Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 10, Issue, 04, pp.68238-68242, April, 2018

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

ISSN: 0975-833X

RESEARCH ARTICLE

APPLYING CLASSIFICATION TECHNIQUES FOR NETWORK INTRUSION DETECTION

***Nancy Awadallah Awad**

Sadat Academy for Management Sciences, Department of Computer and Information Systems, Egypt

ARTICLE INFO

Article History:
Received 10th January, 2018
Received in revised form
15th February, 2018
Accepted 21st March, 2018
Published online 30th April, 2018

Key words:
Classification, Intrusion detection, Machine Learning Algorithms, Decision Tree, K-star NSL-KDD, Decision Table.

ABSTRACT

Organizations are becoming increasingly vulnerable to potential cyber threats which defined as network intrusions. Intrusion Detection is basically providing the security or managing the flow of data, information, managing the access of the system to only authorized user. In adaptive false alarms filter a combination of machine learning algorithms is used to increase the classification accuracy of the system. The experimental results in this research show that the proposed J48, Decision Table and k-star techniques reducing the false alarm rate and improving the accuracy. The new NSL-KDD dataset is used, which is applied with WEKA tool.

Citation: Nancy Awadallah Awad, 2018. "Applying classification techniques for network intrusion detection", *International Journal of Current Research*, 10, (04), 68238-68242.

INTRODUCTION

Network crimes have been expanded quickly accordingly to the improvement of network technologies. The important role of network security is to protect the data from any intruder as these attacks become intrusions form threats for network. But the question that arises is why data mining is applicable to intrusion detection; this is because several causes such as intrusive activities and normal leave evidence in audit data, and intrusion detection is a data analysis process which due that fraud detection, fault/alarm management (data mining applications) are considered successful applications in related domains. In another level, large volumes of network data need to be mined in different techniques such as clustering, classification data mining techniques. Intrusion detection systems (IDSs) are one of the key areas of application of data mining techniques which are used to detect malicious attacks and identify any violation for the policy security of a network. According to the detection mechanism,

There are two classification types of IDS

Misuse and anomaly based IDS (Surabi, 2014). But there are limitations found of traditional signature-based methods, which are manual update of signature database and inability to detect emerging cyber threats.

****Corresponding author:*** **Nancy Awadallah Awad,**
Sadat Academy for Management Sciences, Department of Computer and Information Systems, Egypt.

IDS are additionally essential supplement to preventive security components, for example, firewalls in light of the fact that IDS identify assaults that endeavor framework outline flaws or bugs and IDS give criminological confirmation to advise framework head's responses to digital assaults (Pan *et al*., 2015). K. Hwang reported another trial known as hybrid intrusion detection system (HIDS), which consolidates the upsides of the capacity of peculiarity of the anomaly detection system ability to detect obscure attacks and decrease false-positive rate of misuse IDS. This model recognizes irregularities past the capabilities of signature-based SNORT, a weighted signature conspire is produced to coordinate anomaly detection system with SNORT by removing signatures from anomalies detected (Hwang *et al*., 2007). Researcher in this paper utilize weka tool as an anomaly detector requires some type of machine learning algorithm such as Decision tree (J48), Decision Table, K-star classifiers.

Review of Related Literature

Various techniques that are used to detect the intrusions include data mining clustering, classification, neural network, and statistical methods. Clustering an unsupervised learning technique in which the data is assigned and labeled into groups of similar objects ,it can detect the intrusions from unlabeled data, and furthermore be utilized for both misuse and anomaly detection (Denatious and John, 2012; Chang-Tien *et al*., 2005). Most of the clustering techniques discussed previously utilize several steps to detect intrusions. Chang-Tien Lu *et al* ., said similarity-based and centroid-based clustering are the two

major clustering techniques used in cybersecurity especially in intrusion detection techniques (Chang-Tien *et al*., 2005). Not at all like traditional anomaly detection methods, they utilized an adjusted incremental k-means algorithm to group data occurrences that contain both normal behaviors and attacks (Moorthy and Sathiyabama, 2012). Heuristics is utilized after clustering to consequently label each cluster as either normal or attacks (Mohd Junedul Haque *et al*., 2012). Classification is similar to clustering but much less exploratory than it. Classification partition records into different classes. A Class can be defined as a set or collection of members having certain attributes in common such as botnet is one of the class of cyber attacks. Classifier objective is to classify new records should be, not to explore the data to discover interesting segments (Mohd Junedul Haque *et al*., 2012; Anchuri *et al*., 2011). Jingke Xi (Jingke, 2008) discussed different approaches for outlier detection in data mining perspective. Outlier detection algorithms are categorized into two types, classic outlier approach and spatial outlier approach. In classic outlier approach the outliers are analyzed depending on transaction dataset and in spatial outlier approach the outliers are analyzed based on spatial dataset. Kesavaraj G and Sukumaran S (Kesavaraj and Sukumaran, 2013) presented a study on various classification techniques such as decision tree induction methods, rule-based methods, Bayesian network, neural network and support vector machines and his study states that one of the classification techniques cannot be chosen best among all and they depend on the dataset choice. S. Devaraju et S. Ramakrishnan, proposed association rule mining algorithm (ARMA) for detecting various network.

They generated rules reduce the false positive rate (FPR) and improve the detection rate (Devaraju and Ramakrishnan, 2015). V. Mookerjee *et al*., developed an analytical model in which a variety of factors are considered such as false-positive rate and detection rate. They also discriminated between normal users and hackers that try to penetrate and compromise the firm's information assets (Mookerjee *et al*., 2011). G.V. Nadiammai et M. Hemalatha used several issues: Effectiveness of Distributed Denial of Service (DoS) Attack, data classification, and achieve high level of human interaction (Nadiammai and Hemalatha, 2014). T. F. Ghanem, W. S. Elkilani et H. M. Abdul-kader ,proposed a hybrid approach for anomaly detection using detectors generated based on genetic algorithms and multi-start meta-heuristic method (Ghanem *et al*., 2015). R. W.-w Hu Liang et R. Fei. , used the hierarchical clustering for intrusion detection system which evaluated on KDD Cup99 and accomplished 0.5 % false positive rate (Hu Liang and Fei, 2009). S. M. Sangve et R.a C. Thool , proposed a framework for anomaly network intrusion detection system implemented based on using genetic algorithm, meta-heuristic method and clustering techniques. They checked the framework execution according to detector time and false positive rate and (Sangve and Thool, 2015). M. Zolotukhin *et al*., proposed an algorithm which utilize SSL/TLS protocol for (DoS) detection attacks, the data of network connections is encrypted on the application layer (Zolotukhin *et al*., 2015).

Data Mining based IDS

The entire procedure of finding helpful information and patterns in data is defined as the knowledge discovery in Databases (KDD) terminology .This process aids in creation of a model for intrusion detection system. After the dataset is chosen, the pre-processing techniques are used to clean data (Dewa *et al*., 2016). There are several steps of knowledge discovery process, Firstly, data cleaning for removing any conflicting data. Secondly, data incorporation which consolidates various data sources. Thirdly, data selection where data pertinent to the examination errand are recovered from the database. Fourthly, data transformation where data are changed and solidified into frames proper for mining by pre-shaping synopsis or conglomeration operations. Fifthly, data mining, which is an essential procedure where shrewd systems are connected to separate information designs. Sixthly, pattern evaluation to recognize the right fascinating examples speaking to information in view of intriguing measures. And finally, knowledge presentation perception and information portrayal methods are utilized to display mined learning to users to users.

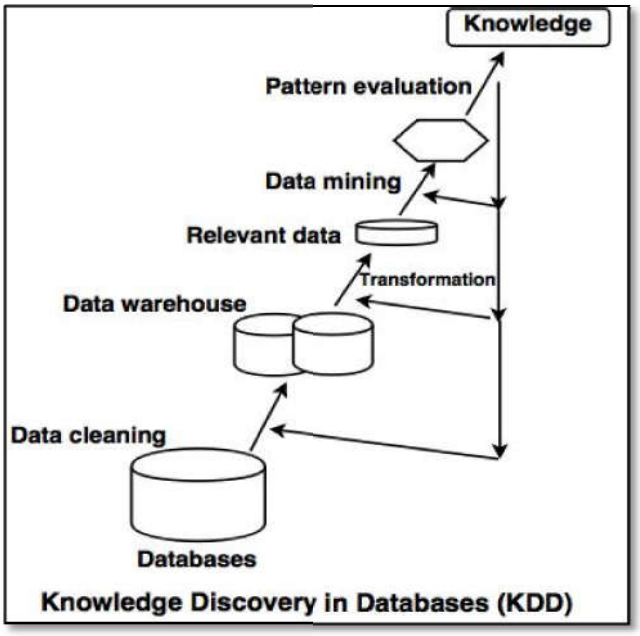

Figure 1. KDD process [20]

Table 1. Intrusion Examples and Description

Intrusion Example	Description
Probing	Attacker collects information about a host via using network services.
Denial-of-Service (DoS)	Attempt to deny the use of resources or services to the authorized users or shut down a computer, a network, or process, or.
Remote to Login (R2L)	Unauthorized access from a remote machine, as. guessing password attacks , gains access to the machine and does harmful operations.
Unauthorized access to local super user (U2R)	Attacker who has access to a local account on a computer system is able to elevate his or her privileges.
Trojan horses / Worms	Attacks that are aggressively replicating on other hosts
Compromises	Attackers use known vulnerabilities as buffer overflows and weak security to gain privileged access to hosts.
Address spoofing	Attacker uses a fake IP address to send malicious packets to a target.

The role of data mining is to utilize algorithms to extricate the information and patterns derived by the KDD from large sets of databases. For using the concept data mining to IDSs several advantages of (1) automatic generation for IDSs detection models, so which due to detect new attacks (2) building IDSs for a wide assortment of figuring conditions.

Data mining approaches consider interruption recognition as examination process, which incorporates four fundamental advances: (1) Capturing parcels exchanged on the system. (2) Extracting a broad arrangement of highlights that can portrays organize association or a host session. (3) Learning a model that can precisely depict the conduct of irregular and ordinary exercises by applying data mining activities. (4) Using the learnt models to distinguish the interruptions (Sànchez-Marrè, 2013). NSL-KDD training dataset is similar to KDD cup 99 and consists of 41 features for 4,900,000 single connections; it is labeled as attack or normal type.

Proposed System Design

The proposed system consists of three modules: Classification module, outlier detection module, and result module. The new NSL-KDD dataset is used and given as input to the classification module where decision tree (J48), K-star, decision table and JRIP classifiers are applied to the dataset. The input from the classification module is given to the outlier detection module where the outliers are detected, which is further classified by classification module and the result is displayed in the form of graph in result module. The overall system design is shown in Figure 2.

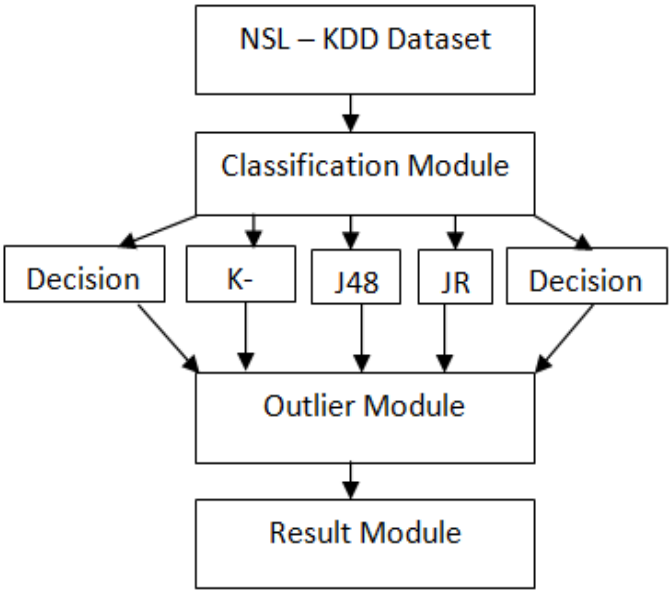

Figure 2. Proposed System

Decision Tree

Decision tree is a data mining technique that can be used as classification and regression tools. It consists of dividing the space of input variables (regressors, predictors, etc) into smaller regions, separating the data points according to the target values y. Hence, the model consists of recursive partition of the data and a simple model for each final region. The recursive partition part is represented as a tree that consists of nodes where each terminal node of the tree, called "leaf" represents a final region.

Starting at the root node of the tree that contains all data points; this node is decomposed in two daughter nodes according to a question asked about the values of regressors. If these daughter nodes are not converted into leaves, same process is applied and they are decomposed each in two nodes, and so on until all nodes are transformed to leaves. Each point moves down in the tree according to its regressors values, until it reaches a leaf.

The predicted value is then determined according to the simple model mentioned above. This model for classification is simply the majority class of the train samples that form that leaf. The difficult tasks of the tree construction is to find the best questions that can lead to an optimal tree (optimal partition), and to find the criteria upon which the node is considered "pure" and converted to a leaf. *Splitting Rules* Starting at the root node, and then repeating the same procedure at each node, the question to be asked is chosen in a way that maximizes the information about y which means minimizing the impurity of the node, when this question leads to two daughter nodes. A score measure is used assess the importance of the variable and their discriminative ability, and thus to be used to split the node.

This score is defined as:

$$Score(S,t) = I(t) - \sum_{i=1}^{2} \frac{Ni}{N} I(ti)$$

Where S is the split used to decompose the node t of size N in two daughter nodes $t1$ and $t2$ of size $N1$ and N respectively and $I(\ .\ )$ is the impurity measure of the node. The impurity measures commonly used are Entropy and Gini index, which in our case (binary classification) are defined as follows:

$$I\,\text{Entropy}\ (t) = -\left(\frac{N+}{N}\log\frac{N+}{N}\right) - \left(\frac{N-}{N}\log\frac{N-}{N}\right)$$

$$I\,\text{Gini}\ (t) = \left[\frac{N+}{N}\left(1-\frac{N+}{N}\right)\right] + \left[\frac{N-}{N}\left(1-\frac{N-}{N}\right)\right]$$

Where N+ and N- represents the number of each class (positive and negative) in the node t.

Stopping Criterion

The choice to stop the procedure mentioned previously on a specific node and converting it to a terminal node depends on several factors:

- I (T) = 0, which means that all the data points in the node are from the same class.
- $score\ (S,\ T) < threshold$, which means that the splitting is not beneficial enough.
- When reaching a maximal size of the tree.

Tree pruning

This construction will always lead to a complex tree that may over fit the train set, and perform poorly on the test set. We can solve this issue by pruning the obtained complex tree. We consider a complexity parameter CP that associates a penalty of having a complex tree. Hence, the quantity we aim to minimize to prune the tree (T) is:

$$RCP\ (T) = \frac{RT}{R(T\ 0)} + CP \times |T|$$

The mis-classification rate of the tree T is (T), and T is the first node of the tree, T is the number of leaves of the tree.

This quantity represents a trade-off between the complexity of the tree and its performance on the train data. J48 (Bhargava *et al*., 2013) , is, an implementation extension of the C4.5 and ID3 algorithm for decision trees, it is used for classification, available through Weka (Hall *et al*., 2009). It deals with several data types input, such as textual, nominal, and numeric. Decision tree algorithm

Split (node, {example}):

A⇓ the best attribute for splitting the {examples}

Decision attribute for this node ⇓ A

For each value of A, create new child node

Split training {examples} to child nodes

For each child node/subset

If subset is pure: STOP

Else: Split (node,{subset}) (Anand Chinchore *et al*., 2016).

K-star Algorithm

K-star is a classification algorithm that utilizes an entropy-based separation work. (John *et al*., 1995). The K-star calculation utilizes entropic measure, in view of likelihood of changing an occurrence into another by haphazardly picking between every single conceivable change. Utilizing entropy as a meter for an example remove is extremely valuable and data hypothesis helps in figuring the separation between the cases. The many-sided quality of a change of one example into another is really the separation between occasions. This is accomplished in two stages. Initially characterize a limited arrangement of changes that will outline example into another (Mahmood and Hussein, 2013).

Performance Measurement

The following evaluating metric terms have been used in this proposed architecture

False positive (FP) for incorrectly recognized, False Negative (FN) for incorrectly rejected, True negative (TN) for correctly rejected and True positive (TP) for correctly identified. An example disarray framework for two class case can be spoken to as appeared in Table 2

Table 2. Binary Matrix (Buczak, Anna, 2015)

Predicted Classes	
True Positive (TP)	False Negative (FN)
False Positive (FP)	True Negative (TN)

The previous metrics are defined:

$$\text{Precision} = \frac{\text{number of true positives}}{\text{total number of positive connections}}$$

Sensitivity (Recall) = *umber of true positives / total number of actual bad connections. Specificity = number of true negatives / total number of actual good connections* $TN / TN + FP$

we can state from the previous matrix, that:

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

The customary F-measure is the symphonious mean of precision and recall:

$$\text{F} - \text{measure} = \frac{2 * Precision * Recall}{Precision + Recall}$$

Table 3. Results of J48, K-star, Decision Table classification techniques

Technique	Parameters						
	Correctly Classified Incidents	Incorrectly Classified Incidents	True Positive Rate	False Positive Rate	Precision	Recall	F-Measure
J48 Without Selection attribute for Classification	98.16 %	1.83 %	0.982	0.054	0.981	0.982	0.982
K-Star Use training set	98.17 %	1.82 %	0.982	0.019	0.982	0.982	0.982
Decision Table Use training data set	99.29 %	0.71 %	0.993	0.007	0.993	0.993	0.993
J48 With attribute selection for classification Use training data KDD train	99.51 %	0.48 %	0.995	0.005	0.995	0.995	0.995
JRIP rules	99.7737 %	0.23 %	0.998	0.002	0.998	0.998	0.998

Decision Table and JRIP Classifiers

Decision Table structures a fundamental decision table larger part classifier. It assesses highlight subsets utilizing best initially search and can utilize cross approval for appraisal .JRIP executes a propositional rule learner called as "Repeated Incremental Pruning to Produce Error Reduction (RIPPER)" and utilizes consecutive covering algorithms for making requested rule lists. The algorithm goes through 4 stages: Growing a rule, Pruning, Optimization and Selection (Veeralakshmi and Ramyachitra).

RESULTS

In this research the new NSL-KDD dataset is utilized, which is an adjusted dataset for KDD Glass 1999 interruption discovery dataset. NSL-KDD is connected with WEKA tool, which comprises of several of machine learning algorithms for Data mining. Two classification techniques are used in this research, J48 and K-star. In J48 technique, researcher compare between using J48 without attribute selection and with using it after attribute selection for classification .K- star technique is used after attribute selection for classification. The experimental results show that the J48 with attribute selection for

classification (pre-processing), JRIP and decision table achieved the required accuracy in comparison using J48 without selection attribute, k-star and technique.

Conclusion

Naturally ,researchers in the network security need data mining techniques in order to achieve the accuracy required to detect intrusions on the network and reduce the rate of detection error .In this research, the J48 , K-Star and Decision table and JRIP classification techniques are utilized and its high accuracy has been implemented with weka tool. J48, JRIP and Decision table classification algorithms are effective methodologies for network intrusion detection.

REFERENCES

Anand Chinchore, Guandong Xu and Frank Jiang, 2016. "Classifying sybil in MSNs using C4.5", 2016 Internat ional Conference on Behavioral, Economic and Socio-cultural Computing (BESC),

Anchuri, P., Zaki, M., Barkol, O., Bergman, R., Felder, Y. and Golan, S. 2011. "Infrastructure Pattern Discovery in Configuration Management Databases via Large Sparse Graph Mining",

Bhargava, N., *et al*., 2013. " Decision Tree Analysis on J48 Algorithm for Data Mining", *International Journal*, 3(6).

Buczak, Anna and Guven, E. 2015. "A Survey of Data Mining and Machine Learning Methods f or Cyber Security Intrusion Detect ion", IEEE Communications Surveys & Tutorials,

Chang-Tien, Lu. and Arnold, P. 2005. Boedihardjo, Prajwal Manalwar "Exploiting Efficient Data Mining Techniques to Enhance Intrusion Detection Systems" *IEEE.,* (1).

Chang-Tien, Lu., Boedihardjo, A.P. and Manalwar, P. 2005. "Exploiting efficient data mining techniques to enhance intrusion detection systems," Information Reuse and Integration, Conf, 2005. IRI -2005 IEEE International Conference on., vol., no., pp.512,517, 15-17.

Denatious, D.K. and John, A. 2012. "Survey on data mining techniques to enhance intrusion detection," Computer Communication and Informatics (ICCCI), *International Conference on* , vol., no., pp.1,5, 10-12 Jan. 2012

Devaraju, S. and Ramakrishnan, S. 2015. " Detection of Attacks for IDS using Association Rule Mining Algorithm " , IETE Journal of Research.

Dewa, Zibusiso and Leandros, A. 2016. "Data Mining and Intrusion Detection Systems", *International Journal of Advanced Computer Science and Applications*,

Ghanem, T. F., Elkilani, W. S., Abdul-kader, H. M. 2015. " A hybrid approach for efficient anomaly detection using metaheuristic methods " , Cairo University, *Journal of Advanced Research,*

Hall, M., *et al*., 2009. The WEKA Data Mining Software: An Update. SIGKDD Explorations, 11(1).

Hwang, K., Fellow, M., Cai, Y., Chen, M., Qin, 2007. " Hybrid Intrusion Detection with Weighted Signature Generation over Anomalous Internet Episodes" , *IEEE.,*

Jingke, Xi. 2008. "Outlier Detection Algorithms in Data Mining," Intelligent Information Technology Application, 2008. IITA '08. Second International Symposium on, vo l.1,no., pp.94, 97, 20-22 Dec. doi: 10.1109/IITA. 2008. 26.

John, G., Cleary, Leonard, E. and Trigg, 1995. "K*: An Instance-based Learner Using an Entropic Distance Measure", 12th International Conference on Machine Learning, 108-114.

Kesavaraj, G. and Sukumaran, S. 2013. "A study on classification techniques in data mining," Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on , vol., no., pp.1,7, 4-6.

Mahmood, D. Y. and Hussein, M. A. 2013. "Intrusion Detection System Based on K-Star Classifier and Feature Set Reduction" , IOSR Journal of Computer Engineering (IOSR-JCE) Volume 15, Issue 5 ,PP 107-112.

Mohd. Junedul Haque, Khalid W. Magld and Nisar Hundewale, 2012. " An Intelligent Approach for Intrusion Detection Based on Data Mining Technique" *IEEE.,* (4).

Mookerjee, V., Mookerjee, R., Bensoussan, A. and Yue, W. T. 2011. " When Hackers Talk: Managing Information Security Under Variable Attack Rates and Knowledge Dissemination ",Information Systems Research , Vol. 22, No. 3, , pp. 606-623.

Moorthy, M. and Dr. Sathiyabama, S. 2012. "A Study of Intrusion Detection using Data Mining" (5).

Nadiammai, G.V. and Hemalatha, M. 2014. "Effective approach toward Intrusion Detection System using data mining techniques", *Egyptian Informatics Journal,*

Pan, S., Morris, T. and Adhikari, U. 2015. " Developing a Hybrid Intrusion Detection System Using Data Mining for Power Systems ", *IEEE.,*

Sànchez-Marrè, M. 2013. "From Data Mining to Big Data & Data Science: a Computational Perspective ".

Sangve, S. M. and Thool, R. C. 2015. "ANIDS: Anomaly Network Intrusion Detection System Using Hierarchical Clustering Technique " , Springer, pp. 274–285,

Surabi, H. 2014. "Hybrid Model for Intrusion Detection using Data Mining Techniques ", Master of Science in Computer Science, *Springer,*

Veeralakshmi, V. and Ramyachitra, D. Ripple Down Rule learner (RIDOR) Classifier for IRIS Dataset. Issues, vol 1, p.79-85.

W.-w Hu Liang, R. and Fei, R. 2009. " An adaptive anomaly detection based on hierarchical clustering ", Information Science and Engineering (ICISE), 2009 1st International Conference on, Changchun, China, Dec. pp. 1626–1629.

Zolotukhin, M., ¨am¨al¨ainen, T. H., Kokkonen, T., Niemel¨a, A. and Siltanen, J. 2015. "Data Mining Approach for Detection of DDoS Attacks Utilizing SSL/TLS Protocol"

****