International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
***Article Received:** 10 September 2023 **Revised:** 20 October 2023 **Accepted:** 30 October 2023*

Improved Feature Selection Algorithm for Intrusion Detection Using Data Mining Approach

Remya Raj B[1], Dr. R. Suganya[2]
[1]Ph.D Research Scholar, Department of Computer Science,
Maruthu Pandiyar College, affiliated to Bharathidasan University
Vallam, Thanjavur, Tamilnadu
remyarajphd2021@gmail.com
[2]HOD & Assistant Professor, Department of Information Technology
Bon Secours College for Woman, affiliated to Bharathidasan University
Thanjavur, Tamilnadu
mailtosuha@gmail.com

Abstract— With the rapid growth of Internet applications, there are more and more intrusions into network systems. In this case, it is necessary to provide security for the network through effective intrusion detection and prevention methods. This can mainly be achieved by creating effective interruption detection systems using efficient algorithms that can identify abnormal activities in network traffic and safeguard network resources from unlawful attack by interlopers. Although many interruption recognition frameworks have been proposed before, existing network intrusion detection has limitations in terms of accuracy and detection time. To overcome these shortcomings, In this paper, we propose a new intrusion detection system by developing a new intelligent feature selection algorithm based on conditional random fields (CRF) to optimize the number of features. Furthermore, algorithms based on existing hierarchical methods (LA) In this paper, we propose another interrupt recognition framework, fostering a book. Compared with the existing methods, the interruption identification framework provides high precision and achieves the efficiency of attack detection. The main advantages of this system are reduced detection time, improved classification accuracy and lower false positive rate.

Keywords- Intrusion detection system, feature selection, false alarms, LA, intelligent CR.

I. INTRODUCTION

Intrusion detection is necessary in today's computing environment because it is impossible to stay aware of current and potential threats and vulnerabilities in computer systems. The networking environment Constantly evolving and changing due to advances in network and Internet technology. To make matters worse, threats and vulnerabilities in the environment is also constantly evolving. a system for detecting intrusions can be utilized to help in managing threats and vulnerabilities in system. Threats occur due to people or groups who have the potential to compromise system. Moreover, the hackers have become a serious threat to many companies in the software field and those in other fields also suffer from this problem. Vulnerabilities are weaknesses in the systems, which are exploited by the hackers to compromise the system. New vulnerabilities are introduced every time when the technology develops and hence brings a new technology, product, or system brings with it a new generation of bugs and unintended conflicts arise. Moreover, the possible impacts from exploiting these vulnerabilities are constantly evolving. An intrusion may cause production downtime, sabotage of critical information, and theft of confidential information, cash, or other assets.

Different methods of attack are listed below: Masquerading: It includes using a stolen username/password or sending a TCP packet with a forged source address. Abuse of Feature: Includes filling up a disk partition with user files or starting hundreds of telnet connections to a host to fill its process table. A bug in a trusted program might allow an attack to proceed. An attacker can exploit errors in security policy configuration that allows the attacker to operate at a higher level of privilege than intended. Social Engineering: An attacker may be able to coerce a human operator it tends to be utilized to help into giving the attacker access. A bug in the execution of the TCP stack on some systems makes it possible to crash the system by sending it a carefully constructed malformed TCP packet. In feature selection in IDS also known as variable selection, attribute selection, or variable subset selection, it is the process of deciding on a relevant subset of features to use to build a model. The central assumption while utilizing an element determination strategy is that the information contains many superfluous or unnecessary characteristics. Redundant features are those which provide no more information than the currently selected features, and irrelevant function does not provide useful information under any circumstances. Include choice procedures are a subset of the more general field of feature extraction. Highlight extraction makes new elements from the qualities of the first elements, Whereas the contains option returns the highlighted subset.

II. LITERATURE SURVEY

Wherever Times is specified, Times Roman or Times New Roman may be used. If neither is available on your word processor, please use the font closest in appearance to Times. Avoid using bit-mapped fonts if possible. True-Type 1 or Open Type fonts are preferred. Please embed symbol fonts, as well, for math, etc.

Deepa V. Guleria et al [1]. These systems, they say, are inefficient and suffer from a high number of false positives. A portion of the normal goes after, for example, DoS,R2L , Probe and U2R affect network resources. Intrusion detection systems are challenged to reliably detect malicious activity and must be able to operate effectively with high volumes of organizational traffic.

Sannasi Ganapathy et al [2]. They describe the development of an efficient intrusion detection system that uses effective algorithms Identify unusual activity in network traffic and protect network resources from illegal infiltration by intruders. Although many interruption recognition frameworks have been proposed previously, existing network intrusion detection has limitations in terms of accuracy and detection time.

Osamah Mohammed Fadhil et al [3]. They described Intrusion detection systems are used to identify and prevent attacks on networks and databases. Harsh Set Trait Decrease Calculation is one of the major theories used for successfully reducing the attributes by removing redundancies. They describe the algorithm for selecting the fewest number of attributes from a KDD dataset.

Hai Thanh Nguyen et al [4]. They Described as playing from top to bottom examination of two fundamental estimates utilized in the channel model: the correlationfeature-selection measure and the minimalredundancy-maximal-relevance (mRMR) measure. We showed that the measures can be intertwined and summed up into a conventional element determination (GeFS) measure. The new approach is based on solved a mixed 0-1 linear programming problem (M01LP) by using the branch and-bound algorithm.

Gotam Singh Lalotra et al [5]. They described An Intelligent Conditional Random Field (ICRF)-based Cuttlefish Feature Selection Algorithm (ICRFCFA) is proposed for efficient decision-making on medical datasets. The proposed highlight determination calculation helps to further improve the expected accuracy faster. Future works in this direction could be the introduction of new rules for effective feature selection.

III. METHODOLOGY

A. *Feature Selection Module*

This module consists of an element choice agent and the data set. The feature selection module collects the data from KDD cup’99 a data set which has 41 features.

• Feature Selection Agent: The feature selection agent selects ideal combination of characteristics from these 41 features based on the rules present in knowledge base.

•Knowledge Base: The knowledge base contains the properties of all the features as facts. In addition to that, it is capable of adding new rules that are generated by the CRF model during training. The feature selection agent selects appropriate features by applying the suitable rules from the rules present in the knowledge base. The fundamental benefit of this feature selection module is that it selects ideal combination of characteristics from the KDD’99 cup set. It contains various rules to characterize the information. These rules are formed based on the LA during training. The guidelines are used to determine the normal data and attacks. Moreover, the rules are useful for classifying the attacks that are detected during testing.

B. *Intrusion Detection Module*

This module comprises of two significant parts namely training agent and decision making agent. The training agent is responsible for framing layers for Probe, DoS, R2L and U2R attacks. The decision making agent is capable of making decision by testing the data and applying rules. The outputs of this module are either normal or attacks. In case of attacks, they are classified as Probe, DoS, R2L and U2R attacks.

C. Training Agent:

This agent trains the data using the LA based on dataset with reduced features. Moreover, the training agent forms the classification rules which will be stored in the knowledge base. In the LA, four layers are considered for identifying four types of attacks.

***Article Received:** 10 September 2023 **Revised:** 20 October 2023 **Accepted:** 30 October 2023*

D. Decision Making Agent:

The decision making agent is responsible for performing the testing by classifying the data using rules selected from the knowledge base. These standards are produced during the preparation stage.using Intelligent Conditional Random Fields (ICRF). This ICRF uses LA to differentiate between normal registrations and four types of attacks, namely Probe, DoS, U2R and R2L. In order to fire the rules effectively, the decision making agent performs rule matching and uses forward chaining inference mechanism for compelling independent direction

E. CRF for Intrusion Detection

A CRF is a probabilistic system for modeling conditional distributions of random variables in any order. Furthermore, CRF is a fair, undirected graphical model that can be utilized to perform grouping naming.

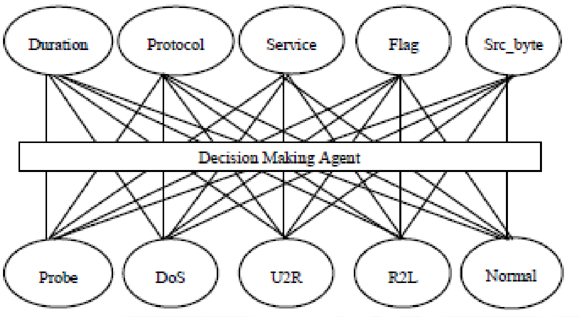


Figure 1.Graphical Representation Of CRF

Let *Xi* be a set of random variables over data sequence to be labeled and *Yi* be the corresponding label sequence, with *i*=1, … *n*. Let *G*= (*V*, *E*) be a graph such that *Y*= *Yi* ™(*V*), so that *Y* is indexed by the vertices of *G*. Then, (*Xi*, *Yi*) is a CRF, when conditioned on *Xi*, the random variables *Yi* obey the Markov characteristic in regard to the graph:

$$P(Y_i \mid X_i, Y_j, j \neq i) = P(Y_i \mid X_i, Y_j, i \sim j)$$

Where *i~j* means that *i* and *j* are neighbours in *G* i.e., a CRF is a random field globally conditioned on *Xi*. The CRF are given by the relation.

$$P_c(Y_i \mid X_j) \alpha \exp\left(\beta_k f_k(e, y \mid e, X_i)\right) + \Sigma \gamma_k g_k(i, y \mid i, X_j)$$

Here, *Xi* is the data sequence, *Yi* is a label sequence. Then, the features *fk* and *gk* are selected by the user. For example, a If the observation Xi, the boolean edge feature fk might be true. is tcp which is returned by the decision agent. The tag *Yi*-1 is “normal” and tag *Yi* is “normal.” Similarly, a vertex feature *gk* if the observation is true, *Xi* is “service=telnet” and tag *Yi* is “attack”.

F. CRF Based Feature Selection Algorithm

In this work, the new ICRFFSA automates feature selection by extending existing CRF-based feature selection algorithms, where we randomly select features for each layer. Each layer is separately prepared to recognize a solitary kind of assault classification like DoS, Probe, U2R and R2L. Contribution values are assigned here for all features in that layer. Based on this cumulative contribution value, we set the threshold to find the exact features for all type of attacks. Selected features are stored in the set *F*. The decision agent takes a decision to select that feature to find the particular attack based on the cumulative contribution value of each feature by applying rules. If the particular feature cumulative contribution value is greater than threshold then, agent chooses the feature for identifying the particular attack

G. Algorithm

```
Intelligent CRF based feature selection.
Input: The set S of all features
Output: F, the set of optimal features
// Let A be the set of features
Begin
F={ }; // Initialize F to all null set.}
for i=1 to n do
Begin
for j=1 to n do
Begin
f=random(S, CRF(s)) //Feature Selection
CV=CV+Cond.prob(fi)//contributed value
D=DA(CV, Decision)
if decision==“yes” then F=F∪(fj)
Val=Check (CV >Threshold(Ai)) and
Constraints (i, j))
if (val==true)
Display (Ai, j, Features(S));
Prevent (Ai, j);
Else
Stop
End
End
End
```

H. Classification Algorithm Using LA

In this work,we integrated the suggested feature choice algorithm called ICRFFSA with the currently used classification algorithm known as LA classifier to perform effective classification. This proposed algorithm receives the trained data with reduced features from the The element determination calculation is approved against the rules and facts present in the knowledge base. Four types of attacks are identified in this A model in view of the principles present in the knowledge base. After identifying the attackers, this classifier also finds the types of attacks

I. MATLAB

MATLAB (Matrix Laboratory) is a fourth-generation programming language and numerical computing environment. Developed by Math Works, MATLAB supports matrix manipulation, plotting functions and data, implementing algorithms, creating user interfaces, and collaborating Projects written in different dialects, including C, C++, Java, and Fortran.

IV. RESULT AND DISCUSSION

A.Description of Kdd'99 Data Set

Benchmark KDD cup 99 Intrusion Detection informational collection is utilized for experiments [3]. The dataset was a collection of simulated raw TCP dump data on a neighborhood. The KDD 99 data set Contains approximately 5 million connected records from training data and approximately 2 million connected records from test data. In our trials, we utilize 10% of the complete preparation information and 10% of the test information (with revised names) gave independently. This leads to 494,020 training and 311, 029 test instances.

B. Feature Selection

Relating to the four assault gatherings (Probe, DoS, R2L and U2R) provided in the KDD 99 dataset and other attacks, We choose different functions for different layers according to the sort of assault detected by the training layer. Thus, we have four separate modules relating to the four assault gatherings, and a fifth module trained on additional assaults not present in the four assault gatherings in the training dataset. We choose different functions to train different layers in the framework. Therefore, we use domain knowledge to select functions for the four attack categories. We now describe why some features are chosen over others in each layer of the layer framework.

C. Probing Attack:

An attacker attempts to scan a network to accumulate data about a computer network or to discover known vulnerabilities, with the apparent purpose of circumventing its security controls. Examples include port scanning, satan, ipsweep, nmap.

D. Denial of Service Attack (DoS):

In this category, an attacker makes some computing or memory resource too busy or full to handle legitimate requests, or denies legitimate users access to the machine. I. Gram. Smurfs, teardrops, earth, back, neptune, pods.

E. Remote to Local Attack (R2L):

When an attacker with no record on the remote computer sends packets over the network to that machine and exploits certain vulnerabilities to acquire nearby access as a client of that machine.eg spy, warezclient, warezmaster, write to ftp, guess password.

F. User to Root Attack (U2R):

In this case, the attacker first gains admittance to an ordinary client account and then exploits certain vulnerabilities to acquire root admittance to the framework.. For example perl, rootkits, buffer_overflow.

G. Other attacks:

These are attacks not present in above four classes. e.g, snmpgetattack, mailbomb, snmpguess ,mscan

The Benchmark KDD' 99 intrusion data set is used for experiments [3]. We use 10% of the total training data and 10% of the test information (with corrected labels), which are provided separately to the system. For our results, we present the precision, recovery and F-value. They are defined as follows:

$$\boldsymbol{Precision = number\ of\ True\ Positives\ /\ number\ of\ True\ Positives + number\ of\ False\ Positive}$$

$$\boldsymbol{Recall = number\ of\ True\ Positives\ \ /\ number\ of\ True\ Positives + number\ of\ False\ Negative}$$

where TP, FP, and FN are the number of True Positives, Positive and negative false alarms, respectively, and corresponds to the relative importance of precision versus recall and is usually set to 1. We divide the training and testing data into different groups; Normal, Probe, DoS, R2L, and U2R. We conduct separate experiments on five attack categories by randomly selecting data corresponding to a specific attack category and normal data only. Therefore, for the five attack categories, we form five independent models respectively through feature selection.

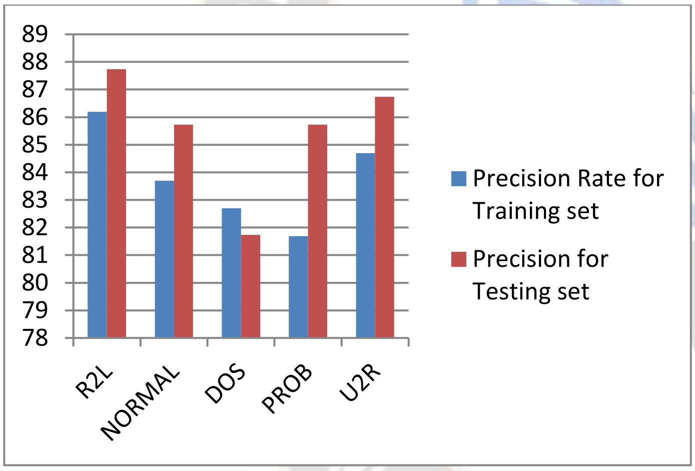
Table 1. Training set performance of Intrusion attack

	Precision Rate	**Recall Rate**
R2L	86.1937	81.6937
NORMAL	83.6937	82.6937
DOS	82.6937	80.6937
PRO B	81.6937	83.6937
U2R	84.6937	83.6937

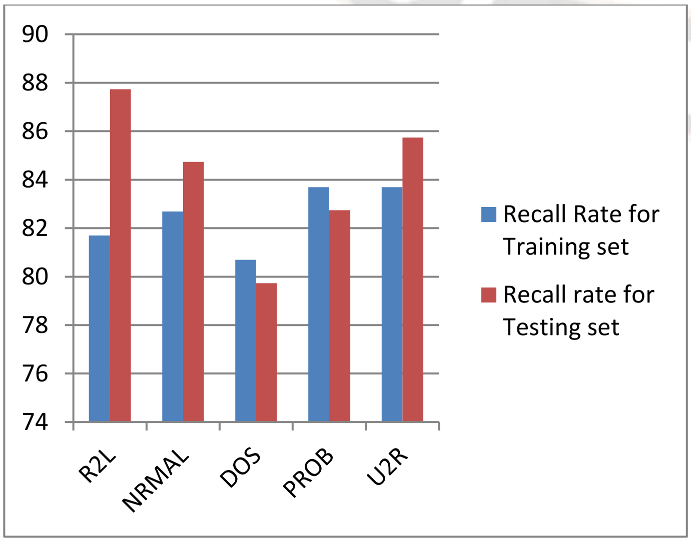
Table 2. Testing set performance of Intrusion attack

	Precision Rate	**Recall Rate**
R2L	87.7336	87.7336
NORMAL	85.7336	84.7336
DOS	81.7336	79.7336
PRO B	85.7336	82.7336
U2R	86.7336	85.7336

A. *Precision Rate*



B. *Recall Rate*



V. CONCLUSION

The new interruption location framework is improves the detection accuracy and time efficiency for building the systems for detecting intrusions. For this purpose, we proposed a the LAICRF model is created in light of the classification algorithm combining ICRFFSA and LA to efficiently identify intrusions. In this work, rule and LA based classification methods have been used that significantly reduce the detection time and hence it increases the detection accuracy. A system for identifying interruptions This document proposes to identify new Internet attacks. Furthermore, another steady component choice calculation is additionally proposed and carried out for successful element determination. The recommended include determination strategy is the Cuttlefish Component Choice calculation in combination and the Extended Chi-square algorithm. The experimental result shows the viability of the suggested system which is achieved detection accuracy in all types of attacks

REFERENCES

[1] Deepa V. Guleria and Chavan M.K, “Intrusion Detection System Based on Conditional Random Fields”, IJCSNS International Journal of Computer Science and Network Security,2013.

[2] Sannasi Ganapathy , Pandi Vijayakumar , Palanichamy Yogesh , and Arputharaj Kannan, “An Intelligent CRF Based Feature Selection for Effective Intrusion Detection”, The International Arab Journal of Information Technology,2015.

[3] Osamah Mohammed Fadhil, “Fuzzy Rough Set based Feature Selection and Enhanced KNN Classifier for Intrusion Detection”,s Journal of Kerbala University,2016.

[4] Hai Thanh Nguyen, Katrin Franke and Slobodan Petrovi´c, “Towards a Generic Feature-Selection Measure for Intrusion Detection”, International Conference on Pattern Recognition,2010.

[5] Gotam Singh Lalotra and R.S.Thakur, “An Intelligent CRF Based Cuttlefish Feature Selection Algorithm For Effective Diagnosis”, International Journal of Pharmacy & Technology,2016.

[6] Yuk Ying Chung and Noorhaniza Wahid, “A hybrid network intrusion detection system using simplified swarm optimization (SSO)”, Applied Soft Computing, Elsevier, vol. 12, pg. 3014-3022, 2012.

[7] Fangjun Kuang, Weihong Xu and Siyang Zhang, “A novel hybrid KPCA and SVM with GA model for intrusion detection”, Applied Soft Computing, Elsevier, vol. 18, pg. 178-184, 2014.

[8] Mohammed A. Ambusaidi, Xiangjian He, Priyadarsi Nanda and Zhiuan Tan, “Building an intrusion detection system using a filter-based feature selection algorithm”, IEEE Transactions on Computers, 2014.

[9] Aikaterini Mitrokotsa and Christos Dimitrakakis, “Intrusion detection in MANET using classification algorithms: The effects of cost and model selection”, Ad Hoc Networks, Elsevier, vol. 11, pg. 226-237, 2013.

Article Received: *10 September 2023* ***Revised:*** *20 October 2023* ***Accepted:*** *30 October 2023*

[10] Seung-Ho Kang and Naju, "A Feature Selection algorithm to find optimal feature subsets for Detecting DoS attacks" IEEE Conference of Decision Making, pp. 12-17, 2015.

[11] Yang Yi, Jiansheng Wu and Wei Xu, "Incremental SVM based on reserved set for network intrusion detection", Expert Systems with Applications, Elsevier, vol. 38, pg. 7698-7707, 2011.

[12] Veronica Bolon-Canedo, Diego Fernandez-Francos, Diego Peteiro-Barral, Amparo Alonso-Betanzos, Bertha Guijarro-Berdinas and Noelia Sanchez-Marono, "A unified pipeline for online feature selection and classification", Expert Systems with Applications, Elsevier, vol. 55, pg. 532-545, 2016.

[13] Shih-Wei Lin, Kuo-Ching Ying, Chou-Yuvan Lee and Zne-Jung Lee, "An intelligent algorithm with feature selection and decision rules applied to anomaly intrusion detection" Applied Soft Computing, Elsevier, vol. 12, pg. 3285-3290, 2012.

[14] Abdulla Amin Aburomman and Mamun Bin Ibne Reaz, "A novel SVM-KNN-PSO ensemble method for intrusion system", Applied Soft Computing, Elsevier, vol. 38, pg. 360-372, 2006.

[15] Ganapathy S., Rajesh Kambattan K., Veerapandian N. and Pasupathy M, "An Intelligent Intrusion Detection System model for MANET's based on Hybrid Feature Selection", Artificial Intelligent Systems and Machine Learning, CiiT, vol. 3, pg. 13, 2011.

[16] Rajesh Kambattan K. and Manimegalai R, "An Effective Intrusion Detection System using CRF based Cuttlefish Feature selection algorithm and MSVM", Asian Journal of Information Technology, vol. 15, pg. 891-895, 2016.