ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

A Network Intrusion Detection System Using Clustering and Outlier Detection

J.Antony Jeyanna [1], E.Indumathi [2] ,Dr.D.Shalini Punithavathani [3]

P.G. Student, Department of Computer Engineering, Govt. college of Engineering, Tirunelveli, Tamilnadu, India[1]

P.G. Student, Department of Computer Engineering, Govt. college of Engineering, Tirunelveli, Tamilnadu, India[2]

Principal, Govt. college of Engineering, Tirunelveli, Tamilnadu, India[3]

ABSTRACT: With the growth of networked computers and associated applications, intrusion detection has become essential to keeping networks secure. A number of intrusion detection methods have been developed for protecting computers and networks using conventional statistical methods as well as data mining methods. It is necessary that the capabilities of intrusion detection methods be updated with the creation of new attacks. This paper proposes a hybrid intrusion detection method that uses a combination of supervised and outlier based methods for improving the efficiency of detection of new and old attacks. The method is evaluated with the benchmark intrusion dataset called the knowledge discovery and data mining Cup 1999 dataset and the new version of KDD (NSL-KDD) dataset. Thus the performance of our method is very good**.**

KEYWORDS**:** intrusion detection; supervised; outlier; NSL-KDD

I. INTRODUCTION

With the enormous growth of network-based computer services and the huge increase in the number of applications running on networked systems, the adoption of appropriate security measures to protect against computer and network intrusions is a crucial issue in a computing environment. Intrusions into or attacks on a computer or network system are activities or attempts to destabilize it by compromising security in confidentiality, availability or integrity of the system.

Intrusion Detection Systems (IDS) capable of detecting attacks in several available environments. This Intrusion Systems resolve ambiguities in passive network monitoring by placing detection systems on the line of attack. IDS are able to give prevention commands to firewalls and access control changes to routers. This can be seen as an improvement upon firewall technologies. It can make access control decisions based on application content, rather than IP address or ports as traditional firewalls do.

An intrusion detection system (IDS) monitors events occurring in a computer system or a network and analyzes them for signs of intrusions. Intrusion detection (ID) is a type of security management system for computers and networks. An ID system gathers and analyzes information from various areas within a computer or a network to identify possible security breaches, which include both intrusions (attacks from outside the organization) and misuse (attacks from within the organization). A network-based IDS (NIDS) often consists of a set of single-purpose sensors or host computers placed at various points in a network. These units monitor network traffic, performing local analysis of that traffic and reporting attacks to a central management console. Network-based intrusion detection is generally implemented using two approaches: rule-based and anomaly-based.

A boundlessness of methods for misuse detection as well as anomaly detection has been applied.The rule-based approach usually does not generate a large number of false alarms since it is based on rules that identify known intrusions but it fails to detect new types of intrusions as their signatures are not known.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

The anomaly detection approach has the ability to examine new or unknown intrusions. Thus given the promising capabilities of anomaly-based network intrusion detection systems (A-NIDS), this approach is currently a principal focus of research and development in the field of intrusion detection.

II. OVERVIEW OF ANOMALY DETECTION TECHNIQUES

Anomaly detection techniques can be classified into three main categories. They are statistical based, knowledge-based, and machine learning-based. In the statistical-based case, the action or reaction of the system is represented from an ergodic viewpoint. On the other side, knowledge-based A-NIDS techniques try to capture the asserted actions from available system data (protocol specifications, network traffic instances, etc.). Finally, machine learning A-NIDS schemes are based on the establishment of an explicit or unstated model that allows the structures analyzed to be classified.

2.1 Statistical-based A-NIDS techniques

In statistical-based techniques, the network traffic function is captured and an outline representing its random behaviour is created. This outline is based on metrics of basic features such as duration, protocol, source & destination ip address, source & destination port, services etc, content and time based features. Two datasets of network traffic are considered during the anomaly detection process: one corresponds to the presently observed outline over time, and the other is for the previously trained statistical outline. As the network events occur, the present outline is determined and an anomaly grade estimated by comparison of the two behaviours. The grade generally indicates the degree of abnormality for a specific event, such that the intrusion detection system will list the occurrence of an anomaly when the grade surpasses a certain threshold.
Furthermore, most of these schemes rely on the assumption of a quasi-stationary process, which is not always realistic.
Statistical-based A-NIDS techniques can further be classified into following categories:
a. Operational Model or Threshold Metric
b. Markov Process Model or Marker Model
c. Statistical Moments or Mean and Standard Deviation Model
d. Multivariate Model
e. Time Series Model

2.2 Knowledge-based techniques

Knowledge based detection Technique can be used for both signature based IDS as well as anomaly based IDS. It accumulates the knowledge about specific attacks and system vulnerabilities. It uses this knowledge to exploit the attacks and vulnerabilities to generate the alarm. Any other event that is not recognized as an attack is accepted. Therefore the accuracy of knowledge based intrusion detection systems is considered good. However their completeness requires that their knowledge of attacks be updated regularly.
Knowledge based detection Technique can further be classified as:
a. State Transition Analysis
b. Expert Systems
c. Signature Analysis
d. Petri Nets

2.3 Machine learning-based A-NIDS schemes

Machine learning based NIDS is one of the classification of anomaly based NIDS. Machine learning techniques are based on establishing an explicit or implicit model that enables the patterns analyzed to be categorized. A singular characteristic of these schemes is the need for labeled data to train the behavioral model, a procedure that places severe demands on resources. In many cases, the applicability of machine learning principles coincides with that for the statistical techniques, although the former is focused on building a model that improves its performance on the basis of previous results. Hence, a machine learning A-NIDS has the ability to change its execution strategy as it acquires new information. Although this feature could make it desirable to use such schemes for all situations, the major drawback is their resource expensive nature.
Machine Learning Based detection Technique can further be classified as:

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

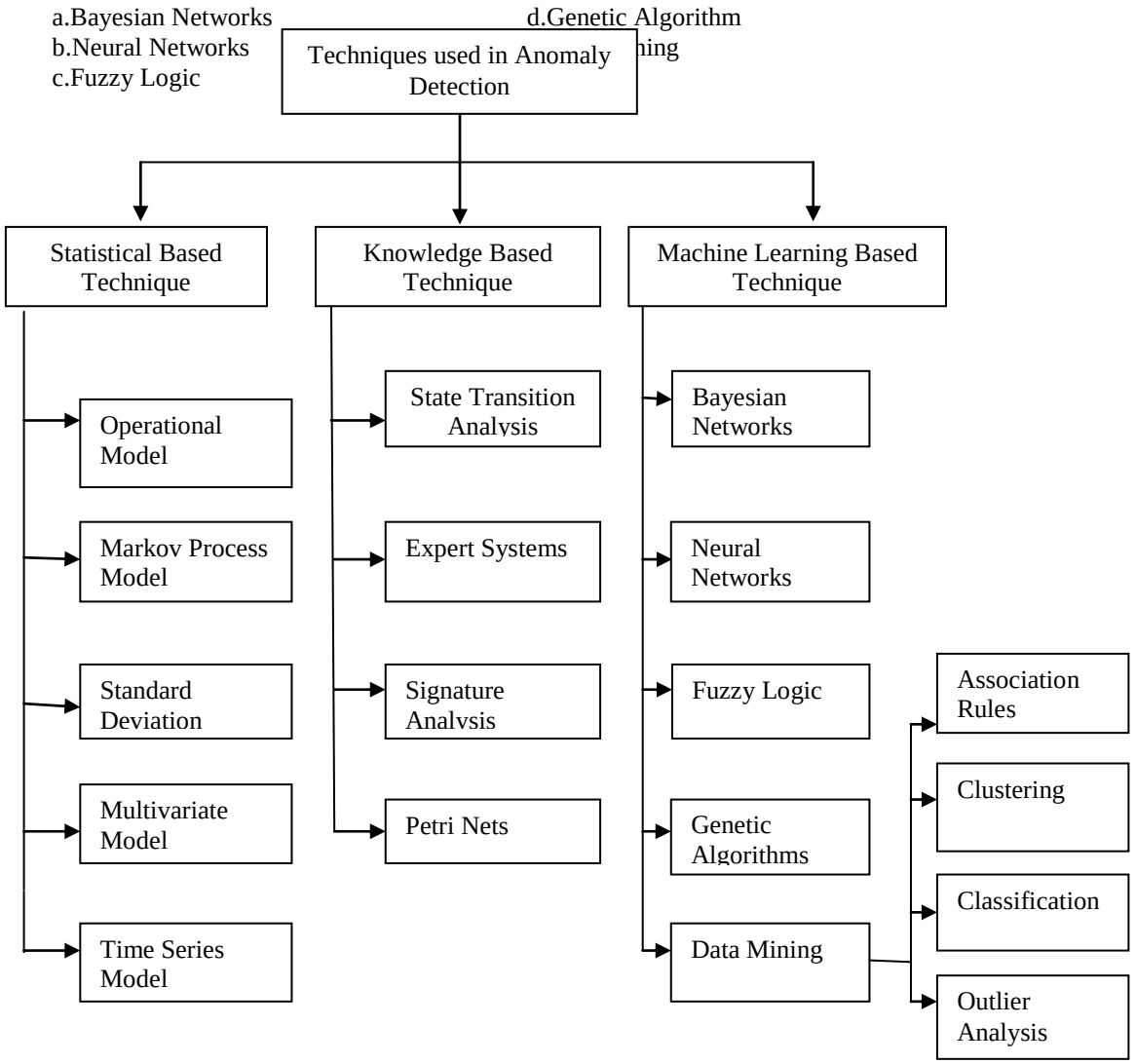


Fig 2.1 Anomaly Detection Techniques

III. RELATED WORK

Network attacks include four main categories: DoS (single source as well as distributed source), Probe, U2R and R2L following the standard classification of network intrusions as given in [16]. In the paper[1], the benefit of hybrid SVM via GA was illustrated also the paper has proven that by enhancing SVM with GA can reduce false alarms and mean square error (MSE) in detecting intrusion. In the paper[2], an intrusion detection model based on hybrid fuzzy logic and neural network was proposed. This new model has ability to recognize an attack with high detection rate and low false negative.In the paper[3], the two issues such as Accuracy and Efficiency by using Conditional Random Fields for hybrid intrusion detection system was presented. In the paper[4], a hybrid method of C5.0 and SVM was developed. The hybrid C5.0–SVM approach gave the best performance for probe, U2R and R2L attacks.In the paper[5], an intrusion detection model based on genetic algorithm and neural network was addressed. This approach uses evolution theory to information evolution in order to filter the traffic data and thus reduce the complexity. In the paper[6], a novel hybrid model for intrusion detection was proposed. The framework composed of

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

TAN and REP which can be effortlessly implemented in real time and is able to detect U2R and R2L attacks .In the paper[7], feature selection method is proposed. This method increases the efficiency of a given intrusion detection model and reduces the dataset looking for overlapping categories and also filters the desired features.In the paper[8], a simplified particle swarm optimization (SSO) is proposed. SSO is an optimization method that has a strong global search capability and is used for dimension optimization. **I**n the paper[9], a hybrid framework based on clustering and association was developed. This result better in terms of high detection and low false alarm rate.In the paper[10], an intrusion detection model based on hybrid neural network and SVM was presented. The key idea is to aim at taking advantage of classification abilities of neural network for unknown attacks and the expert based system for the known attacks. In the paper[11], a hybrid IDS is proposed. It is obtained by combining packet header anomaly detection (PHAD) and network traffic anomaly detection (NETAD). By combining anomaly-based and misuse based IDSs shows that the hybrid IDS is a more powerful system. In the paper[12], a hybrid approach for adaptive network intrusion detection which involves a hybrid model combining HMM based model with Naive Bayesian (NB) based approach was proposed. But it holds some difficulties that might arise when implementing HMM model in real time.In the hybrid system[13] the advantages of low false-positive rate of signature-based intrusion detection system (IDS) and the ability of anomaly detection system (ADS) was combined to detect novel unknown attacks. This technique leads to fast and accurate intrusion detection. In the paper[14], a design of fuzzy logic-based system for effectively identifying the intrusion activities within a network are proposed. The amount of data retained for processing i.e., attribute selection process was reduced. In the paper[15], an IDS is based on a general and enhanced flexible neural tree FNT. The FNT structure is developed using an evolutionary algorithm and the parameters are optimized by a particle swarm optimization algorithm.

IV. PROPOSED SYSTEM

A two-level IDS is proposed that is capable of detecting network attacks with a high degree of accuracy. The implementation of this IDS uses two levels of attack detection: a supervised method, and an outlier-based method. The selection of supervised or outlier-based classifier at a particular level for a given dataset is based on the classification accuracy of the individual classifier for a given dataset. The first level of classification categorizes the test data into three categories DoS, probe, normal and Rest (unclassified). U2R and R2L connections are classified as Rest at this stage. The main purpose at level 1 is to extract as many DOS and Probe connections as possible accurately from the data using a supervised classifier model. At level 2, the Rest category is classified as U2R and R2L attacks using an outlier detection model.

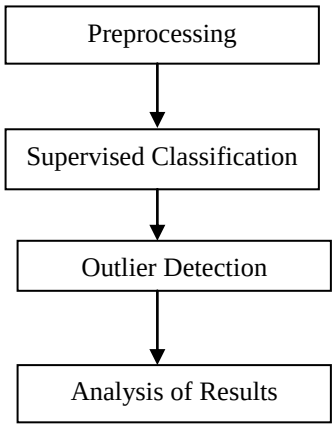


Fig 4.1 Block Diagram

This IDS has a high effective overall performance for known as well as unknown attacks. As in the fig4.1. it is a combination of the a supervised classifier based on a categorical clustering algorithm and a supervised outlier

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

detection algorithm based on symmetric neighbourhood relationship. The results are analysed with the benchmark intrusion dataset called NSL-KDD data set.

4.1 Preprocessing:

Each record in the dataset is described by 41 attributes, some of which are discrete, some continuous and some categorical values. Continuous valued attributes are discretized by taking logarithm to the base 2. Nominal valued attributes are mapped to discrete numerical codes (Serial numbers beginning with zero).The class label attribute is removed from the dataset and stored separately in a different file.

4.2 Supervised classification

The classification technique on [17] which creates a set of representative clusters from the available labeled training objects. Let the dataset to be clustered contain n objects, each described by d attributes A_1,A_2, . . . , Ad having finite discretevalued domains D_1,D_2, . . . , D_d , respectively. A data object is represented as X = {x_1, x_2, . . . , x_d }. The j th component of object X is xj and it takes one of the possible values defined in domain Dj of attribute Aj . Referring to each object by its serial number, the dataset can be represented by the set
N = {1, 2, . . . , n}. Similarly, the attributes are represented by the set M = {1, 2, . . . , d}.

The similarity between two data objects X and Y is the sum of per attribute similarity for all the attributes. It is computed as

$$\mathbf{sim(X, Y) = d \sum_{j=1} s(xj, yj),}$$

where s(xj, yj) is the similarity for the j th attribute defined as

$$\mathbf{s(xj, yj) = \{1 \text{ if } |xj - yj| \leq \delta j, 0 \text{ otherwise}\},}$$

where δj is the similarity threshold for the j th attribute. For categorical attributes δj = 0 and for numeric attributes δj ≥ 0.We use a subspace-based incremental clustering technique.

A cluster is a set of objects that are similar over a subset of attributes only. The minimum size of the subset of attributes required to form a cluster is defined by the threshold MinAtt. Let the subset of defining attributes be represented by Dattributes = {a_1, a_2, . . . , anoAttributes} such that Dattributes ⊆ M and noAttributes is the size of Dattributes. A cluster is represented by its profile. The profile representation is similar to that of an object. All objects in a cluster are similar with respect to the profile. The cluster profile is defined by a set of values, Values = {v_1, v_2, . . . , $v_{noAttributes}$} taken over the corresponding attributes in Dattributes, that is v1 ∈ Da1 is the value for attribute
a_1 ∈ M, v_2 ∈ Da_2 is the value for attribute a_2 ∈ M.
It involves

Subspace-based Incremental Clustering Technique.

- Cluster C = {Olist, Profile}

Olist->Index of the rows present in the cluster

- Profile = {noAttributes, Dattributes,Values}.

noAttributes->Number of features
Dattributes->Name of the features

To obtain appropriate values for δj and δaj , the following two
tasks are performed.

(i) Preprocessing of the dataset using logarithm to the base 2 normalization to remove bias. It discretize continuous-valued attributes by taking logarithm to the base 2 and then converting to integer. This is done for each attribute value z using the computation: if (z > 2)z = _(log2(z)) + 1. Before taking the logarithm, the attributes that take fractional values in the range [0, 1] are multiplied by 100 so that they take values in the range [0, 100].

(ii) Use of a heuristic method to identify appropriate range for δj and δaj by exhaustive experiments using benchmark and real-life datasets.

4.2.1 The CatSub+ algorithm

This algorithm on [17] which minimizes the dependency on input parameters by providing the probable range of parameter values based on a heuristic method. Further, CatSub+ is cost effective, since it works on subset of relevant

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

features selected using an information theoretic method. CatSub+ starts with an initially empty set of clusters. It reads each object Xi sequentially, inserts it in an existing cluster based upon the similarity between Xi and the clusters, or a new cluster is created with Xi if it is not similar enough, as defined by the threshold MinAtt, to be inserted in any one of the existing clusters. Search for a cluster for inserting the present object is started at the last cluster created and moves toward the first cluster until the search is successful. If successful, the object is inserted in the cluster found and the search is terminated. At the time of inserting the object in the found cluster C, the values of the defining attributes of the cluster (C.noAttributes) are set according to the computed similarity measure between the cluster and the object. The sets of C.Dattributes along with C.Values are also updated. If the search is not successful, a new cluster is created and the object itself made the representative object of the cluster, i.e. the full set of attributes becomes Dattributes while the full set of values of the object becomes corresponding Values of the new cluster profile. Initially, CatSub+ is trained with a fixed number of known clusters. Once the clusters and corresponding profiles are built for the known classes, newer instances are incrementally inserted in any one of the clusters. Before the initiation of cluster formation and respective profile building, all the unselected objects are marked as unprocessed. Similarity thresholds minAtt and minSize are assigned high values and they are gradually decreased in steps. In each iteration, the remaining unprocessed objects are clustered using the similarity measure, with reference to δ. If it fails to insert an object in any of the preexisting known clusters (created in the previous iteration), then a new cluster is created with the object. When the clustering process ends in the present iteration, cluster profiles are extracted from each of the clusters having at least minSize objects in it and the objects in such a cluster are marked as processed. All insignificant clusters, whose sizes are less than minSize, are deleted. The remaining new clusters become known clusters for the next iteration after making them empty by deleting their object lists. Then the threshold values minSize and minAtt are reduced so that the next iteration can create larger clusters instead of fragmented clusters. By reducing the thresholds, more generalization is allowed. The algorithm iterates so long as there are unprocessed objects remaining. To ensure termination of the algorithm, minSize is reduced to minSize/2 so that the ultimate value of minSize becomes 1, after which no objects will remain unprocessed. The threshold minAtt is loosened by setting minAtt = minAtt − α, where α is a small integral constant such as 1 or 2. Reduction of minAtt below a certain level (MIN) is not allowed. It remains constant atMIN.
Generalization beyond the MIN level will make data objects belonging to two different classes indistinguishable. When the clustering process terminates, the set of profiles found in the profile file becomes the final cluster profiles for use in the prediction process.

4.3 Outlier Mining based classification

Outlier Mining method on [17] which based on symmetric neighborhood relationships. For each object of the dataset, a forward neighbor outlier factor is estimated by finding the nearest neighbor set and the forward nearest neighbor set of the data objects to identify outliers.
In the dataset D = {d_1, d_2, . . . , d_n} of n objects, let di and dj be two arbitrary objects in D. It used Euclidean distance to evaluate the distance between objects di and dj , denoted as dist (di, dj). The Nearest Neighbor Set of k objects for an object p (NNk(p)) is the set of k nearest neighbor objects of p where k > 0. In dataset D of |D| objects, |NNk(p)| = k

(i) if ∀p _∈ NNk(p) and (ii) dist(o, p) < dist(´o, p) where o and ´o are kth and(k + 1)th nearest neighbors of p, respectively. Forward Nearest Neighbor Set of k objects of object p is the set of objects whose NNk contains p, denoted as FNNk(p). In dataset D of |D| objects where p and q are FNNk(p) = {q ∈ D | p ∈ NNk(q), p ≠ q}.
A score for each object of the dataset is computed based on |NNk| and |FNNk| of the object. The score is termed Forward Neighbor Outlier Factor of k objects (FNOFk) and it decides the strength of linkage of the object with other objects. The Forward Neighbor Outlier Factor of k objects for an object p is the ratio of the remaining number of objects of FNNk(p) of the dataset D (except object p) to the number of dataset objects (except object p), denoted as FNOFk(p).

4.3.1 GBBK Algorithm:

The outlier detection algorithm on [17] which named the GBBK algorithm consists of two functions: getFNOFk(D, k) and getNNk(D, p, k). The function getFNOFk(D, k) computes the distances among all objects using Euclidean distance, sorts all distances and searches for the shortest k distances. The function getNNk(D, p, k) searches for forward nearest neighbor objects for each of k objects returned by the function getFNOFk(D, k) for any object and computes a score.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

V. PERFORMANCE ANALYSIS

5.1 Supervised classification:

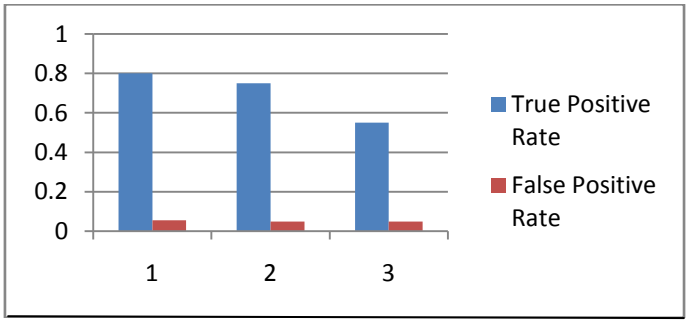
When the training set is 450 data objects and the testing set is 180 data objects .The high detection rate is 0.97778 percentage.

5.2 Outlier Classification:

When the testing set is 935 data objects

Index	Threshold	True Positive Rate	False Positive Rate
1	0.97	0.8	0.055
2	0.98	0.75	0.049
3	0.99	0.55	0.049

Table 5.4 Outlier Classification



VI. CONCLUSION AND FUTURE ENHANCEMENT

The proposed system presented a two-level hybrid intrusion detection method based on supervised and outlier methods. This method exhibits very good performance in detecting rare category attacks as well as large-scale attacks of both new and existing attacks when tested with a NSL KDD datasets. In further studies, this system will strive to create a more effective ensemble approach based on faster and efficient classifiers so as to make a significant contribution in the study of the intrusion detection.

REFERENCES

[1]Kayvan Atefi1, Saadiah Yahya2, Ahmad Yusri Dak3, and Arash Atefi4 "A HYBRID INTRUSION DETECTION SYSTEM BASED ON DIFFERENTMACHINE LEARNING ALGORITHMS" Proceedings of the 4th International Conference on Computing and Informatics, ICOCI 201328-30 August, 2013 Sarawak, Malaysia. Universiti Utara Malaysia (http://www.uum.edu.my)Paper No.022

[2]Muna Mhammad T. Jawhar & Monica Mehrotra "Design Network Intrusion Detection System using hybrid Fuzzy-Neural Network" International Journal of Computer Science and Security, Volume (4): Issue (3)

[3]Sandip Ashok Shivarkar Mininath Raosaheb Bendre "Hybrid approach for Intrusion detection using conditional random fields" International Journal of Computer Technology and Electronics Engineering (IJCTEE) Volume 1, Issue 3

[4]Vahid Golmah "An Efficient Hybrid Intrusion Detection System based on C5.0 and SVM" International Journal of Database Theory and Application Vol.7, No.2 (2014), pp.59-70

[5]Parveen Kumar,Nitin Gupta "A Hybrid Intrusion Detection System Using Genetic-Neural Network"
International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 National Conference on Advances in Engineering and Technology (AET- 29th March 2014)

[6]Mradul Dhakar+ and Akhilesh Tiwari "A Novel Data Mining based Hybrid Intrusion Detection Framework" ISSN 1746-7659, England, UK Journal of Information and Computing Science Vol. 9, No. 1, 2014, pp. 037-048 (Received June 23, 2012, accepted October 12, 2013)

[7]Witcha Chimphlee,Abdul Hanan Abdullah,Mohd Noor Md Sap"A Rough fuzzy hybrid algorithm for Computer Intrusion Detection"International Journal of Arab techonologies,VOl 4,No 3,July 2007 International Journal of P2P Network Trends and Technology (IJPTT) – Volume 3 Issue 8 - Sep 2013

[8]S. Revathi1, A. Malathi2 "Network Intrusion Detection Using Hybrid Simplified Swarm Optimization Technique"ISSN: 2249-2615 Page 375 Manish Somani1, Roshni Dubey2" Hybrid Intrusion Detection Model Based on Clustering and Association

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 2, February 2015

[9]Manish Somani1, Roshni Dubey2 "Hybrid Intrusion Detection Model Based on Clustering and Association" International Journal of Advanced Research in Electrical,Electronics and Instrumentation Engineering (An ISO 3297: 2007 Certified Organization)Vol. 3, Issue 3, March 2014

[10]Wesam K. AL-Rashdan, Reyadh Naoum, Wafa,S. Al_Sharafat, Mu'taz Kh. Al-Khazaaleh
"Novel Network Intrusion Detection System using Hybrid Neural Network (Hopfield and Kohonen SOM with Conscience Function)" IJCSNS International Journal of Computer Science and Network S 10 ecurity, VOL.10 No.11, November 2010 Manuscript received November 5, 2010 Manuscript revised November 20, 2010

[11]M. Ali Aydın *, A. Halim Zaim, K. Gokhan Ceylan "A hybrid intrusion detection system design for computer network security"Department of Computer Engineering, Faculty of Engineering, Istanbul University, 34320 Avcilar, Istanbul, TurkeyComputers and Electrical Engineering 35 (2009) 517–526

[12] R Rangadurai Karthick, R Rangadurai Karthick, Balaraman Ravindran,"Adaptive Network Intrusion Detection System using a Hybrid Approach", 978-1-4673-0298-2/12/$31.00 c 2012 IEEE

[13] Kai Hwang, Fellow, IEEE, Min Cai, Member, IEEE, Ying Chen, Student Member, IEEE, and Min Qin "Hybrid Intrusion Detection with Weighted Signature Generation over Anomalous Internet Episodes" IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, VOL. 4, NO. 1, JANUARY-MARCH 2007

[14] Chapke Prajkta P. , Raut A. B."Hybrid Model For Intrusion Detection System" International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume1 Issue 3 Dec 2012 Page No. 151-155

[15] Yuehui Chen,1,† Ajith Abraham,2,* Bo Yang1,‡ "Hybrid Flexible Neural-Tree-Based Intrusion Detection Systems"1School of Information Science and Engineering, Jinan University,Jinan 250022, P.R. China 2School of Computer Science and Engineering, Chung-Ang University,Seoul, Korea

[16] Lippmann, R. *et al.* (2000) Evaluating Intrusion Detection Systems: The 1998 DARPA Off-Line Intrusion Detection Evaluation. *Proc. DARPA Information Survivability Conf. and Exposition (DISCEX) 2000*, Los Alamitos, CA, USA, pp. 12–26.

[17] Prasanta Gogoi1, D.K. Bhattacharyya1,∗ , B. Borah1 and Jugal K. Kalita2 "MLH-IDS: A Multi-Level Hybrid Intrusion Detection Method" © The Author 2013. Published by Oxford University Press on behalf of The British Computer Society