Journal of AL-Qadisiyah for computer science and mathematics
Vol.9 No.1 Year 2017

Anomaly Detection by Using Hybrid Method

Mohamed H. Ghaleb
Colge of Computer and information ticnology , University of Wassit
E-mail: Mghaleb@uowasit.edu.iq

Recived : 27\12\2016 Revised : 9\4\2017 Accepted : 13\4\2017

Abstract:

In this paper a new approach has been designed for Intrusion Detection System (IDS). The detection will be for misuse and anomalies for training and testing data detecting the normal users or attacks users.

The method used in this research is a hybrid method from supervised learning and text recognition field for (IDS). Random Forest algorithm used as a supervised learning method to choose the features and k-Nearest Neighbours is a text recognition algorithm used to detect and classify of the legitimate and illegitimate attack types.

The experimental results have shown that the most accurate results is that obtained by using the proposed method and proved that the proposed method can classify the unknown attacks. The results obtained by using benchmark dataset which are: KDD Cup 1999 dataset.

Keywords: IDS, Random Forest (RF), k-Nearest Neighbour (kNN)

1. INTRODUCTION

Generally, the major focus of the network attacks is to increase the threat against the commercial business and our daily life, so that, it become a serious problem for the researchers to find a suitable solution for these types of attacks.

The main objective of using the Intrusion Detection System (IDS) in the network is to monitor the behaviour of the network by determining whether is the incoming requests is an attack threats the network or not.

The (IDS) alerts the network or system administrator about suspicious activity when it happens and also monitors the networks traffic about those attacks.

The system objectives are covering the privacy, integrity, and availability of critical network information system [1-3].

Two fundamental methodologies have created by specialists for interruption recognition: abuse and Anomaly Intrusion Detection. The primary approach spoke to the particular examples of interruptions that endeavor known framework vulnerabilities.

On the inverse perspective, inconsistency recognition accept that each movement is meddling range unit basically strange. This proposes on the off chance that we have a tendency to may set up a conventional movement profile for a framework; we could, in principle, hail all framework states variable from the built up profile as interruption tries.

These two assortments of frameworks have their own upsides and downsides [4].

The previous will discover surely understood assaults with a truly high precision through example coordinating on understood marks, however can't discover novel assaults as an aftereffect of their marks aren't by the by realistic for example coordinating.

The last will discover novel assaults however for the most part for some such existing frameworks; have a high cautioning rate as a consequence of it's difficult to think of sensible customary conduct profiles for secured frameworks [5,6].

We have built a model not only limited to reducing feature of a rapid and significant to increase detection of known and unknown attack detection accuracy.

We tend in our experiments to use the information that arises from the MIT Lincoln Laboratory. A reference data set, developed for the bureau's intrusion detection system assessments also examine the attack in four varieties, probe ,denial of service, root to native and user to root, distinguish with traditional.

The rest of the research is planned as follows. Section 2 is specialized to present the related works for proposed model by using Algorithms of machine learning. Section 3 displays the datasets that we have used in this research. In section 4 we described the machine learning algorithms.

The algorithms of that machine learning which presented in Section 2 have been used in our proposed system, while Section 5 describes the experimental results got by using WEKA tool [7]. As for section 6 is specialized to present IDS description.

Section 7 displays the details of the proposed system at last section 8 presents conclusion for this paper.

2. RELATED WORK

An IDDM (Intrusion Detection utilizing Data Mining Techniques) [8] is an ongoing NIDS for abuse and oddity discovery.

It connected affiliation rules, Meta guidelines, and trademark rules. Jiong Zhang and Mohammad Zulkernine [9] utilize irregular woods for interruption location framework.

Irregular woods calculation is more precise and proficient on huge dataset like system movement. We likewise utilize this information mining system to choose components and handle imbalanced interruption issue.

The most related work to our own is done likewise by them [10]. They utilize Random Forests Algorithm over lead based NIDSs. In this way, novel assaults can be identified in this system interruption identification framework.

Rather than the already proposed information mining based IDSs, we utilize arbitrary backwoods for inconsistency interruption location. Arbitrary timberlands calculation is more exact and productive on extensive dataset like system movement.

We likewise utilize the information mining methods to choose elements and handle imbalanced interruption issue [11].

Irregular Forest (RDF) likewise mean to treat new occasions that are not mentioned in all current techniques of machine learning [12],

And k-Nearest Neighbor (k-NN) calculation is one of those calculations which are extremely easy to see yet works staggeringly well practically speaking.

k-NN technique was utilized as a supporter strategy for multi-class characterization [13][14].

3. DATASETS DESCRIPTION

A Since 1999, KDD'99 [15] has been the most broadly utilized information set for the assessment of abnormality recognition strategies. This information set is constructed in light of the information caught in "DARPA'98" IDS assessment program [16].

"DARPA'98" is around four gigabytes of packed crude (paired) tcpdump information of seven weeks of system movement. The 14th days of test information have around two million association records.

KDD preparing dataset comprises of roughly 4,900,000 single association vectors which contains 41 highlights and is named as typical or an assault, with precisely one particular assault sort. The reproduced assaults will be in one of the accompanying four classes:

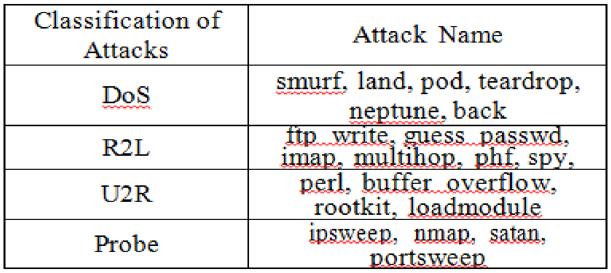
Denial of Service Attack (DoS) [17]: is an assault where the aggressor creates some figuring of memory asset excessively occupied or too full, making it impossible to treat honest to goodness asks for, or denies genuine clients access to a machine.

(1) User to Root Attack (UtoR) [18]: is a class of endeavor in which the assailant begins with access to an ordinary client account on the framework (maybe picked up by sniffing passwords, a word reference assault, or social designing) and can misuse some defenselessness to pick up access root to the framework.

(2) Remote to Local Attack (RtoL) [19]: This occurs when an attacker can send packets to a machine on a system, but those who do not have a record on this machine can cause some defenselessness to increase nearby access as a client of this machine.

(3) Probing Attack [20]: is an endeavor to accumulate data about a system of PCs for the evident motivation behind bypassing its security controls.

Table (1) demonstrated the four classes and their comparing assaults on every classification.

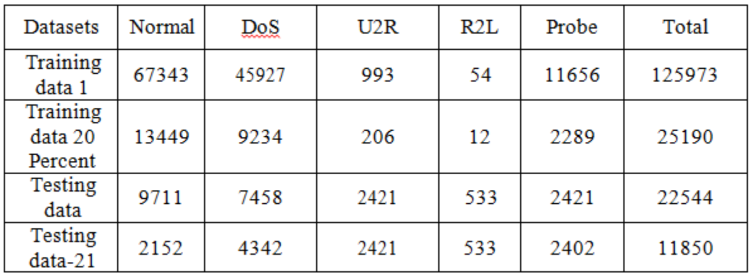


Classification of Attacks	Attack Name
DoS	smurf, land, pod, teardrop, neptune, back
R2L	ftp write, guess passwd, imap, multihop, phf, spy,
U2R	perl, buffer overflow, rootkit, loadmodule
Probe	ipsweep, nmap, satan, portsweep

TABLE (1): Attacks Classification in KDD Dataset

Note that the test information is not from an indistinguishable likelihood circulation from the preparation information, and it incorporates particular assault sorts not in the preparation information which make the undertaking more sensible.

Some interruption specialists believe that most of the novel attacks are differences in known assaults and the mark of known attacks can be sufficient to capture novel differences.



Datasets	Normal	DoS	U2R	R2L	Probe	Total
Training data 1	67343	45927	993	54	11656	125973
Training data 20 Percent	13449	9234	206	12	2289	25190
Testing data	9711	7458	2421	533	2421	22544
Testing data-21	2152	4342	2421	533	2402	11850

TABLE (2): Number of Attack Type On Connections Of KDD Datasets

4. MACHINE LEARNING ALGORITHMS USED IN THE PROPOSED FRAMEWORK

To conquer the impediments of the control based frameworks, various IDSs utilize information mining procedures. Information mining is the examination of (regularly expansive) observational information sets to discover examples or models that are both justifiable and valuable to the information proprietor [21].

Information mining can effectively separate examples of interruptions for abuse location, set up profiles of typical system exercises for peculiarity discovery, and manufacture classifiers to recognize assaults, particularly for the unlimited measure of review information. Information mining-based frameworks are more adaptable and deployable [22].

In the course of recent years, a developing number of research ventures have connected information mining to interruption recognition with various calculations. We propose a way to deal with utilize arbitrary woods and k-Nearest Neighbor in interruption location.

For example, those had been connected to expectation, likelihood estimation, and example investigation in mixed media data recovery and bioinformatics. Generally, to the best of our insight, Random Forests calculation has not been totally connected to recognize novel assaults (obscure assaults) in programmed interruption identification.

Luckily, we can take preferences from k-NN that can arrange in more absolutely and an imperative example perceiving strategy in light of agent points [23].

A. Random Forests (RDF)

The Random Forests [24] is a group of unpruned characterization or relapse trees. Irregular timberland produces numerous order trees. Every tree is developed by an alternate bootstrap test from the first information utilizing a tree order calculation.

After the woods are framed, another protest that should be characterized is put down each of the tree in the woodland for order. Every tree gives a vote that shows the choice of tree of the question class. The timberland picks the class with the most votes in favor of the question.

The principle components of the irregular backwoods calculation are recorded as takes after:

- It runs proficiently on substantial information sets with many components.
- It can give the evaluations of what components are essential.
- It has no ostensible information issue and does not over-fit.
- It can deal with lopsided information sets.

B. k-NN: k-Nearest Neighbor

k-NN order is a straightforward and simple to actualize characterization technique [25]. In spite of its straightforwardness, it can perform well by and large. k-NN is especially appropriate for multi-modular classes and in addition applications in which a protest can have many class names.

For instance, for the task of capacities to qualities in light of expression profiles, a few analysts found that k-NN outflanked SVM, which is a considerably more modern grouping scheme [25].

The 1-Nearest Neighbor (1NN) classifier is a vital example perceiving strategy in view of agent focuses [26]. In the 1NN calculation, entire prepare tests are taken as agent focuses and the separations from the test tests to every illustrative point are processed.

The test tests have a similar class mark as the agent direct closest toward them.

The k-NN is an augmentation of 1NN that decides tests through finding the k closest neighbors.

C. Feature Selection

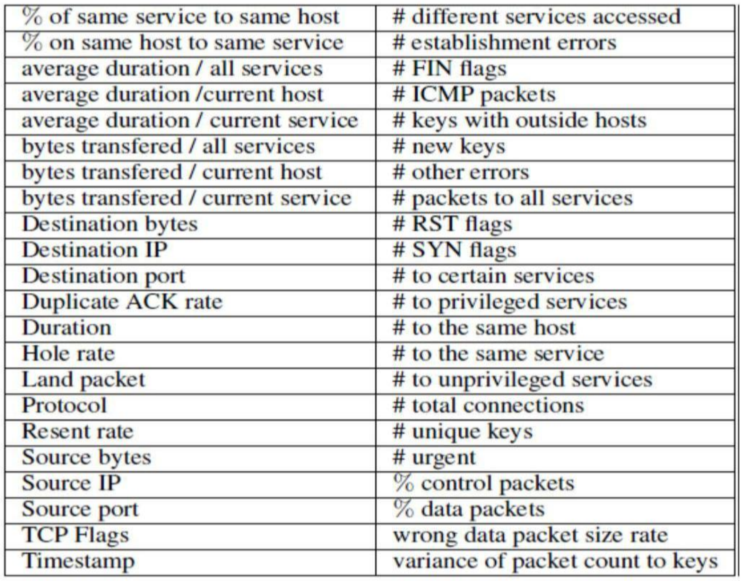
In complex arrangement spaces, a few information's may frustrate the characterization procedure. Components may contain false relationships, which obstruct the way toward distinguishing interruptions.

Further, a few components might be repetitive since the data they include is contained in different elements [27]. Additional components can build calculation time, and can affect the exactness of IDS. Highlight choice enhances order via hunting down the subset of elements, which best characterizes the preparation information [28].

The elements under thought rely on upon the kind of IDS, for instance, organize based IDS will dissect arrange related data, for example, bundle goal IP address, signed in time of a client, sort of convention, span of association and so forth. It is not known which of these components are excess or immaterial for IDS and which ones are significant or fundamental for IDS.

There does not existing any model or capacity that catches the relationship between various elements or between the distinctive assaults and components. In the event that such a model existed, the interruption identification process would be basic and direct. In this paper we utilize information digging strategies for highlight determination.

As shown in Table 3 the features that are extracted from the dataset to be applied in the process of the data mining process.



% of same service to same host	# different services accessed
% on same host to same service	# establishment errors
average duration / all services	# FIN flags
average duration /current host	# ICMP packets
average duration / current service	# keys with outside hosts
bytes transfered / all services	# new keys
bytes transfered / current host	# other errors
bytes transfered / current service	# packets to all services
Destination bytes	# RST flags
Destination IP	# SYN flags
Destination port	# to certain services
Duplicate ACK rate	# to privileged services
Duration	# to the same host
Hole rate	# to the same service
Land packet	# to unprivileged services
Protocol	# total connections
Resent rate	# unique keys
Source bytes	# urgent
Source IP	% control packets
Source port	% data packets
TCP Flags	wrong data packet size rate
Timestamp	variance of packet count to keys

TABLE (3): Extracted Features For Applying Data Mining To The IDSs

5. Intrusion Detection System (IDS)

In Figure 1 shows, the methods employed in the system have been described, and explain how to apply these methods to detect new types of attacks with true positive rate, false positive rate for detection of network intrusion.

Two phases are recognized in this system is to process of identifying the abnormal and normal instances [29]. The first phase is the training phase; the function of the phase is to reduce the features that are not related. Next phase is detection phase.

Since normal operations specific and show expected behavior, we can use knowledge-based misuse of IDS, while unexpected activity (assuming that sneaking would be unusual) is designed continuously and progress Cannot be seen as a knowledge based attack, therefore the novel attacks is performed by using anomaly IDS detection [30].

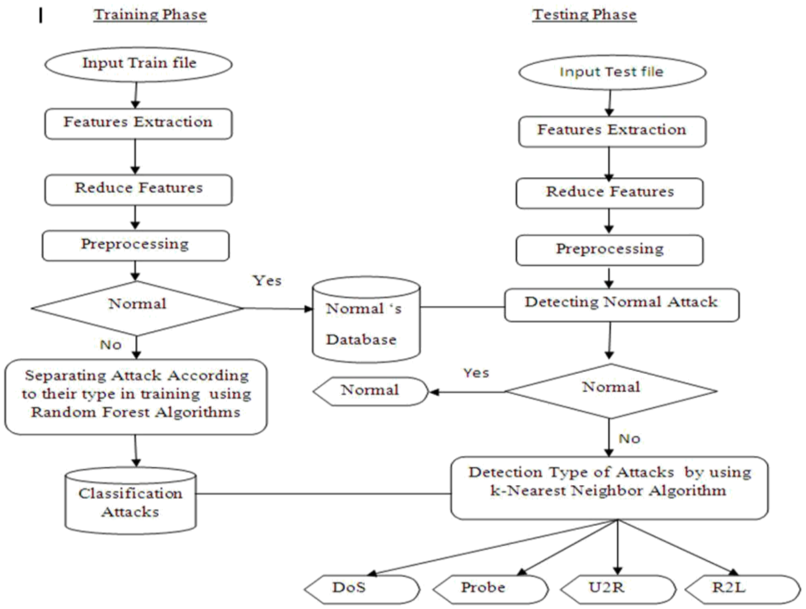


Figure (1): The General Flow of ID Systems

The experimental results are reported through the KDD'99 datasets. And they showed that the proposed system gives better performance than results from the KDD'99 contest.

6. Proposed Model

The main objective from this work is to propose another model for more accurate and recognition rate as shown in Figure 2 using the knowledge of the data flow used as a part of WEKA algorithm-based software.

In this proposed framework, as taking everything into account, will utilized it as a part of the separating procedure of preprocessing state and it will build the trees furthermore select the irregular elements. Subsequent to preprocessing state, we will utilize the k-NN calculation, design acknowledgment strategy for order state to distinguish the approaching assaults.

The outcomes with content that express the Ture Positive, False Positive Rate, Precision, Recall furthermore perplexity lattice we can extricate.

7. Experimental Results

The experimental results detect unknown attacks for intrusion detection by the KDD'99 datasets.

Experimental results are classified in terms of the classes which obtained better level of discrimination from others in the training set [31][32].

Random Forest algorithm in the proposed system reduced some features in dataset at each connection. By using corrected KDD dataset the system will try to detect various anomaly attacks.

The training time will be reduced through the proposed system and will be increased the accuracy of the system's classification.

WEKA tool will be used to obtain experimental results.

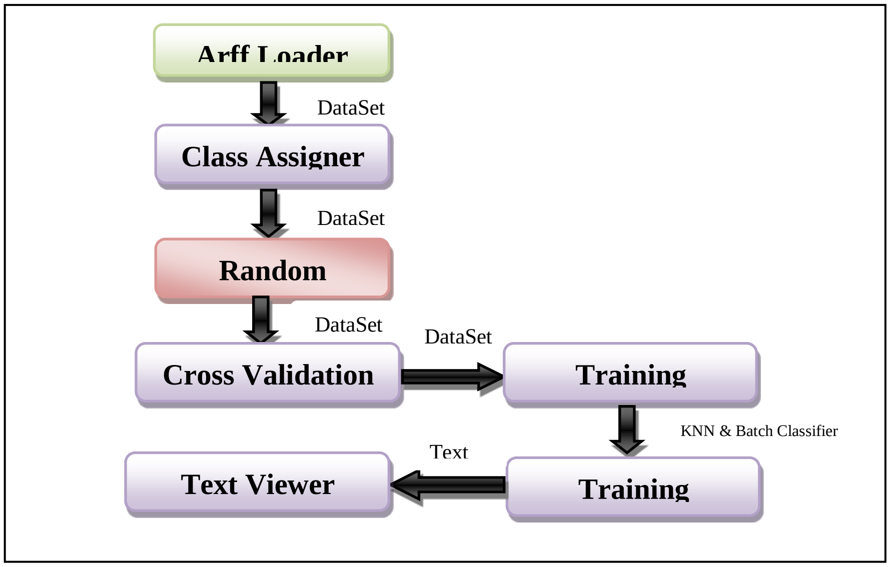


Figure (2): The proposed Model

For classification the system uses the reduced features (default 6 in WEKA) and 10 trees in the experiments process.

Figure (3) shows how the accuracy of the proposed system is better than Random Forest, K-Nearest Neihbor and Naïve Bayes.

Since the test datasets "Testing data" and "Testing data-21" have with different statistical distributions than either "Training data" or "Training data_20 Percent", the accuracy decrease Cross Validation results with those training files, but to detect the unknown attack, the results in the test file containing the most unknown attack types (novel attacks) from other data sets get more detection random forest rate can be compared with other methods, such as Shows in Figure 3.

The proposed model can be used in more accurate attack detection according to these results of Figure (3).

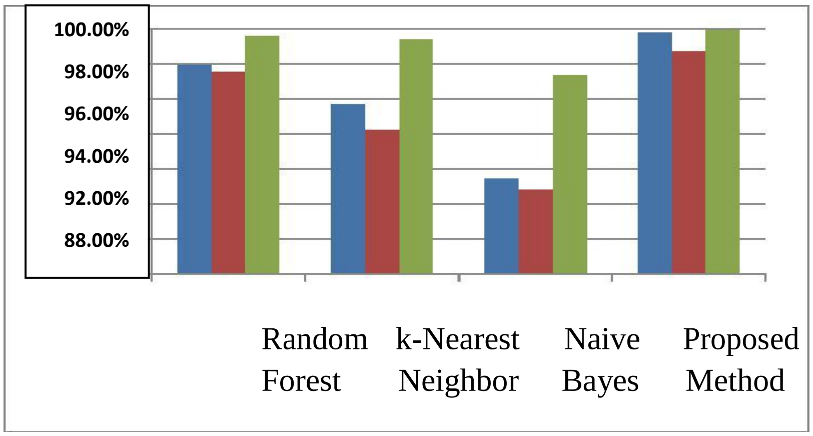


Figure (3): The comparison of accuracy between the Proposed Method and (Random Forest, k-NN & Naive Bayes)

8. Conclusion and Future Direction

Late explores utilized choice trees, counterfeit neural systems and a probabilistic classifier and reported, as far as location and false caution rates, however it was still false positives and insignificant alarms in identification of novel assaults.

This research contained a review of the different information mining strategies that have been proposed to upgrade of abnormality interruption location frameworks. Also, we connected the characterization strategies for ordering the assaults (interruptions) on DARPA dataset.

The outcomes demonstrating the Random Forest execution is superior to different classifiers. In any case, Random Forest takes more time than different classifiers. Then again, k-Nearest Neighbor is likewise the great demonstrating calculation in our tests.

Subsequently, we can amplify this trial by consolidating those two calculations; the framework may hope to get the more exact and recognition rate to distinguished interruption.

Arbitrary Forest will handle the separating stage while the k-NN will be used as a classifier.

REFERENCES

[1] W. Lee and S. J. Stolfo, “Data Mining Approaches for Intrusion Detection”, the 7th USENIX Security Symposium, San Antonio, TX, January 1998.

[2] K.T.Khaing and T.T.Naing, “Enhanced Feature Ranking and Selection using Recurisive Featue Elemination and k-Nearest Neighbor Algorithms in SVM for IDS”, Internaiton Journal of Network and Mobile Technology(IJNMT), No.1, Vol 1. 2010.

[3] M. Bahrololum, E. Salahi and M. Khaleghi, "Anomaly Intrusion Detection Design using Hybrid of Unsupervised and Supervised Neural Network", International Journal of Computer Network & Communications(IJCNC), Vol.1, No.2, July 2009.

[4] L. Breiman, “Random Forests”, Machine Learning 45(1):5–32, 2001.

[5] V. Marinova-Boncheva, "A Short Survey of Intrusion Detection System" , 2007.

[6] Tamas Abraham, “IDDM: Intrusion Detection Using Data Mining Techniques”, DSTO Electronics and Surveillance Research Laboratory, Salisbury, Australia, May 2001.

[7] WEKA software, Machine Learning, http://www.cs.waikato.ac.nz/ml/weka/, The University of Waikato, Hamilton, New Zealand.

[8] KDD’99 datasets, The UCI KDD Archive, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html , Irvine, CA, USA, 1999.

[9] KDD Cup 1999. Available on:http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, December 2009.

[10] Lan Guo, Yan Ma, Bojan Cukic, and Harshinder Singh, “Robust Prediction of Fault-Proneness by Random Forests”, Proceedings of the 15th International Symposium on Software Reliability Engineering (ISSRE'04), pp. 417-428, Brittany, France, November 2004.

[11] Ting-Fan Wu, Chih-Jen Lin, and Ruby C. Weng, “Probability Estimates for Multi-class Classification by Pairwise Coupling”, The Journal of Machine Learning Research, Volume 5, December 2004.

[12] Yimin Wu, High-dimensional Pattern Analysis in Multimedia Information Retrieval and Bioinformatics, Doctoral Thesis, State University of New York, January 2004.

[13] Bogdan E. Popescu, and Jerome H. Friedman, Ensemble Learning for Prediction, Doctoral Thesis, Stanford University, January 2004.

[14] Eleazar Eskin, Andrew Arnold, Michael Prerau, Leonid Portnoy, and Salvatore Stolfo. “A Geometric Framework for Unsupervised Anomaly Detection: Detecting Intrusions in Unlabeled Data.” Applications of Data Mining in Computer Security, 2002.

[15] M. Mahoney and P. Chan, “An Analysis of the 1999 DARPA/Lincoln Laboratory Evaluation Data for Network Anomaly Detection”, Proceeding of Recent Advances in Intrusion Detection (RAID)-2003, Pittsburgh, USA, September 2003.

[16] Leo Breiman and Adele Cutler, Random forests, http://statwww.berkeley.edu/users/breiman/RandomForests/cc_home.h tm, University of California, Berkeley, CA, USA.

[17] David J. Hand, Heikki Mannila, and Padhraic Smyth, Principles of Data Mining, The MIT Press, August, 2001.

[18] MIT Lincoln Laboratory, DARPA Intrusion Detection Evaluation, http://www.ll.mit.edu/IST/ideval/,MA, USA.

[19] J.Zhange and M. Zulkerline, “Network Intrusion Detection using Random Forests”,2011.

[20] T. Lappas and K. Pelechrinis Data Mining Techniques for (Network) Intrusion Detection Systems”.

[21] J. Zhang and M. Zulkernine, ”Anomaly Based Network Intrusion Detection with Unsupervised Outlier Detection”, Symposium on Network Security and Information Assurance Proc. of the IEEE International Conference on Communications (ICC), 6 pages, Istanbul, Turkey, June 2006.

[22] S. Thirumuruganathan , “A Detailed Introduction to K-Nearest Neighbor (KNN) Algorithm”, World Press, May 17, 2010.

[23] X Wu, V Kumar, J Ross Quinlan, J Ghosh, “Top 10 Data mining Algorithm”, Knowledge and Information Systems, Volume 14, Issue 1, pp 1-37 ,2008 – Springer

[24] S. Mukkamala, A.H. Hung and A. Abraham, “Intrusion Detection Using an Ensemble of Intelligent Paradigms.” Journal of Network and Computer Applications, Vol. 28(2005), 167-182.

[25] S. Chebrolu, A. Abraham, and J.P. Thomas, “Feature Deduction and Ensemble Design of Intrusion Detection Systems.” International Journal of Computers and Security, Vol 24, Issue 4,(June 2005), 295-307

[26] A.H. Sung and S. Mukkamala, “The Feature Selection and Intrusion Detection Problems.” Proceedings of Advances in Computer Science - ASIAN 2004: Higher- Level Decision Making. 9th Asian Computing Science Conference. Vol. 321(2004) , 468-482.

[27] Kim, G., Lee, S., & Kim, S. (2014). A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. Expert Systems with Applications, 41(4), 1690-1700.

[28] S. Mukkamala, A.H. Sung and A. Abraham, “Modeling Intrusion Detection Systems Using Linear Genetic Programming Approach.” LNCS 3029, Springer Hiedelberg, 2004, pp. 633-642.

[29] Tsai, C. F., Hsu, Y. F., Lin, C. Y., & Lin, W. Y. (2009). Intrusion detection by machine learning: A review. Expert Systems with Applications, 36(10), 11994-12000.

[30] A. Abraham and R. Jain, “Soft Computing Models for Network Intrusion Detection Systems.” Soft Computing in Knowledge Discovery: Methods and Applications, Springer Chap 16, 2004, 20pp.

[31] Tsai, C. F., & Lin, C. Y. (2010). A triangle area based nearest neighbors approach to intrusion detection. Pattern recognition, 43(1), 222-229.

[32] A. Abraham, C. Grosan, and C.M. Vide, “Evolution Design of Intrusion Detection Programs.” International Journal of Network Security, Vol. 4, No. 3, 2007, pp. 328-339.

كشف المتطفلين باستخدام طريقة هجينة

محمد حسين غالب عبد الخالق
جامعة واسط
كلية علوم الحاسوب وتكنلوجيا المعلومات
mghaleb@uowasit.edu.iq

المستخلص :
في هذا البحث تم تصميم طريقة جديدة في انظمة الكشف عن الدخلاء (المتطفلين) للشبكة الحاسوبية الالكترونية، عملية الكشف كانت لسيئي الاستخدام للشبكة من خلال استخدام بيانات تجريبية وتدريبية صنفت عالميا للتمييز بين المستخدمين الاعتياديين والمستخدمين اللذين يهاجمون الشبكة. الطريقة المستخدمة في هذا البحث هي طريقة هجينة بين خوارزمية التمييز العشوائي (supervised learning random forest) والتي استخدمت في تحديد الخصائص المهمة في الكشف عن المستخدمين السيئين وخوارزمية (K-nearest Neighbours) والتي استخدمت لعملية الكشف والتصنيف لانواع الهجومات المعروفة والغير معروفة. اضهرت النتائج ان الطريقة المقترحة اعطت دقة عالية في التصنيف واثبتت بان لها فعالية في تصنيف الهجومات الغير معروفة وان العينات المتقدمة كانت عينات عالمية من شركة (KDD Cap 1999) والتي تحتوي على انواع مختلفة من الهجومات .