Classification of Intrusion Detection using PSO-SVM and Improved Decision Tree

Stuti Tiwari
Department of Computer Science & Engineering
stutitiwari19@gmail.com

Amit Dubey
Department of Computer Science & Engineering
amitdubey@oriental.ac.in

Abstract — Intrusion Detection is an efficient way of detecting the abnormal behavior of packets in the network, Although in data mining there are various effective decision tree based algorithms are implemented for the classification and detection of Intrusions in KDDCup99 Dataset. Here an efficient technique is implemented for the classification and detection of Intrusions in KDDCup99 Dataset using Feature Selection and Decision Tree based algorithms. The Proposed methodology works in two Stages Feature Selection using Particle Swarm Optimization with Optimization of PSO by Support Vector Machine and then Classification of Intrusion using Horizontal Partition Decision Tree. The Proposed Methodology implemented is more efficient in comparison with Decision Tree based algorithms.

I. INTRODUCTION

Current Intrusion Detection Schemes have numerous known shortcomings, such as: low accuracy (registering high False Positives and False Negatives); little real-time presentation (dispensation a great quantity of circulation in real time); incomplete scalability (storage a great quantity of user shapes and bout signatures); an incapability to notice new bouts (knowing new attacks when they are hurled for the first period); and feeble system-reactive capabilities (efficiency of response). This makes the extent of IDS an beautiful investigation arena. In recent years, investigators have explored methods such as false intelligence, autonomous agents, and distributed systems for detecting intrusion in network environments. An Intrusion Discovery System (IDS) can be distinct as a grouping of software and/or hardware mechanisms that screens computer organizations and increases an alarm when an interruption ensues [1].

Nevertheless, most computer schemes are still vulnerable to bouts from hackers, so it is vital to found a second line of resistance for these schemes in the procedure of an Intrusion Discovery System (IDS). IDS [2], [3], [4] play a significant character in attaining the survivability of material schemes and safeguarding their security from attacks. They aim to protect the availability, confidentiality, and honesty of dangerous network information systems by analyzing what occurs or has occurred during an interruption, and endeavoring to classify signs that a computer has been distorted. They can also take suitable movements to sever net influences, best proceedings, raise alarms, and repeat scheme administrators to take good events.

IDS are usually classified as host-based or network-based. Host-based systems [5], [6], base their choices on material gotten from a solitary crowd (usually log files, network traffic to and from the host, or information on processes running on the host), while network-based schemes find data by nursing system circulation between hosts, and are usually run on a separate machine.

Securing company data, activities and communications become an open challenge and particularly now where teamwork and on request facilities cultivate up. As introduced, the appearance of new needs involved the emergence of new vulnerabilities and cyber criminality organizations. The first step to deal with this threat is to understand what security means in our company/organization context. There are a number of security issues for a computer network environment [7].

As distinct by Heady et al. [8], an interruption is any set of movements that effort to encompass the honesty, discretion or handiness of a reserve. Intrusion leads to defilements of the safety strategies of a computer organization, such as unlawful admittance to secluded material, spiteful break-in into a computer organization, or version organization untrustworthy or unfeasible.

II. LITERATURE REVIEW

Gad El Rab et al. in [9] planned an attack cataloging that appears cooperative for the IDS assessment course. This organization has five scopes; (1) Dismissal source: designates the introduction opinion of attack, (2) Honor boom: mentions to the raised admission added by an attacker to the organization incomes, (3) vulnerability: stipulates the brow beaten system susceptibilities, (4) Carrier: Gad El Rab et al. in [10] planned an attack nomenclature that appears obliging for the IDS assessment procedure incomes by which the bout spreads the prey; either via system traffic or finished a local deed, (5) Board: denotes to the attack objectives. Although this classification is interesting, it does not shelter all dimensions of attacks from the perspective of the IDS evaluation, especially for the wireless environment. Describes the auxiliary incomes by which the dose spreads the victim; either via system circulation or finished a local action.

In this paper author e.t.al Kumar in [11] planned an dose classification that appears as a protection centric one. This organization was based on give somewhat the once-over bout autographs, to help finally in meaning and construction signature-based IDS. The novelist confidential the bout crosses in the following aspects: Reality, Arrangement, Regular Countenance decorations, and other designs that contain all other intrusion signatures that cannot be represented directly in one of the earlier categories.

Here author Killourhy et al. has [12] secret the doses from the standpoint of the anomaly-based IDS protector. This organization is founded on detecting the irregularities of bout appearance: Distant Character, Negligible Distant Arrangement, Latent Sequence, and Non-Anomalous Sequence. In foreign symbol, the bout appearance comprises a scheme call which not ever looks in the standard record. For negligible distant development, the attack seeming confines an organization call evolution which not at all develops noticeable in the standard certification, but all subsequences emerge in the normal documentation.

Cloud computing has produced noteworthy attention in both academe and manufacturing, but it's motionless an developing example [13]. Essentially, it purposes to combine the financial usefulness perfect with the evolutionary expansion of many present methods and figuring machineries, counting dispersed amenities, claims, and material structures involving of puddles of processers, systems, and stowing possessions.

In this newspaper we emphasis on the uncovering of the bargained machines in a system that are used for distribution spam mails, which are normally mentioned to as spam zombies [14].

Here they propose a new method based on hybrid defense-in-depth intrusion detection framework based on our previous work NICE [15] to detect and monitor the vulnerability and attacks in the cloud. For better detection of attacks, our framework incorporates attack graph analytical model to describe the detected vulnerabilities in each VM and creates the vulnerability dependency based on VM's reach ability in the virtual network.

Detection of cooperated machineries presently takes place largely at host level and/or network level. At the host level, while anti-virus and anti-spyware systems are effective in catching and preventing the spread of known threats [16], majority of users do not keep these security software updated or properly configured.

III. PROPOSED METHODOLOGY

The Proposed Methodology implemented here is based on the concept of applying Feature Selection from the KDDCup99 Dataset and then Classifying Normal Packets and Abnormal Packets based on Improved ID3 based Decision Tree.

1. Load an input KDDCup99 Dataset.
2. Apply PSO based Feature Selection and optimization is done using SVM for the Selection of Most Dependent attributes from the Dataset.
3. Apply Improved Horizontal Partition based Decision Tree algorithm for the classification of KDDCup99 Dataset.

FEATURE SELECTION USING SVM OPTIMIZED BY PSO

The methodology implemented here using the concept of feature selection from the KDDCup99 dataset so that the selection of features is done more accurately using the concept Support Vector Machine (SVM) which is a supervised machine learning approach and is based on the statistical learning theory optimized by Particle Swarm Optimization (PSO). The feature selection using SVM (Support vector machine) provides a hyper plane that efficiently separates the various classes in the KDDCup99 dataset. The input training dataset consist of tupples$\{x_i, y_i\}$, where i=1→ n and 'x' denotes the input vector (attributes of the training dataset) and 'y' contains the class labels {+1,-1}. SVM contains a hyper plane of the form$wo.p + bs = 0$, where 'p' is the dynamic point lying on the considered hyper plane and 'bs' denotes the bias value of the distance of hyper plane from origin and 'wo' denotes orientation of hyper plane. The feature selection process repeats till the optimum hyper plane is detected.

Particle Swarm optimization (PSO) is stochastic population based optimization technique which is based on the simulation behavior of birds within a group of flocks. PSO (particle swarm optimization) contains a number of particles from which the problem can be solved. It contains personal best solution denotes as 'Pbest' which provides position of particle so that the maximum value used by the attributes can be predicted used for classification of attributes. It also contains local best particle denoted as 'Lbest' indicating position of entire swarm best position. The leader particle is used for the best search space from the set of attributes. The Vector velocity contains the direction of the particle it needs to move to another position from the current position. The weighted factor 'W' is the inertia is used to control the particle movement position depending on the previous particle position. Particle Swarm Optimization (PSO) contains two learning constant factors C1 & C2 which denotes the Cognitive learning factor and Social Learning Factor.

PSEUDO CODE FOR FEATURE SELECTION PROCESS USING PSO BASED SVM

Start with the Initialization of Population
While! (Ngen || Sc)
For p=1 :Np
If fitness Xp> fitness pbestp
Update pbestp = Xp
For $k \in NXp$
If fitness Xk>gbest
Update gbest = Xk
Next K
For each dimension d

$$v_{pd}^{new} = w * v_{pd}^{old} + c_1 * rand_1 * \left(pbest_{pd} - x_{pd}^{old}\right) + c_2 * rand_2 * \left(gbest_d - x_{pd}^{old}\right)$$

If $v_{pd} \notin (V_{min}, V_{max})$

$$v_{pd} = \max\left(\min\left(V_{max}, v_{pd}\right), V_{min}\right)$$

$$x_{pd} = x_{pd} + v_{pd}$$

Next d
Next p
Next generation till stop

The particles are first encoding into a bit string S=F1F2….Fn, n=1,2…m and the bit {1} represents for the selected feature from the dataset and the bit string {0} is the non-selected feature from the dataset. The evaluation parameters can be computed using SVM. Let us suppose in the dataset the available feature set is 10 then set {F1F2F3…..F10} is then analyzed using PSO and selection of any number of features say 5 a dimensional evaluation of these 5 features is computed using SVM. Each particle in PSO is renewed using adaptive computation of SVM, hence on the basis of which pbest is chosen. Now for the final feature selection each of the particle is then updated according to operation.

$$v_{pd}^{new} = w * v_{pd}^{old} + c_1 * rand_1 * \left(pbest_{pd} - x_{pd}^{old}\right) + c_2 * rand_2 * \left(gbest_d - x_{pd}^{old}\right)$$

$$S\left(v_{pd}^{new}\right) = \frac{1}{1 + e^{-v_{pd}^{new}}}$$

If $\left(rand < S\left(v_{pd}^{new}\right)\right)$

$$x_{pd}^{new} = 1 \; else \; x_{pd}^{new} = 0$$

Table 1. Various Notations used in Pseudo Code

Parameter	**Summary**
Ngen	Number of generations or iterations
Sc	Stopping Criteria
Np	Number of particles
Xp	Current position of pheromone
Pbestp	Pheromone with best fitness
Xk	Current particle position
Gbest	Best fitness value
K	Current particle number
v_{pd}^{new}	Updated particle velocity
v_{pd}^{old}	Current particle velocity
rand1	Random number 1
rand2	Random number 2
a1	Acceleration factor 1
a2	Acceleration factor 2
V_{max}	Maximum Velocity

The renewed features are then calculated using Eq. and hence on the basis of renewal calculation of ‘S’ and depending on the previous value of ‘S’ the features are selected as {1} otherwise {0} means the feature is not selected.

1. Initialization of Input Parameters such as No. of Particles and Dimension and Cross Folds validation.

Since PSO (Particle Swarm Optimization) is applied here for the optimization of selection of features, hence the parameters of PSO needs to be initialized such as number or particles and their position velocity and the cross folds used by SVM (Support vector machine).

2. For each of the particle and objects chose a random instance values from the dataset and their respective moving position.

As soon as the initialization step is over the particles or features of the KDDCup99 dataset is chosen so that the best or dependent attributes are selected.

3. The Particle Current Position can be considered as the Best Position means as the best instance value.

The random feature selected assumes to be the best attribute of the dataset and so is the fitness value as best and selection of features starts from this feature of the dataset.

4. For Each of the particle for max_iterations

Now compute for all the features of the KDDCup99 Dataset.

5. Selection of the particle position in the dataset as an instance values as (x, y).

The feature to be selected (Particle) moves along ‘X’ and ‘Y’ axis for the next best feature from the dataset depending upon the fitness value.

6. Now Initialize the Input Parameters of SVM

Here initialization of parameters such as type of kernel used and margin width of SVM and number of iterations is done.

7. ker = '@linearKernel' OR ‘GaussianKernel’

Selection of kernel as linear or Gaussian (RBF) on the basis of which SVM iterates and select features.

8. p1 = x;

Here ‘x’ is selected as the particle or feature of the dataset.

9. C = y;

‘y’ is assumed to be the class index of SVM.

10. trnX=X;

Select the ‘X’ as the training value of the SVM.

11. trnY=Y;

Select Y as the training classes of SVM.

12. tstX=X';
13. tstY=Y';
14. Training is performed on the basis of (trnX, trnY, C)

The selection of features starts with the basic input to SVM as the training values and class index.

15. On the basis of Training Parameters as (trnX, trnY, tstX, ker, alpha, bias, actfunc);

A predefined function is defined which computes the features on the basis of above parameters.

16. Selection of ‘Y’ as the features values can be predicted.

HORIZONTAL PARTITION DECISION TREE

Input Layer – Contribution layer encompasses of all the gatherings that are complicated in the computation process. They individually compute the Material Gain of each feature and send Intermediate result to UTP. This process is done at every period of decision tree.

Output Layer – The UTP exists at the 2nd layer i.e. the computation layer of our protocol. UTP gathers only transitional consequences from all gatherings not data and analyze the total evidence gain of each quality. Then find the attribute with highest information gain and then create the root of decision tree with this attribute and send this attribute to all parties for further calculation. This process is also done at every point of decision tree.

Informal Algorithm

Input Layer

1. Party independently analyzes Expected Material of every characteristic.

The contribution dataset taken here is first divided into a number of parties. Here parties are the various users who can calculate dataset attributes. The Information describes here is the impression of particular class in the dataset and is given by:

$$I(y,n) = -\frac{y}{(y+n)}\log\left(\frac{y}{y+n}\right) - \frac{n}{(y+n)}\log\left(\frac{n}{y+n}\right)$$

Where, I is the information to be computed for the classes 'y' and 'n'.

2. Party exclusively analyses Entropy of each feature.

After calculating the Information of dataset by each of the party. Entropy is computed based on the classes and attributes. The entropy can be computed using:

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

3. Party separately analyses Information Gain of each characteristic.

Finally after the calculation of the Entropy of each of each of the attribute Gain of each of the attribute is computed on the basis of parties. The Information Gain Computed here computes the dependency factor of attribute in the whole dataset.

4. Calculation of information gain from Han and Kamber and Pujari.
5. Accept there are two modules, *P* and *N*
6. Let the established of specimens *S* comprehend *p* essentials of class *P* and *n* fundamentals of class *N*
7. The quantity of evidence, desirable to choose if an subjective instance in *S* fits to *P* or *N* is distinct as

$$I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

8. Assume that using feature A set *S* will be segregated into gangs $\{S_1, S_2, \ldots, S_v\}$
9. If S_i comprehends p_i instances of *P* and n_i samples of *N*, the entropy, or the projected evidence required to organize substances in all subtrees S_i is

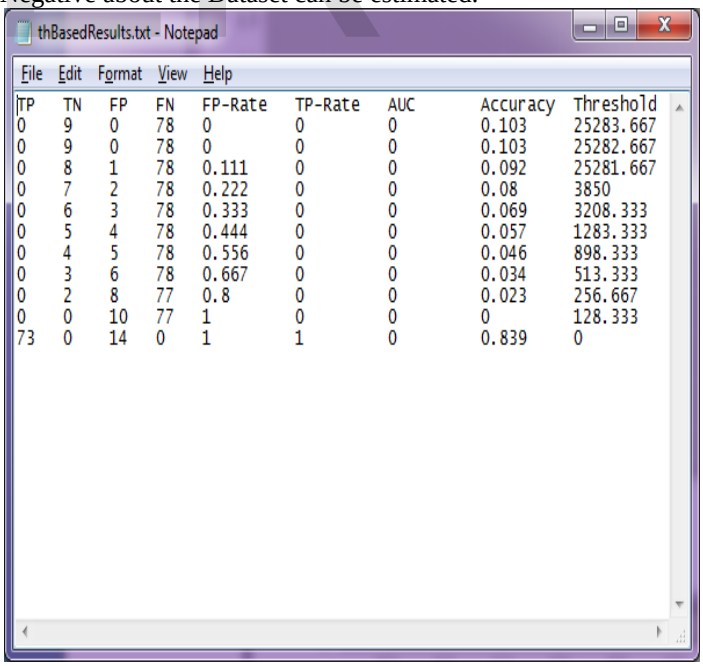
$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

10. The training material that would be expanded by dividing on A

$$GAIN(A) = I(p, n) - E(A)$$

IV. RESULT ANALYSIS

The Table shown below is the result estimation of the Threshold based results through which True Positive and True Negative about the Dataset can be estimated.



TP	TN	FP	FN	FP-Rate	TP-Rate	AUC	Accuracy	Threshold
0	9	0	78	0	0	0	0.103	25283.667
0	9	0	78	0	0	0	0.103	25282.667
0	8	1	78	0.111	0	0	0.092	25281.667
0	7	2	78	0.222	0	0	0.08	3850
0	6	3	78	0.333	0	0	0.069	3208.333
0	5	4	78	0.444	0	0	0.057	1283.333
0	4	5	78	0.556	0	0	0.046	898.333
0	3	6	78	0.667	0	0	0.034	513.333
0	2	8	77	0.8	0	0	0.023	256.667
0	0	10	77	1	0	0	0	128.333
73	0	14	0	1	1	0	0.839	0

Table 1 Result Estimation of Threshold results

The Figure shown below is the Analysis of Graph for the values of true Positive and False Positive Rate. Also shows

that when true positive rate increases the false positive decreases.

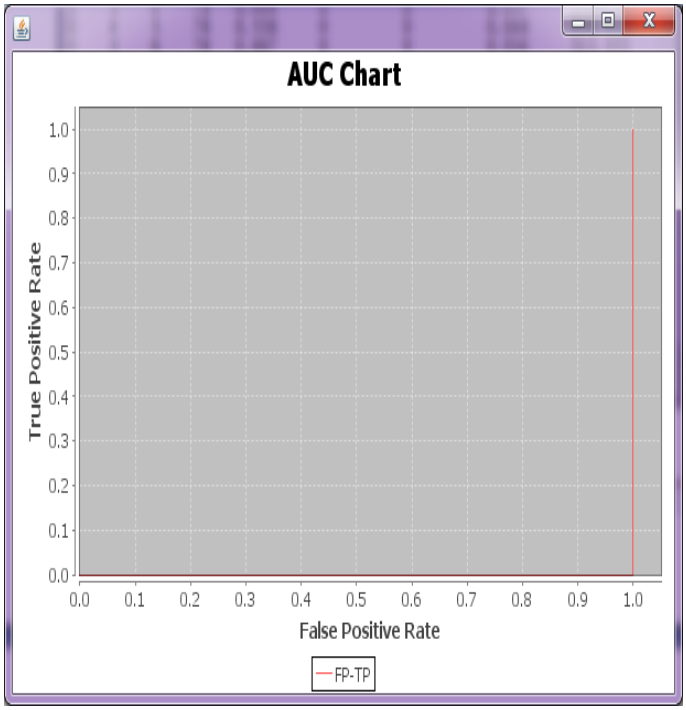


Figure 2 Result Analysis between TP and FP rate

The Table shown below is the analysis and comparison of various Decision Tree Algorithm on the basis of various parameters. The Various Methodology implemented for the Detection of Intrusions on KDDCup99 Dataset provides efficient results, but the proposed methodology proves to be more efficient in comparison with other Decision tree based algorithms.

S. No	Bayes Net	OneR	J48	Proposed
Correctly Classified Instance	97.3732	90.9139	99.742	99.86
Incorrectly Classified Instance	2.6268	9.0861	0.258	0.175
Kappa Statistics	0.9571	0.8478	0.9957	0.997
Mean Absolute Error	0.0024	0.0079	0.0003	0.0002
Root Mean Squared Error	0.0434	0.0889	0.0145	0.0084
Relative Absolute Error	4.5617	15.0351	0.656	0.486
Root Relative Squared Error	26.7982	54.8395	8.9515	6.375

Table 2 Performance Comparison of various Decision Tree based Algorithms

V. CONCLUSION

Detection of intrusion in network is necessary since the intrusion may create harm or attack any application which needs to be detected and prevented. Although there are various algorithms implemented for the detection of intrusions, but the classification of these intrusion is also an important concern since the type of attack depends on the intrusion. Since Intrusion Detection is a technique of finding the unwanted packets from the packets that shows abnormal behavior during the flow of the packets.

The Proposed Methodology implemented for the Detection and Classification of Intrusion in KDDCup99 dataset using Feature Selection by Particle Swarm Optimization with Optimization of SVM and then Classifying KDDCup99 Dataset using Horizontal Partition based Decision Tree. The Various result analysis shows that the proposed methodology is better in comparison with other Decision Tree based algorithms.

REFERENCES

[1] R. Heady, G. Luger, A. Maccabe, and M. Servilla, "The architecture of a network level intrusion detection system", Technical Report CS90-20, Department of Computer Science, University of New Mexico, August 1990.

[2] Th. Verwoerd and R. Hunt, "Intrusion Detection Techniques and Approaches", Journal in Computer Communications, Page(s): 1356-1365. 2002.

[3] Common Intrusion Detection Framework Working Group, "A CISL" Tutorial. http://www.gidos.org/tutorial.html. (March 2009).

[4] B. Mukherjee, L.T. Heberlein, and K.N. Levitt, "Network Intrusion Detection" IEEE Network, Page(s): 26-41, Vo1.8, No.3, May-June 1994.

[5] P. Lichodzijewski and A. Zincir, "Host-Based Detection Using Self-Organizing Maps", Proceedings of the 2002 International Joint Conference on Neural Networks, Vol. 2, Page(s): 1714-1719, 2002.

[6] M. Yasin and A. Awan, "A Study of Host-Based IDS using System Calls", Proceedings of the International Conference on Networking and Communication 2004, Page(s): 36- 41, June 2004.

[7] Larry J. Hughes, Jr. Actually Useful Internet Security Techniques, New Riders Publishing, Indianapolis, IN, 1995.

[8] R. Heady, G. Luger, A. Maccabe, and B. Mukherjee. A Method To Detect Intrusive Activity in a Networked Environment. In *Proceedings of the 14th National Computer Security Conference*, pages 362-371, October 1991.

[9] M. Gad-El-Rab, A. Abou El Kalam, & Y. Deswarte, "Defining Categories to Select Representative Attack Test-Cases", In Proceedings of ACM Workshop on Quality of Protection, Alexandria VA, USA. 2007.

[10] S. Smaha, "Haystack: An intrusion detection system", Proceedings of the 14th Conference on Aerospace Computer Security Applications, Page(s): 37–44, 1988.

[11] S. Kumar, "Classification and Detection of Computer Intrusions", Ph.D.Thesis, Purdue University.

[12] K. S. Killourhy, R. A. Maxion, & K. M. Tan, "A Defense-Centric Taxonomy Based on Attack Manifestations",

In Proceedings of International Conference on Dependable Systems and Networks (DSN'04), pp. 102-111, 2004.
[13] H. Takabi, J. B. Joshi, and G. Ahn, “Security and privacy challenges in cloud computing environments,” IEEE Security & Privacy, vol. 8, no. 6, pp. 24–31, Dec. 2010.
[14] Z. Duan, P. Chen, F. Sanchez, Y. Dong, M. Stephenson, and J. Barker, “Detecting spam zombies by monitoring outgoing messages,” IEEE Trans. Dependable and Secure Computing, vol. 9, no. 2, pp. 198–210, Apr. 2012. IEEE TRANSACTIONS ON DEPEDABLE AND SECURE COMPUTING
[15] C.-J. Chung, P. Khatkar, T. Xing, J. Lee, and D. Huang, “NICE: network intrusion detection and countermeasure selection in virtual network systems,” IEEE Transactions on Dependable and Secure Computing, vol. 10, no. 4, pp. 198–211, Jul. 2013.
[16] P. Salvador, A. Nogueira, U. Franca, and R. Valadas, “Framework for zombie detection using neural networks,” in Fourth International Conference on Internet Monitoring and Protection, 2009. ICIMP ’09, 2009, pp. 14–20.