International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 21 (2017) pp. 10929-10937

An Advanced Data Processing Based Fusion IDS Structures

C. Ramachandran

Assistant Professor, Department of Computer Applications,
Kumaraguru College of Technology, Saravanampatti, Coimbatore, Tamil Nadu, India.
Orcid Id: 0000-0003-0795-0650

Abstract

An Intrusion can be definite as any perform or act that attempt to crack the veracity, privacy or accessibility of a source this may enclose of a on purpose not permitted challenge to admittance the information, influence the data, or create a system defective or not viable. With the growth of computer networks at an alarming rate through the past decade, safety has become one of the serious issues of computer systems. IDS, is a discovery apparatus for detecting the invasive activities concealed among the usual activities. The radical organization of IDS has concerned analysts to work dedicatedly enabling the system to deal with industrial advancements. Therefore, in this view, a variety of constructive schemes and models have been planned in order to attain improved IDS. This paper proposes a novel fusion model for intrusion detection. The proposed structure in this paper may be predictable as another step towards development of IDS. The structure utilizes the vital data mining categorization algorithms helpful for incursion detection. The fusion structure will lead to proficient, adaptive and intellectual intrusion detection.

Keywords: Intrusion Detection, Feature Selection, Data mining, Conventional IDS.

INTRODUCTION

In modern years, with the incredible expansion in networked computer resources, a assortment of network-based applications have been developed to provide services in dissimilar areas such as ecommerce services, social media services, banking services, administration services, etc. The increase in the number of networked machines has lead to an increase in unconstitutional action, not only from exterior attacks, but also from internal attacks, such as people gaining unprivileged access for personal gain. Intrusion detection system (IDS) detects unauthorized intrusions into computer systems and networks.

Incidents may be malware attacks (such as worms, virus), attackers gaining unconstitutional entrance to system throughout Internet or user of the system gaining unprivileged root access of the system for which they are not approved.

An IDS monitors network traffic of a computer system like a network sniffer and collects network log data. The composed network data is analyzed by intrusion detection model or technique for rule violations. When some regulation contravention is detected the IDS alerts the network supervisor by raising alarm. Fig. 1 illustrates the design of Intrusion detection system.

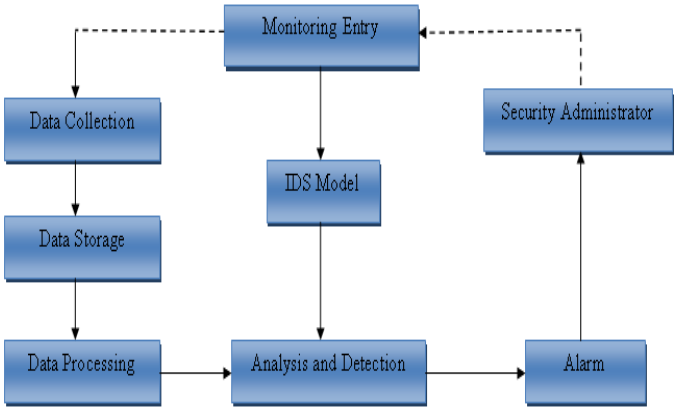

Figure 1: Design of Intrusion detection system

Usually IDSs are deployed to observe a scheme or a group in explore of any uncharacteristic circumstance. In this observation if any kind of invasive challenge is detected, the monitoring scheme i.e. IDS sets up an alarm which is a suggestion of the existence of intrusion. In order to detect intrusions in an efficient manner, various appreciable models have registered their presence in the literature. The currently available models involve procedure of a choice of novel algorithms which are likely to detect these intrusions distinguishably. Among these, algorithms based on data mining have been a point of desirability for researchers since of their wide probability in detect intrusions. These algorithms aid in improving precision of the system along with efficient detection rate and less false alarm rate. The algorithms faithful for categorization are the most attractive algorithms for discovery.

In the data mining organization techniques, Tree Augmented Naïve Bayes (TAN) and Reduced Error Pruning (REP) algorithms have come out as the most important recognition algorithms in IDS. This replica is a combinational system which aims at surmounting the shortcomings faced by two algorithms independently with fascinatingly enlarged precision of the recognition.

INTRUSION DETECTION SYSTEM

The adverseness of abnormalities on web has brought up the safety concerns leading to the successful completion of abnormalities discovery scheme named as Intrusion Detection System (IDS). Intrusions may be defined as the unconstitutional challenge for in advance admittance on a protected system or network. Intrusion detection is the route of exploit to detect apprehensive action on the network or a device. Intrusion Detection System (IDS) is a significant detection used as a countermeasure to protect data reliability and system accessibility from attacks.

The IDS has been a well-known feature for detecting intrusions sufficiently. The IDS is unspecified as hardware or software or arrangement of both that allows monitoring of the system traffic in search of intrusions. An intrusion detection system (IDS) inspects all inbound and outbound network action and identifies apprehensive patterns that may specify a network or system attack from someone attempting to break into or cooperation a system. It has profitably helped the analysts to learn about the various probable attacks.

Intrusion Detection Process

Intrusion detection on the basis of their detection process are categorize into Misuse / Signature-based intrusion detection and Anomaly-based intrusion detection.

Misuse Detection

Misuse detection compares the user activities to the known impostor activities on web. The idea of misuse detection is to represent attacks in the form of a pattern or a signature so that the same attack can be detected and banned in future. The IDS searches for distinct signatures and if a match is found, the system generates an alarm indicting the attendance of interference. As it works on the base of predefined signatures, it is incapable to detect new or before unknown intrusions.

Anomaly Detection

Anomaly intrusion detection identifies deviations from the normal procedure performance patterns to identify the intrusion. It is a method which is based on the informative of traffic anomalies. It estimates the deviation of a user activity from the normal performance and if the deviation goes further than a preset entrance, it considers that activity as an intrusion. It is since of this threshold concept anomaly can detect new intrusions in calculation to the before known intrusions. Nevertheless anomaly is able to detect new intrusion but the force for participation of limiting factor results in high proportion of false positive rate.

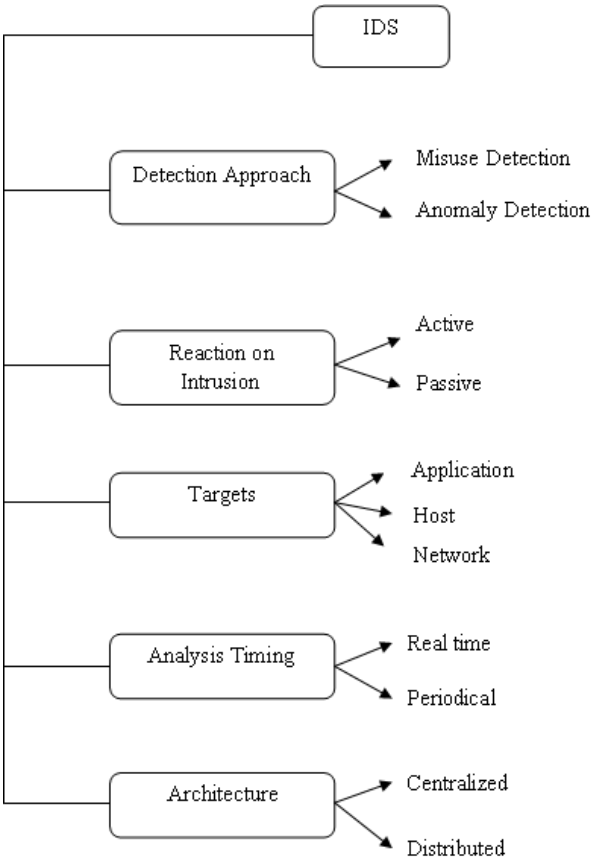

Figure 2: Approaches of intrusion detection

INTRUSION DETECTION APPROACHES

On the basis of the data analyzed and stored it is classified into Host-based Intrusion Detection System (HIDS) and Network-based Intrusion Detection System (NIDS).

Host-based Intrusion Detection System

Host-based IDS examine host-bound review sources such as operating system audit trails, system logs, and application logs. It is a software application which is installed onto a system in order to defend it from intruders. The audit data which is to be analyzed is collected from the host in the network. HIDS are OS dependent and thus need some previous planning before accomplishment and are proficient in detecting buffer overflow attacks.

Network-based Intrusion Detection System

Network-based IDS examine network packets that are captured on a network. In NIDS, discovery software is installed in a system in order to detect intrusions. NIDS collects data straight from the network in form of packets and are analyzed for detecting intrusions. It is OS self-sufficient and cannot scan protocol or content if the traffic in encrypted. It provides better safety against refutation of service attacks

Understanding Taxonomy Of Ids

A common description of Taxonomy is the practice or principle of categorization. Taxonomy may serve several purposes in design. Initially, it can explain the current global condition, supplementary in cleansing multifaceted situations and presenting it in a clearer deliberate approach. Furthermore using taxonomy to categorize a number of substances enables recognition of missing objects early in design, which allows users to develop the extrapolative qualities of a good taxonomy. Finally, a good taxonomy presents users with ideas, further explaining experiential current occurrence.

There has been much taxonomy presented for IDSs. The first real documented IDS taxonomy seems to be the one proposed by Debar et al. Since then many more taxonomies have been published, most particularly is the one proposed by Axelsson followed by another proposed by Halme and Baue.

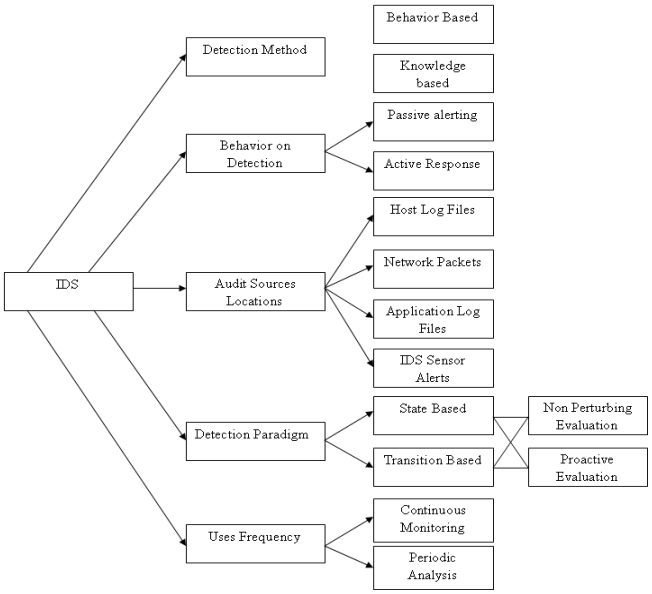

Figure 3: Updated IDS taxonomy

The famous taxonomies can be used in order to demonstrate general relationships in IDSs". Debar explains the meaning of understanding the system before creating an IDS and expanding on various mechanisms used in IDS to enable a prearranged approach in design. It is supposed that subsequent this practice results in improvement of well-organized Intrusion Detection Systems. There is no clear suggestion as to whether this is a necessary approach to creating IDSs", though the sheer sum of researchers who include this method into their design may serve as evidence, suggesting the significance of taxonomy.

The Halme and Bauer taxonomy named "A taxonomy of Anti-Intrusion Techniques", focuses on uncharted methods in hostility intrusion behavior, it focuses on this rather than dealing with IDSs". The taxonomy reveals six anti-intrusion approaches avoidance, pre-emption, deflection, prevention, detection, and/or separately countered. Axelssons" taxonomy proposes an improved advance, as it deals solely with the IDSs.

This commences with categorization of the section principles, followed by the operational tasks of the IDS. The taxonomy aims to analyze current IDSs, which accordingly allows progress in researching the chosen field enabling categorization to help aid improve knowledge of the field. In Figure 3 the taxonomy follows a series of steps, initially the categorization of the detection standard and then certain operational aspects of the intrusion detection system. In Figure 3, beginning starts during first identifying the different types of invasive behaviors generated by an intruder.

Intrusion Detection System Categories

By additional analyzing Axelssons taxonomy, two methods of IDS may be exposed when viewed from another perception as illustrated in Table 1. The first is building a taxonomy using principals of IDSs", where categorization is based upon the following discovery methods; abnormality detection, Signature discovery and Hybrid detection. The author establishes IDS categorization, based upon scheme individuality, time of discovery and response to detecting intrusions as obtainable in Table 1.

Table 1: General Ids Taxonomy

Anomaly	Self Learning	Non-time Series
		Time Series
	Programmed	Default Deny
		Descriptive Stats
Signature	Programmed	State Modeling
		Expert System
		String Matching
		Simple Rule-Based
Signature inspired	Self Learning	Automatic Feature Selection

DATA MINING IN INTRUSION DETECTION SYSTEM

Data Mining refers to the method of extracting efficient, efficient, concealed, useful, and the comprehensible pattern from a large unfinished, noise, non-stable and chance data. In intrusion detection system, the information deals from numerous sources such as system traffic or logs, system logs, application logs, alarm messages, etc. Due to varied data source and arrangement, the difficulty increased in auditing and examination of data. Data Mining has enormous benefit in data removal from large volumes of data that are noisy and active, thus it is of great significance in intrusion detection system.

k-Means

K-means is a partitioning method in clustering method of data mining. K-Means clustering method is used to separation the preparation data into k clusters with the help of Euclidean distance similarity. It is an algorithm to collection or to categorize the objects based on attributes/features into k number of clusters. Euclidean Distance equation to find distance between two objects is: D(a,b)= D(b,a)= |a-b|= Basic steps for clustering the data by k-means are:

- Select a number (k) of cluster centers - centroids (random)
- Assign every object to its nearest cluster center (e.g. using Euclidean distance)
- Move each cluster center to the mean of its assigned objects
- Repeat steps 2,3 until convergence (change in cluster assignments less than a threshold)

CART (Classification and Regression Trees)

Classification tree analysis is used to recognize the "class" to which the data belongs. Regression tree analysis is where the data is incessant and tree is used to predict its value. The term Classification and Regression Tree (CART) study is used to refer to both of the above events. Classification and regression trees are machine- learning methods for constructing prediction models from data. The Classification and Regression Trees (CART) methodology is technically called as binary recursive partitioning. The process is binary because parent nodes are always split into exactly two child nodes and recursive because the process is repeated by treating each child node as a parent. The key elements of CART analysis are a set of rules for splitting each node in a tree; deciding when tree is complete and assigning a class outcome to each terminal node.

The main steps of CART are:

1. Rules for splitting data at a node based on value of a variable

2. Stopping when a branch becomes a leaf/terminal node and cannot be split further

3. Finally a prediction for target variable in each leaf/terminal node.

4. CART does not rely on data belonging to a particular type of distribution.

5. It is not significantly impacted by outliers in input data.

IDS Architecture in Data Mining

A large amount of data mining techniques can be used in intrusion detection, each with its own exact benefit.

Classification: Creates a classification of tuples. It could be used to detect creature attacks, but as described by previous sample experiments in the literature indication it is produce a high false alarm rate. This problem may be reduced by applying fine-tuning techniques such as boosting.

Association: Describes relationships within tuples. Detection of irregularities may occur when many tuples display before unseen associations.

Grouping: Groups tuples that exhibit parallel properties according to pre-described metrics. It can be used for universal analysis similar to categorization, or for detecting outliers that may or may not characterize attacks. Figure.4

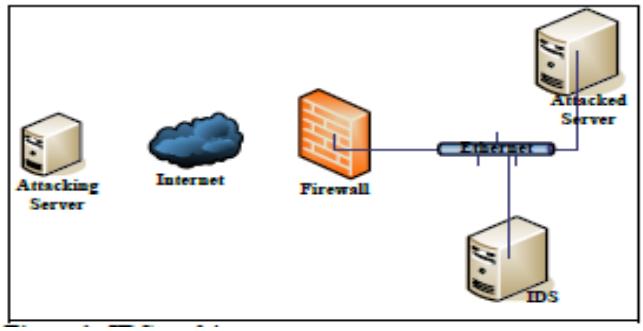

Figure 4: IDS Architecture

Attacks Detected By IDS

Following are the four types of attacks on ground being detected by IDS:

Denials-of-Service (DoS)

Denials-of Service attacks have the objective of restraining or denying services provided to the user, computer or network. Attacker tries to prevent legitimate users from using a service. It is typically done by creation the possessions either too busy or overflow, which as an importance results in the defiance of services requested by the legitimate users.

Probing or Surveillance

Probing or Surveillance attacks have the goal of gaining associate of the subsistence or pattern of a computer method or system. The attacker subsequently tries to harm or rescue information about the possessions of the casualty system.

User-to-Root (U2R)

User-to-Root attacks are attempts by a non-privileged user to increase managerial privileges. In the attack, users take benefit of system leak to get access to authorized purview or administrator's purview, such as: Buffer Overflow is among them. The goal of gaining root or super-user access on a exacting computer or a system is to cooperation the vulnerabilities of the system.

Remote-to-Local (R2L)

Remote-to-Local attack is the type of intrusion attack where the remote intruder constantly sends packets to a local machine with reason to expose the machine vulnerabilities and develop privileges which a local user would have on the computer.

PROPOSED METHODOLOGY

The proposed system is a hybrid intrusion detection structure based on the permutation of two classifiers i.e. Tree Augmented Naïve Bayes (TAN) and Reduced Error Pruning (REP). The TAN classifier is used as a base classifier while the REP classifier is used as a Meta classifier. The Meta categorization is the learning method which learns from the Meta data and judge the rightness of the categorization of each occurrence by base classifier. The judgment from each classifier for each class is treated as a feature, and then builds another classifier, i.e. a meta-classifier, to make the final decision.

The working of hybrid structure can be understood in following algorithmic steps:

Step 1: Input dataset

Step 2: Execute preprocessing of the dataset

Step 3: Choose TAN as the base categorization algorithm

Step 4: Select REP algorithm for Meta categorization

Step 5: Perform categorization on base classifier for Meta Rules

Step 6: Set the obtained Meta rules as input for Meta categorization

Step 7: Achieve re-classification using Meta classifier .The major design of with this process is to enlarge the general categorization production ensuing in enhanced result than any other existing method. The two classifiers indulged in the proposed system.

Tree Augmented Naïve Bayes Algorithm

The Tree Augmented Naïve Bayes (TAN) is a Bayesian system knowledge method and it is the addition to simple Naïve Bayes classifier. Naive Bayes is probabilistic classifier arrangement based on Bayes theorem having naive (strong) independence assumptions. This arrangement encodes the strong provisional self-government supposition amongst attributes i.e. the class nodule is the parent node for each and each attribute node with no parent node distinct for it. Thus the joint likelihood as represented by:

$$p(c, v_1, v_2, \dots, v_n) = p(c)\prod_i p(v_i|c)$$

The TAN network is a development on Naive Bayes that relaxes the strong provisional independence supposition. It allows extra boundaries between the attributes of the system in order to capture correlations among them. Similar to Naive Bayes, each attribute can have class node called augmenting edge pointing to it. The augmenting edges encode arithmetical dependencies between attributes. Thus, the joint likelihood in TAN depends on probabilities trained not only on class in-fact also on an attribute parent node as well.

$$p(c, v_1, v_2, v_3, \dots\dots, v_n) = p(c)\prod_{i=1}^{n} p(v_i|\, pa_{v_i}, c)$$

From these system structures and equivalent joint distributions, we can calculate class predictions

$$\hat{C}(V) = argmax_c P(C|V) \propto P(C)\prod_i P(V|C)$$

In Naive Bayes, the network organization is known earlier while TAN system structures needs to be cultured. The learning is done by scheming conditional common information function among two attributes. This purpose by decision the maximal weighted across tree in a graph is used in view to build the maximum-likelihood tree. The following are the steps for this procedure as discussed in:

1) Compute the provisional common information given C among every pair of separate variables,

$$I(X_i; X_j|C) = \sum_{x_i, x_j, c} \hat{P}(x_i, x_j, c)\log\frac{\hat{P}(x_i, x_j|c)}{\hat{P}(x_i|c)\hat{P}(x_j|c)}$$

Reduced Error Pruning Algorithm

REP is a quick choice tree learner classifier of data mining method. It uses a justification data set for estimating simplification fault. The mistake is pruned in the method for each node in the tree. Fundamentally the node with the uppermost abridged error rate is pruned. Pruning can be unstated as a method whose object is to decrease the size of decision tree by removing parts of a tree that allow better classification of instances. Pruning consequently reduces the categorization difficulty with increased correctness. Therefore, a pruning of a tree is a sub tree of the original tree with just zero, one or more interior nodes changed into leaves.

The pruning in REP is always done at leaves where each node is replaced with it's the majority popular class. First, the preparation data are split into two subsets: a growing set (usually 2/3) and a pruning set (1/3). The rising phase is used to produce the rules for constructing categorization tree while pruning phase performs pruning.

Next an error rate is calculated for each node, predictable as the number of instances that are misclassified on a substantiation set by propagating errors rising from the leaf nodes. Lastly, if the dissimilarity is a decrease in error then the sub-tree below the node can be measured for pruning. In position to the conversation of REP, following are the

algorithmic steps undertaken.

Step 1: Choose dataset

Step 2: Divide the input dataset into two subsets, rising set and substantiation set.

Step 3: Repeat the pruning phase i.e. step 4 and 5 for each node in the tree

Step 4: Estimate the crash on the confirmation set i.e. error rate for every node.

Step 5: Eliminate the node which maximally improves the precision of the confirmation set. set i.e. the node with maximum compact error rate.

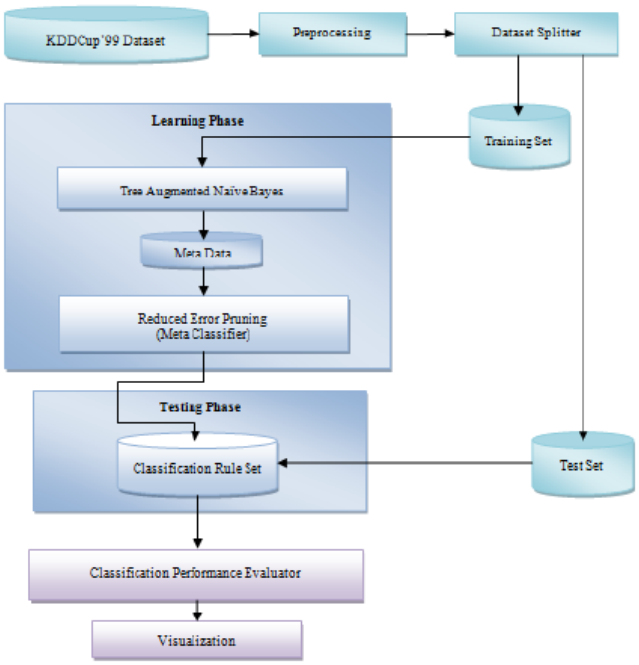

Figure 5: Fusion Intrusion Detection structure

Detailed Description Of The Fusion Ids Structure

This section describes about all the modules incorporated in the fusion IDS structure shown in fig. 1. Following is the brief discussion about each module:

KddCup'99 Dataset

The kddcup'99 dataset is a benchmark dataset which is originated by processing the tcp dump section of DARPA 1998 assessment dataset. The KDDCup'99 dataset was originated by processing the tcpdump segment of DARPA 1998 assessment dataset. The data set consists of 41 features and a divide characteristic that labels the link as 'normal' or a type of attack. The data set contain a total of 24 attack types that fall into 4 major categories (DoS, Probe, R2L and U2R).

Preprocessing

In the preprocessing component the class label presents in the 42nd feature of KddCup'99 dataset is recast into five main categories for the sake of declining difficulty of presentation assessment of the planned replica. As the unique KddCup'99 dataset having 22 types of attack labels, it was very difficult to assess the presentation of the categorization model. Hence the attack labels are customized to their individual category for the ease of analysis. Lastly five major classes are formed as the class label i.e. DoS, Probe, R2L, U2R and Normal.

Dataset Splitter

The Dataset Splitter unit partitions the dataset into two parts conventional from the preprocessing unit. To division the dataset into two parts a technique named holdout is used. In this method, the given data are accidentally partitioned into two free sets, a training set and a test set. The 66% of the data is owed to the preparation set and the residual 44% of the dataset is owed to the testing set. The training set is used to obtain the future structure while the test set is used to assess the correctness of the resultant replica. When the KddCup'99 dataset passed during the data splitting component then it gets separated into the preparation set.

Learning Phase

The knowledge point involves two steps for generating the categorization rules. In the primary step, the knowledge of bottom classifier i.e. TAN using the homework dataset is achieved. The result of this base classifier is implicit as the Meta data for the second step. This meta-level preparation set is collected by using the base classifiers' predictions on the justification set as quality principles, and the true set as the target. From these predictions, the meta-learner adapts the characteristics and performance of the base classifier and computes a meta-classifier which is a model of the original training data set. This meta-classifier in next step fetches the predictions from the base classifier for classifying an unlabeled occurrence, and then makes the final categorization choice.

Testing Phase

The categorization systems that are generated in knowledge stage are stored for the presentation estimate of fusion intrusion detection structure. In this stage, the Testing Set generated in Data Splitting unit is used as input to review the presentation. The result of this component is additional forwarded to next module i.e. Classifier recital assessor module.

Classifier Performance Evaluator

The Classifier Performance Evaluator module calculates a variety of categorization performance measures in order to judge the correctness of the fusion IDS structure. These measures are as follows:

True Positive Rate (TPR): TPR = $\frac{TP}{TP+FN}$

False Positive Rate (FPR): FPR= $\frac{FP}{TN+FF}$

Where, TP (True Positive), FN (False Negative), FP (False Positive) and TN (True Negative).

Experimental Analysis

This part describes the experimental analysis of the developed fusion intrusion detection structure and its assessment with a variety of other techniques current in the situation. It has been noticed that the outcomes of the fusion IDS structure excelled most of the algorithms in value of presentation (importantly accuracy).

Table 2: Performance Comparison of TAN and K2

Class	TAN		K2	
	TPR	FPR	TPR	FPR
DoS	0.992	0.000	0.976	0.002
Probe	0.985	0.000	0.970	0.002
R2L	0.975	0.001	0.969	0.001
U2R	0.867	0.001	0.856	0.005
Normal	0.987	0.001	0.985	0.002

Following Table 2 and 3 is the comparison of the two algorithms i.e. TAN and REP utilized in the fusion IDS structure with deference to the regularly favored bayes net based K2 algorithm.

Table 3: Performance Comparison of REP and K2

Class	REP		K2	
	TPR	FPR	TPR	FPR
DoS	1.000	0.001	0.988	0.000
Probe	0.978	0.000	0.978	0.005
R2L	0.982	0.000	0.959	0.001
U2R	0.667	0.000	0.810	0.005
Normal	0.999	0.000	0.985	0.002

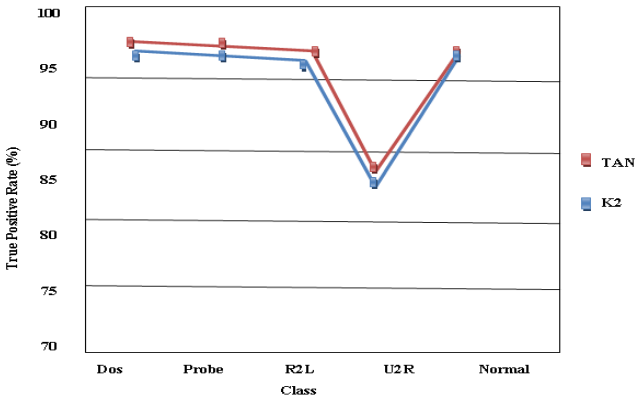

Figure 6 : Class-wise comparison of accuracy in K2 and TAN

Figure 6 and Figure 7 showing the number correctly classified instances and number of incorrectly classified instances and show the class-wise Accuracy comparison among K2, TAN, REP.

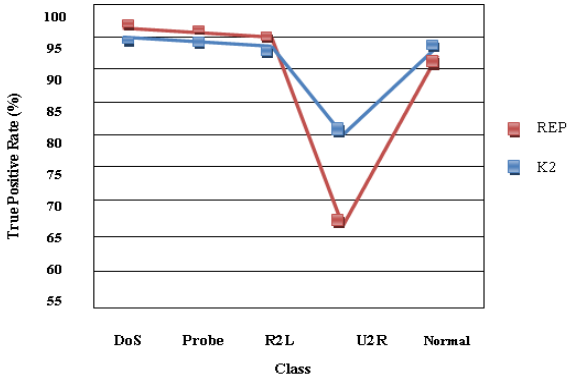

Figure 7 : Class-wise Comparison of accuracy in K2 and REP

CONCLUSION

This paper proposes an envisioning structure for intrusion detection i.e. fusion Intrusion Detection System. The developed structure is an intelligent, adaptive and efficient intrusion detection structure. The experimental analysis is performed on the developed IDS structure and is compared with other techniques current in the situation. The resultants obtained communicate that the developed fusion structure is extremely efficient to defeat the deficiencies establish in prior work. As the structure uses two data mining techniques (i.e. TAN and REP) to type the categorization rules, it can be naturally implemented in real time and is capable to distinguish and adjust new types of invasive behavior. Also experimental evaluation shows that the developed structure has abridged the fake alarm rate and improved the accuracy up to important expand which is a major anxiety in case of interruption discovery device. In calculation to this, the organization is able to notice U2R and R2L attack more professionally than previous findings, boosting up the

detection method. In prospect, some more work can be through in order to detect U2R and R2L attacks more precisely which may tend to additional improve the scheme efficiency.

REFERENCES

[1] Wenke Lee and Salvatore J.Stolfo. "Data Mining Approaches for Intrusion Detection". Proceedings of the 7th USENIX Security Symposium San Antonio, Texas, January 26-29, 1998.

[2] Sattarova Feruza Yusufovna. "Integrating Intrusion Detection System and Data Mining". Proceedings of International Symposium on Ubiquitous Multimedia Computing, IEEE (2008), pp. 256-259.

[3] Zillur Rehman, S S Ahmedur Rehman, Lawangeen Khan. "Survey Reports on Four Selected Research Papers on Data Mining Based Intrusion Detection System". University of West Indies at Mona, 2009.

[4] Parekh S.P., Madan B.S. And Tugnayat R.M. "Approach For Intrusion Detection System Using Data Mining". Journal of Data Mining and Knowledge Discovery, ISSN: 2229–6662 & ISSN: 2229–6670, Volume 3, Issue 2 (2012), pp.-83-87.

[5] Srinivas Mukkamala, Andrew H. Sung, Ajith Abraham. "Intrusion detection using an ensemble of intelligent paradigms". Elsevier, Journal of Network and Computer Applications 28 (2005) pp.-167–182.

[6] Sandhya Peddabachigari, Ajith Abraham,Crina Grosan, Johnson Thomas. "Modeling intrusion detection system using hybrid intelligent systems". Elsevier, Journal of Network and Computer Applications 30 (2007), pp.114-132.

[7] Daejoon Joo, Taeho Hong, Ingoo Han. "The neural network models for IDS based on the asymmetric costs of false negative errors and false positive errors". Elsevier, Expert Systems with Applications 25 (2003), pp.- 69–75.

[8] Farah Jemili, Dr. Montaceur Zaghdoud, Pr. Mohamed Ben Ahmed. "A Framework for an Adaptive Intrusion Detection System using Bayesian Network". Intelligence and Security Informatics, IEEE (2007).

[9] Mrutyunjaya Panda and Manas Ranjan Patra. " Network Intrusion Detection Using Naïve Baye". IJCSNS International Journal of Computer Science and Network Security, VOL.7 No.12, December (2007).

[10] Su-Yun Wu, Ester Yen. "Data mining-based intrusion detectors". Elsevier, Expert Systems with Applications 36 (2009), pp. 5605-5612.

[11] Wei-Hao Lin and Alexander Hauptmann. "Meta-classification: Combining Multimodal Classifiers". Springer, Mining Multimedia and Complex Data, LNAI 2797 (2003) pp. 217–231.

[12] Alexandra M. Carvalho, Arlindo L. Oliveira and Marie-France Sagot. "Efficient learning of Bayesian network classifiers: An extension to the TAN classifier". Proceedings of Advances in Artificial Intelligence, Springer, Volume 4830, (2007), pp 16-25.

[13] Ajit Singh. "Tree-augmented naive bayes". Homework 2 Problem 7 of Probabilistic Graphical Models, Fall 2006.

[14] Tapio Elomaa and Matti Kääriäinen. "An analysis of Reduced Error Pruning". Journal of Artificial Intelligence Research 15 (2001), pp. 163-187.

[15] Johannes Fürnkranz. "Pruning Algorithms for Rule Learning". Machine Learning, 27 Kluwer Academic Publishers (1997), pp. 139-172.

[16] KddCup99 dataset, available at http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, 1999.

[17] Jiawei Han & Micheline Kamber. "Data Mining: Concepts and Techniques". Second Edition, Morgan Kaufmann Publishers, 2006.

[18] Andreas L., Prodromidis and Salvatore J. Stolfo. "A Comparative Evaluation of Meta-Learning Strategies over Large and Distributed Data Sets". In workshop on Meta Learning, Sixteenth Intl. Conf. Machine Learning, 1999.