Enhance Network Intrusion Detection System Using Bee Algorithm

Soukaena H. Hashem[1], Shatha Habeeb[2], and Besmah M. Khalil[3]

[1]University of Technology, Computer Sciences, Baghdad, Iraq (soukaena_hassan@yahoo.com)
[2]University of Technology, Computer Sciences, Baghdad, Iraq (shathahabeeb@yahoo.com)
[3]Al-Rafidain University College, Baghdad, Iraq (beseng1977@yahoo.com)

Abstract

Intrusion detection systems have sequential steps begin with selecting training and testing dataset, the preprocessing dataset, selecting most important features, and finally constructing the most reliable classifier. This research focuses on building a reliable Network Intrusion Detection System (NIDS) to detect traditional and modern attacks with minimum number of features. The proposal creates dataset depending on KDD. The proposal will inject KDD with new sessions to represent most modern attacks. This update requires adding new features for the dataset, since these features are critical to detect these modern attacks. The proposal considers updated dataset without any assumptions says that the dataset is idealism, so there are preprocessing steps to be done to make it consistence for training and constructing the classifier. Meta heuristic bee's algorithm will be used as Feature Selection technique with the support of two of statistical ranking filters. The ranking of features with bee give an optimized ordering to the most critical and intrinsic features in the space of training and constructing classifier. The results obtained by constructing the most reliable classifiers Interactive Dichotomizer 3 classifier (ID3), Naïve Bayesian Classifier (NB), Artificial Neural Network (ANN) and Support Vector Machine (SVM) depending on both updated dataset and bee's ranked features sets give effective efficiency in reducing false alarms and increasing detection rates.

Keyword: IDS, feature reduction, KDD training data, bee's algorithm.

تحسين نظام كشف التطفل الشبكي باستخدام خوارزمية النحل

سكينة هـ. هاشم[1] وشذى حبيب[1] وبسمة م. خليل[2]

[1]الجامعة التكنولوجية، علوم الحاسبات، بغداد، العراق
[2]كلية الرافدين الجامعة، بغداد، العراق

خلاصة

يتكون تحسين نظام كشف التطفل الشبكي باستخدام خوارزمية النحل ومجموعة بيانات KDD أنظمة كشف التطفل من خطوات متتابعة تبدأ مع اختيار مجموعة البيانات الخاصة بالتدريب والاختبار ، ثم المعالجة الأولية لمجموعة البيانات واختيار الخصائص المهمة، وأخيرا بناء المصنف الأكثر موثوقية. يركز هذا البحث على بناء نظام كشف تطفل شبكي لكشف وتحديد الهجوم التقليدي والحديث مع اقل عدد من الخصائص. يقوم المقترح بخلق مجموعة بيانات تعتمد على الـ KDD . المقترح سوف يقوم بضخ جلسات ارتباط شكلية الى KDD تمثل المجوعات الأكثر حداثة. يتطلب هذا التحديث إضافة خصائص جديدة لمجموعة البيانات حيث أن هذه الخصائص تكون مهمتها تحديث الهجمات الحديثة. يأخذ المقترح بنظر الاعتبار مجموعة البيانات المحدثة بدون أي افتراضات تقول أن مجموعة البيانات مثالية ولذلك هناك خطوات معالجة أولية سوف يتم إيجازها لجعل مجموعة البيانات متناسقة في رحلة بناء واختيار المصنف. استخدمت خوارزمية النحل كتقنية اختبار الخصائص مع دعم من اثنين من مرشحات الترتيب الاحصائية في فضاء بناء واختيار المصنف. تعتمد النتائج المستحصلة بواسطة بناء المصنف الأكثر موثوقية *ANN* ,*NB* ,*ID3* على كلاهما مجوعة البيانات المحدثة وخوارزمية النحل كتقنية لترتيب الخصائص حسب أهميتها أعطت نتائج مهمة جدا . في تقليص الإنذارات الخاطئة وزيادة نسبة الكشف.

كلمات مفتاحية: IDS، إختزال الخصائص، مجموعة بيانات KDD، خوارزمية النحل.

1. Introduction

Intrusion Detection is a security service that monitors and analyzes system events to find, and provide (real-time) warning of unauthorized access attempts to resources. The intrusion detection systems are classified as: Host-based IDS (HIDS): monitor single host activity, Distributed Host-based IDS combining info from multiple hosts, and Network-based IDS (NIDS) monitor network traffic.

There are two approaches of IDS, often used in combination, these are: anomaly detection which defines normal behavior threshold detection and profile based signature detection that defines proper behavior Sequence of events [1, 2].

The **Bees Algorithm** is a new population-based search algorithm that mimics the food foraging behavior of swarms of honey bees. In its basic version, the algorithm performs a kind of *neighborhood search* combined with *random search* and can be used for optimization problems [3,4].

KDD 2000 is a training data that consists of the first seven weeks of traffic with approximately 4.9 million connections and the testing data consists of the last two weeks of traffics with approximately 300,000 connections. It injected with new types of attacks that were not exist in training data. Each record consists of 41 features of various types as well as a class label that is either normal or one of attack types. The classes in KDD dataset can be categorized into five main classes (one normal and four main intrusion classes: probe, Denial of Service (DOS), User to Root (U2R), and Remote to Local (R2L)). These four classes are divided into 22 different attacks which they are: DOS (back, land, Neptune, pod, smurt and teardrop), R2L (ftp_write, guess_password, imap, multihop, phf, spy, warezclient, and warezmaster), U2R (buffer_overflow, perl, loadmodule, and rootkit) and Probing (ipsweep, nmap, portsweep, and satan) [5-10].

2. Related Works

Pietraszek [5], has proposed machine learning method for IDS alert classification, in order to reduce the amount of false positives. Viinkka et al [6], have suggested the use of time series modeling for modeling regularities in large alerts volumes. Vaarandi [7], proposed IDS alerts classification algorithm which distinguishes important alerts from redundant ones. The author improved his work by proposing algorithms that suggest an IDS alert classification method which is based on frequent itemset mining and data clustering algorithm. Eunhye Kim, et al [8], statistical feature construction scheme is proposed in which factor analysis is orthogonally combined with an optimized k-means clustering technique. Also SOM is performed for unsupervised anomaly detection. Dewan Md. Farid , et al [9], a new learning algorithm for adaptive network intrusion detection using naïve Bayesian classifier and decision tree is presented. It performs balance detection and keeps false positives at acceptable level for different types of network attack. Also eliminate redundant attributes as well as contradictory examples from training data that make the detection model complex. Lee W., et al [10], famous datasets used in traditional and newest IDS is KDD CUP1999. In that dataset the intrusion data characterized into three sets of features, these are: basic features, content features, and traffic features. So this dataset describes network connection using of total 41 features that cover all the types of attacks to the greatest extent possible [10].

3. The Proposed Policy to Enhance NIDS

This research enhances IDS across enhancing two important stages which they are: selecting training and testing dataset and optimize feature space to include intrinsic features. Algorithm1 will explain the outlines of sequential stages, to enhance NIDS:

Algorithm1: Enhanced NIDS

Input: DARPA KDD, sessions present most modern attacks, and new features.

Output: Effective NIDS

Process:

1. Creating updated dataset to have various sessions from KDD (normal and all variations of attacks).
2. Inject the proposed dataset by sessions present most modern attacks.
3. Adding new features related to the injected session that present attacks not exist in KDD.
4. Preprocessing the created dataset since it will be a mixture of many resources and contains new features added to dataset. So there are many problems will appear such as noise, in complete attributes and missing values.
5. Proposing Bee algorithm for ranking the features depending on averaging two ranking methods. By applying proposed bee algorithm on features will register three cases 44 features, top 22 features and top 11 features.
6. Construct four classifiers such as: ID3, NB, NN and SVM on preprocessed updated dataset three times depending on the three cases considered with bee ranking. That is to evaluate the proposed IDS with various classifiers.
7. Allow the enhanced NIDS to be adaptive by reporting the stranger sessions and analyzing them to extract the new attacks appear in them. Then if there is a new feature must added to dataset must repeat all steps above, else just add the session to dataset and classify it with it is classifications.

End

3.1. Dataset Creation (inject sessions and features)

The created dataset used for training and testing most of its sessions taken from KDD. About quarter of the created dataset is injected by connection sessions that have most modern attacks. The proposed created dataset will be divided into two subsets, one for training and second for testing. These two subsets have 400,000 records, 300,000 records for training and 100,000 for testing. Most of these records are selected in very precise manner to have various types of normal and intrusion connections. The new types of attacks taken into account are:

1. Financial malware that has the ability to hijack customer's online banking sessions in real time using their session ID tokens.
2. Types of worms such as Conficker.
3. Java Script Obfuscation and Zeus Botnet Kit.

*These types of attacks could be taken under one name called **Extended Attack** which is collect most new attacks that not correlated with the famous four types of attack in DARPA dataset. The proposal increases the no. of features which seems important to be added because it related to the new attacks added as a connection session to the dataset. These added features are:*

*Connection-based traffic features are obtained using some knowledge of connection domain, such as **type of connection (wire or wireless), connection security (encrypted or not encrypted) and connection multimedia (image, video, sound and text).***

*By this proposed feature the no. of depended features will be **44 features,** and no. of general classes will be **6 instead of 5**. These Classes are: Normal connections, Denial of Service (DoS), Remote to User (R2L), User to Root (U2R), Probing (Probe), and Extended Attacks. For more explanation see table (1).*

3. 2. Dataset Preprocessing

In addition to the injected sessions there was features development along with all parts of dataset (parts taken from DARPA and parts injected to it). The proposed dataset has ratio of noise in its data records, this noise presents the most challenging issues in ID application which is aim to detect the intrusions using data mining techniques. Noise removal of dataset at the learning time is to avoid over-fitting the dataset. Treating noise can be done as in the following:

1. Treating missing attribute values by replacing their values with the most frequent attribute value in the dataset. But missing values in the proposal presented by the three features added in connection sessions injected, which they don't found in the sessions taken from KDD2000.
   - Connection types in all traditional KDD will fill with (wire, encoded 0).
   - Connection security 50% in traditional KDD will fill with (encrypted, encoded 0) and other 50% will fill with (unencrypted, encoded 1).
   - Connection multimedia 25% in traditional KDD will fill with (text, encoded 0), 25% will fill with (image, encoded 1), 25% will fill with (sound, encoded 2) and 25% will fill with (video, encoded 3).
2. Treating redundant examples by removes redundancy by keeping only a unique example in the dataset (some new sessions may redundant because it presents an old attack with new vision). By doing so, it will speeds significantly up the learning process.
3. Treating incomplete attribute problem by avoiding the essential attributes of a problem is not used to describe in the dataset (by adding the three proposed features, this problem was solved).
4. Treating misclassified examples by labeled with a true classification instead of wrong classification (in the proposal the injection of session must be real, mean by real the injected sessions taken from network connected with Internet and deal with all connection types, media and encrypted/unencrypted sessions).

3.3. Feature Selection (propose bee algorithm as ranking method)

The most important step in building IDS is how to characterize the important features they will based in increasing detection rate and optimized trigger alarms (reduce false positive alarms, reduce low important alarms, and reduce false negative alarms). By optimizing features the data space will also optimized, so the training dataset and training time will be more efficient for classification that work under real time environment.

The proposal presents the metaheuritic algorithm (Bee) as a feature ranking algorithm that by making the following assumptions:

1. The weights of features will be taken by its correlation to the 6 classes; this correlation will be measured by average of two ranking methods *Chi-Square and Gain Ratio.*
2. Some terminologies in Bee algorithm will be replaced according to the proposal of feature selection, these are:
    - n the scout bees will be; n no. of features
    - m sites and e best sites will be; m selected features and e best features
    - nep no. of bees recruited will be; nep weight given to e best features
    - nsp no. of bees recruited will be; nsp weight given to (m-e) features
    - Patches will be; features set.
    - Neighborhood for features will be; other features in the same type (as in the proposal there are 6 types) then features in other feature type's subset.

So after interpretations in the two points above the proposal bee algorithm for feature ranking will be introduced in the following Algorithm2.

Algorithm2: Proposed Bee Algorithm for Feature Ranking

Parameters

1. *n*: number of all known features
2. *m*: number of features selected out of *n visited* features
3. *e*: number of best features out of *m* selected features
4. *nep*: weight given for best *e* features (rich)
5. *nsp*: weight given for other (*m-e*) selected features (poor)
6. *ngh*: initial size of features set which includes features and its neighborhood features and stopping criterion

Process

1. Initialize population with random features. (*n* features are placed randomly in the search space).
2. Evaluate fitness of the population. Fitness calculation for features obtained from average of two ranking measures *Chi-Square and Gain Ratio.*
3. While (stopping criterion not meet). While no more new ranking for features. // forming new population.
4. Select features for neighborhood search. (Feature that have the highest fitness are chosen as "selected" and features from same type subset are chosen for neighborhood search (after complete the features from same type subset algorithm will begin with the other feature type subset)).
5. Weighted selected feature (more weights for features in best *e* features) and evaluate fitness.
6. Select the fittest feature from each feature set. (For each feature set, only the feature with the highest fitness will be selected to form the next feature population).
7. Assign remaining features to search randomly and evaluate their fitness.
8. End While.

End Process.

3. 4. Classifier Constructing

Always IDS have database either has all signatures of known attack which support the misuse intrusion detection or has all the normal behavior which support the anomaly intrusion detection. The proposal support IDS with database has both normal and attacks in all its variations to decide if that attack or not, if it was attack then it determines its type.

The research record detecting intrusions using most of strong data mining algorithms used in last year: Decision Tree (DT) ID3, Naïve Bayesian (NB), Neural Network (ANN), Support Vector Machine (SVM). These learning algorithms implemented in WEKA environment to evaluate the optimization of updated KDD and proposed feature selection.

4. Discussion and Experimental Work

The number of features increased to be 44 features and types of connection increased to be 6 general classes. Now will display the number of training and testing examples, as depended in the updated dataset, see Table 2.

The Proposed Feature Ranking is to use an intelligent approach (bee's algorithm) which is differing from traditional approach where the best subsets are chosen upon iterative evaluation experiment. This approach is supported with measures that calculate the correlation to quantify each with class (normal traffic or intrusion traffic (all the 6 classes)). So, the feature will has a rank represent the feature importance in intrusion detection, three ranked features subsets were involved, these are 44 features set, 22 features subset and 11 features subset. In order to evaluate the performance of updated dataset and proposed bee's algorithm feature selection for network intrusion detection. Ideally, IDS should have an attack Detection Rate (DR) of 100% along with False Positive (FP) of 0%. Nevertheless, in practice this is really hard to achieve. The most important parameters involved in the performance estimation of IDS are shown in Table 3.

The results obtained from constructing the four classifiers (ID3), (NB), (NN), and (SVM), on the updated KDD 2000 dataset and proposed bee's feature selection are very consistence and convergence with results in previous works [5, 6, 7, 8, 9, and 10]. The results in the following Tables (4, 5, and 6) present DR and FP measures with each classifier relating to the six classes. Each of these tables show the results of classifiers applied on updated dataset but each one consider case of the three features subsets cases.

5. Conclusions

From results obtained in implementing the enhanced NIDS reached to the following conclusions:

1. Updating KDD by a proposed created dataset to has new injected sessions, make it reliable and novel since it will contain most modern attacks not appear in KDD2000.
2. Because of injection there is three features added to be 44 features. This makes dataset suffer from missing values. But by applying preprocessing to dataset make the constructed classifier dependable and truth.
3. Optimizing no. of features to consider the critical feature will make the classifier constructing optimized in time and space. Also make the classifier work more speed as real-time system, since no. of features will be checked much less than original numbers of all features.

From Tables (3, 4, and 5), the results obtained are more consistence with previous related work and enhanced, especially with classifiers in 11 top features. This ensures the validity of updating KDD and using bee's algorithm as dependable intelligent ranking.

Table 1. Proposed Dataset Update KDD

Session ID	*Traditional features (41)*	*Added (3) Features*
1 *.* *.* *.*	*Traditional KDD2000 sessions*	*Filled with proposed encoded values*
. *.*	*Injected Sessions for modern intrusion* *Already have (44) features*	

Table 2. Number of examples for training and testing

Connection Types	Training examples	Testing examples
Normal	65,000	15,000
Denial of Services	85,000	35,000
Remote to User	73,000	15,000
User to Root	27,000	5,000
Probing	40,000	20,000
Extended	10,000	10,000
No. of Examples	300,000	100,000

Table 3. IDS parameters and their meaning

Parameters	Meaning
True Positives (TP) – Detection Rate (DR)	Attacks occur and alarm raised
False Positives (FP)	No attack but alarm raised

Table 4. Comparison of the results using 44 features

Method	Normal	DOS	U2R	R2L	Probe	Extended
ID3 (DR%)	99.70	99.76	99.25	99.27	99.30	99.12
ID3 (FP%)	0.08	0.04	0.11	6.81	0.40	5.83
NB (DR%)	99.25	99.69	72.25	99.11	99.13	99.05
NB (FP%)	0.06	0.04	0.14	8.02	0.45	6.83
NN (DR%)	99.30	99.50	85.04	99.01	99.09	89.17
NN (FP%)	0.07	0.03	0.50	9.81	0.60	4.83
SVM (DR%)	99.80	99.50	99.30	99.48	99.66	99.76
SVM (FP%)	0.09	0.05	0.18	8.81	0.45	7.83

Table 5. Comparison of the results using 22 features

Method	Normal	DOS	U2R	R2L	Probe	Extended
ID3 (DR%)	99.95	99.96	99.97	99.97	99.98	99.96
ID3 (FP%)	0.03	0.02	0.05	4.81	0.40	4.83
NB (DR%)	99.95	99.96	99.97	99.97	99.98	99.96
NB (FP%)	0.02	0.01	0.03	3.08	0.29	3.67
NN (DR%)	99.95	99.96	99.97	99.97	99.98	99.96
NN (FP%)	0.04	0.02	0.12	5.34	0.23	3.51
SVM (DR%)	99.95	99.96	99.97	99.97	99.98	99.96
SVM (FP%)	0.01	0.04	0.03	3.23	0.09	3.28

Table 6. Comparison of the results using 11 features

Method	Normal	DOS	U2R	R2L	Probe	Extended
ID3 (DR%)	99.97	99.97	99.97	99.97	99.98	99.98
ID3 (FP%)	0.03	0.02	0.05	4.81	0.40	4.83
NB (DR%)	99.98	99.99	99.99	99.99	99.98	99.98
NB (FP%)	0.02	0.01	0.03	3.08	0.29	3.67
NN (DR%)	99.96	99.96	99.97	99.97	99.98	99.98
NN (FP%)	0.04	0.02	0.12	5.34	0.23	3.51
SVM (DR%)	99.97	99.98	99.98	99.97	99.98	99.98
SVM (FP%)	0.01	0.04	0.03	3.23	0.09	3.28

References

1. Vaarandi R., “Real-Time Classification of IDS Alerts with Data Mining Techniques,” in Proc. of 2009 MILCOM Conference.
2. Pietraszek T., “Using Adaptive Alert Classification to Reduce False Positives in Intrusion Detection,” in Proc. of 2004 RAID Symposium, pp. 102-124.
3. Pham DT, Ghanbarzadeh A, Koc E, Otri S, Rahim S and Zaidi M. “The Bees Algorithm. Technical Note”, Manufacturing Engineering Centre, Cardiff University, UK, 2005.

4. Karaboga. D, “An idea based on honey bee swarm for numerical optimization. Technical Report TR06, Erciyes University, Engineering Faculty, Computer Engineering Department, 2005.
5. Pietraszek T., “Using Adaptive Alert Classification to Reduce False Positive in Intrusion Detection”, in proc. of 2004 RAID Symposium, PP. 102-124.
6. Viinkka J., Debar H., Me’ L., et al, “Processing Intrusion Detection Alert Aggregates with Time Series Modeling”, Information Fusion Journal, Vol. 10(4), 2009, PP. 312-324.
7. Vaarandi R., “Real-Time Classification of IDS Alerts with Data Mining Techniques”, in Proc. of 2009 MILCOM Conference, 7 pp.
8. Eunhye Kim, et al, “Feature Construction Scheme for Efficient Intrusion Detection System”, Journal of Information science and Engineering 26, 527-547 (2010).
9. Dewan Md. Farid , et al, “Combining nave Bayes and Decision Tree for Adaptive Intrusion Detection”, International Journal of Networking Security and it is Applications (IJNSA), Vol. 2, No. 2, April 2010.
10. Lee W., et al,” A data Mining Framework for Building Intrusion Detection Models”, Proceeding of IEEE Symposium on Security and Privacy, 1999, pp 120-132.