNN classifier which will trained and tested over KDD'99 in three cases (all features, 25 top features and 8 top features) to standardize the most accurate model.

Algorithm (3) proposed-KNN Classifier
Input: KDD'99 with three cases of features (all, 25 and 8).
Output: Optimal KNN (classifier model).

Steps:

For I = 1 to 3 (try to repeat the classifier 3 times according the available 3 cases)

1. From the training datasets (according case) scan the sessions
2. Initialize K=3 (in the proposal)
3. All of features must be normalized so that, the values range from 0 to 1 (for more accuracy since the values are variants). This done using

$X=X/(1+X)$ ……. *(7).*

4. Do the reverse reduction of overall sessions according the k=3.
5. Testing; will begin by take all sessions from the testing datasets (according the case). Then Find KNN in the training datasets based on similarity functions, in the proposal Euclidian distance

$D(session\ i, session\ j) = \sqrt{\sum_{i=1}^{n}(session\ i - session\ j)^2}$ …….. (8)

Where n no. of features and will be variable according the case (41, 25 and 8).

6. Detect the class prediction done by detect the maximum class introduces in the KNN.
7. Set the KNN as the classifier model to classify testing dataset of both modern KDD and CWDS.

End For

End

3. Experimental Work and Results

This section will explain the experimental work and results of GARNIDS, with KDD'99 the number of features is 41 features and types of sessions is 5 as a general classes. Table (1) displays the number of training and testing samples of modern dataset.

Table (1) Number of samples for training and testing

Connection Types	Training	Testing
Normal	95,000	20,000
Denial of Services	85,000	40,000
Remote to User	103,000	20,000
User to Root	57,000	20,000
Probing	70,000	30,000
Total Number	410,000	130,000

According the three cases of learning will explain the results of proposal, so first of all must explain the results of the GAR algorithm. To explain the results in details will show all features and GAR features in KDD'99, see table (2).

Table (2) DD using Gain on KDD'99 Features

No. of Features	Original Sequence of Feature	Features according GAR (25)	Features according GAR (8)
1.	Duration	Service	Service
2.	Protocol_Type	Duration	Duration
3.	Service	Land	Land
4.	Src_Bytes	Urgent	Urgent
5.	Dst_Bytes	num_failed_logins	num_failed_logins
6.	Flag	Flag	Flag
7.	Land	num_root	num_root
8.	Wrong_Fragment	num_file_creations	num_file_creations
9.	Urgent	num_shells	
10.	(hot)	num_outbound_cmds	
11.	(num_failed_logins)	is_hot_login	
12.	(logged_in)	srv_diff_host_rate	
13.	(num_compromised)	dst_host_count	
14.	(root_shell)	dst_host_srv_count	
15.	(su_attempted)	dst_host_same_srv_rate	
16.	(num_root)	dst_host_diff_srv_rate	
17.	(num_file_creations)	dst_host_same_src_port_rate	
18.	(num_shells)	dst_host_serror_rate	
19.	(num_access_files)	dst_host_rerror_rate	
20.	(num_outbound_cmds)	Protocol_type	
21.	(is_hot_login)	count, srv_count	
22.	(is_guest_login)	root_shell	
23.	(count)	same_srv_rate	
24.	(serror_rate)	Src_Bytes	
25.	(rerror_rate)	Dst_Bytes	
26.	(same_srv_rate)		
27.	(diff_srv_rate)		
28.	(srv_count)		
29.	(SRV_serror_rate)		
30.	(SRV_rerror_rate)		
31.	(SRV_diff_host_rate)		
32.	(DST_host_count)		
33.	(DST_host_srv_count)		
34.	(DST_host_same_srv_rate)		
35.	(DST_host_diff_srv_rate)		
36.	(DST_host_same_src_port_rate)		
37.	(DST_host_srv_diff_host_rate)		
38.	(DST_host_serror_rate)		
39.	(DST_host_srv_serror_rate)		
40.	(DST_host_rerror_rate)		
41.	DST_host_srv_rerror_rate		

The experiments with the testing session (130,000 sessions), as listed in table (1), introduce that the subsets of features extracted by GAR are the higher detection rate, higher accuracy and minimum false alarms, see table (3).
The DR computed by;

$$DR = TP/ (TP+FN)*100 \quad (9)$$

False Alarm Rate (FAR) calculated by;

$$FAR = FP/ (TN+FP) *100 \quad (10)$$

The classification accuracy measures the proportion of correctly classified cases;

$$Accuracy= (TP+TN) / (TP+FP+ TN+ FN) *100 \quad (11)$$

Table (3) Results of testing the proposal KNN with three cases

Classifier	Feature	TP (intrusion)	TN (normal)	FP (F-intrusion)	FN (F-normal)	DR	FAR	Accuracy
KNN	41 features	95000	15000	5000	15000	0.863	0.25	0.846 (85%)
	25 features	100000	17000	3000	10000	0.909	0.15	0.900 (90%)
	8 features	108000	19000	1000	2000	0.981	0.05	0.976 (98%)

From table (3), will see higher accuracy, higher detection rate and less alarms are with the subset of 8 features. By using tcpdump in real-time to validate the proposal of KNN with the subset of 8 features for 200,000 sessions with 50,000 normal and 150,000 attacks, the results were as in table (4) below.

Table (4) Results of testing the proposal KNN with tcpdump

Classifier	Feature	TP (intrusion)	TN (normal)	FP (F-intrusion)	FN (F-normal)	DR	FAR	Accuracy
KNN	8 features	140000	45000	5000	10000	0.933	0.10	0.925 (93%)

4. Conclusion

From results obtained in implementing the GARNIDS reached to the following conclusions:

1. Using KDD'99 as a learning dataset and using tcpdump as a validation and verification tracer make the results obtained by testing most reliable, since the validation give very near results of testing, see table (3) and table (4).
2. Using Gain as Domination Degree procedure and hybrid it with AR to modify the traditional confidence which depend on frequency only to depend on both frequency and DD give a higher correlation for features with each other's and for features and classes.
3. Using KNN as classifier give a good results in the three cases of features, which proof the efficiency of the proposed KNN for intrusion detection system.

References

1. Hashem S. H. (2013), Efficiency Of SVM And PCA To Enhance Intrusion Detection System, *Journal Of Asian Scientific Research, 3(4):381-395*.
2. Hashem S. H. and Ali I. (2013), A Proposal To Detect Computer Worms (Malicious Codes) Using Data Mining Classification Algorithms, *Eng. &Tech. Journal .Vol31,Part (B), No. 2.*
3. Hashem S. H., Majeed S. K., and Gbashi I. K. (2013), Propose HMNIDS Hybrid Multilevel Network Intrusion Detection System, *IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 5, No 2, ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784, www.IJCSI.org*
4. Zhou Q. and Zahao Y. (2013), The Design and Implementation of Intrusion Detection System based on Data Mining Technology, *Journal of Applied Sciences, Engineering and Technology 5(14): 3824-3829, ISSN: 2040-7459; e-ISSN: 2040-7467 © Maxwell Scientific Organization.*
5. Lee W., Stolfo S. J. Mok K. W. (1999), A data Mining Framework for Building Intrusion Detection Models, *Proceeding of IEEE Symposium on Security and Privacy, pp 120-132.*
6. The UCI KDD Archive, Information and Computer Science, KDD Cup 1999 Data, *University of California, Irvine, 1999, available at: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.*
7. Hashem S. H., Majeed S. K., and Gbashi I. K. (2013), Proposal To Wnids Wireless Network Intrusion Detection System, *Ijsr - International Journal Of Scientific Research Volume : 2 | Issue : 10 | October 2013 • Issn No 2277 – 8179.*
8. Barhoo T. S. and ElShami E. (2011), Detecting WLANs' DoS Attacks Using Backpropagate Neural Network, *Journal of Al Azhar University-Gaza (Natural Sciences), 2011, 13 : 83-92,* http://www.alazhar.edu.ps/journal123/natural_Sciences.asp?typeno=1
9. Tulasi R. L. and Ravikanth M. (2011), Impact Of Feature Reduction On The Efficiency Of Wireless Intrusion Detection Systems", *International Journal Of Computer Trends And Technology- July To Aug Issue, Issn: 2231-2803 Http://Www.Internationaljournalssrg.Org Page 171.*
10. Jiawei Han, and Micheline Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmaan Publishers, 2006.
11. Shaimaa A. Hassan, "A Technique for Mining Association Rules in Multidimensional Databases", M.Sc. thesis, University of Technology Department of Computer Sciences, 2008.
12. Y. Angeline Christobel & P. Sivaprakasam" Improving The Performance Of K-Nearest Neighbor Algorithm For The Classification Of Diabetes Dataset With Missing Values"- International Journal Of Computer Engineering & Technology Volume 3, Issue 3, October - December 2012.