IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

An Effective Intrusion Detection System using CRF based Cuttlefish Feature Selection Algorithm and MSVM

A. Baby[1], Dr. S. Ravichandran Ph.D.,[2]

M.Phil Research Scholar, Computer Applications Dept, H.H The Rajah's College (Autonomous), Pudukkottai, India[1]

Head, Dept. of Computer Applications, H.H. The Rajah's College (Autonomous), Pudukkottai, India[2]

Abstract: In this we propose an effective intrusion detection system for improving the detection accuracy. In this proposed system, we propose a new feature selection algorithm called enhanced cuttlefish feature selection algorithm (ECFSA) for effective feature selection and Intelligent Agent based Enhanced Multiclass Support Vector Machine (IAEMSVM) classification algorithm is used for classification. The experimental results of the proposed system show that this system produced high-detection rate when tested with KDD cup 99 dataset.

Keywords: CRF – Conditional Random Field, CuttleFish Feature Selection, Multiclass Support Vector Machine.

I. INTRODUCTION

Intrusion detection is needed in today's computing environment because it is impossible to keep pace with the current and potential threats and vulnerabilities in computing systems. The networking environment is constantly evolving and changing due to the advances in web and internet technologies. To make matters worse, threats and vulnerabilities in the environment is also constantly evolving. An intrusion detection system can be used to assist 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. An intrusion may cause production downtime, sabotage of critical information, and theft of confidential information, cash, or other assets. It is these evaluation metrics which distinguish between the three main categories of feature selection algorithms: wrappers, filters and embedded methods. Other popular approach is the Recursive Feature Elimination algorithm, commonly used with Support Vector Machines to repeatedly construct a model and remove features with low weights.

II. LITERATURE SURVEY

Deepa V. Guleria et al [1]. They described systems are inefficient and suffer from a large number of false alarms. Some of the common attacks such as DoS, R2L, Probe and U2R affect the network resources. Intrusion detection system has challenges to detect malicious activities reliably and should able to perform efficiently with large amount of network traffic.

Sannasi Ganapathy et al [2]. They described developing efficient intrusion detecting systems that use efficient algorithms which can identify the abnormal activities in the network traffic and protect the network resources from illegal penetrations by intruders significantly reduce the detection time and hence it increases the detection accuracy.

Ambusaidi et al [6], filter based feature selection could handle linearly and nonlinearly dependant data features. Classification is done by SVM classifier. Though ANN used to detect attacks in IDS but provide the less accuracy due to its design to solve this ICA was used to fuse the complex intrusion input and hence attain renowned characteristics (that is, self-determining components, ICs) about the original data. By the use of ICs, the intricate of the ANN structure design could be condensed. Then, the PSO was employed to optimize the structural parameters of the ANN. Adel Sabry Eesa, Zeynep Orman,, uses the cuttlefish algorithm (CFA) as a search tactic to determine the best subset of features and the decision tree (DT) classifier as a judgment on the selected features that are produced by the CFA.

This dealt about the proposed system. In this section, we have discussed in detail about cuttlefish algorithm (CFA). The adjacent two intervals of the set could be merged, that $\chi 2$ value is computed from the adjacent two intervals and the threshold value difference is also greater than other $\chi 2$ value. When two adjacent intervals have a maximal difference in the calculated $\chi 2$ value and threshold should be merged first.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

III. PROBLEM DESCRIPTION

This dealt about the proposed system. In this section, we have discussed in detail about cuttle fish algorithm (CFA). Feature Selection We have used the existing cuttle fish algorithm and Chi-square for effective feature selection and classification. This cuttle fish algorithm is supporting dynamic decision over the feature selection process based on the environments and Extended chi-square algorithm to check the inconsistency level derived from previous level and allow the most relevant features to next level. The adjacent two intervals of the set could be merged, that $\chi 2$ value is computed from the adjacent two intervals and the threshold value difference is also greater than other $\chi 2$ value. When two adjacent intervals have a maximal difference in the calculated $\chi 2$ value and threshold should be merged first.We have used the existing and efficent classification algorithm called Intelligent Agent based Multiclass Support Vector Machine (IAEMSVM) algorithm for effective classification. This technique uses the clustering technique, intelligent agent and decision tree for improving the classification accuracy.

IV. METHODOLOGY

A. Feature Selection

To trace user policy violations to monitor the network events for intrusion, IDS have to process large amount of data in real time. This process is done until the stopping criteria are met. A feature selection algorithm can be seen as the combination of a search technique for proposing new feature subsets, along with an evaluation measure which scores the different feature subsets.

B. Intrusion Detection Module

This module consists of two major components 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.

• 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.
• 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 rules are generated using the Intelligent Conditional Random Field (ICRF) during the training phase. This ICRF uses a LA for distinguishing the normal records and the 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 effective decision making.

C. CRF for Intrusion Detection

CRF are a type of probabilistic system that is used to model the conditional distribution of random variables of any order. Moreover, a CRF is an unbiased and undirected graphical model that can be used to perform sequence labeling.

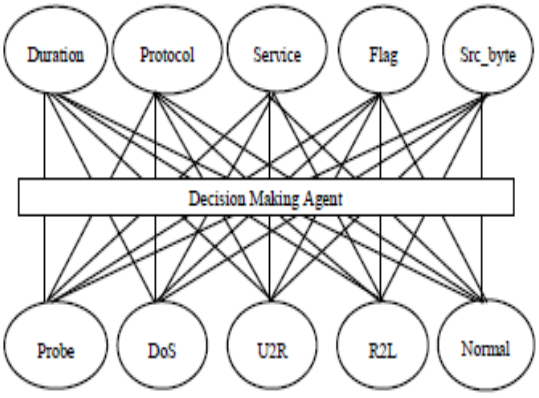


Figure I. Graphical representation of CRF

Let Xi be a set of random variables over data sequence to be labelled and Yi be the corresponding label sequence, with i=1, … n. Let G= (V, E) be a graph such that Y= Yi^{TM}(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 property with respect to the graph:

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

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

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.

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

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 boolean edge feature fk might be true if the observation Xi 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 is true if the observation Xi is "service=telnet" and tag Yi is "attack".

D.CRF Based Feature Selection Algorithm

Feature selection is the process of selecting appropriate features from the underlying data set such as KDD'99 cup data set for building models. In the CRF based feature selection algorithm, each feature is added to class of values depending on their dependency information. Therefore, to improve the efficiency of feature selection, we propose an intelligent agent and CRF based feature selection in this paper.

In this paper, a new ICRFFSA to perform feature selection automatically by extending the existing CRF based feature selection algorithm in which we select features for every layer randomly. Every layer is individually trained to detect a single type of attack category 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.

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

E.Classification Algorithm using LA

In this paper, we integrated the proposed feature selection algorithm called ICRFFSA with the existing classification algorithm known as LA classifier to perform effective classification. This proposed algorithm receives the trained data with reduced features from the feature selection algorithm and they are validated based on the rules and facts present in the knowledge base. Four types of attacks are identified in this model based on the rules present in the knowledgebase. After identifying the attackers, this classifier also finds the types of attacks.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

F. Algorithm Used

Intelligent CRF based Cuttlefish Feature Selection Algorithm Input: Datasets Output: Selected Features Step 1: Initialize the population (features) with random subset. Step 2: Evaluate fitness of the population using EMSVM. Step 3: Store the best subset in B. Step 4: Remove one feature from B using ICRF [3]. Step 5: Sort the original features in descending order based on the fitness value which is calculated according to [4]. Step 6: Randomly selected features is split into two and store into a set.

The main steps of the CFA are:

Step1:Initialize the population with random solutions, calculate and keep the best solution and the average value of the best solutions.

Step2:Use interaction operator between chromatophores and iridophores cells, to produce a new solution based on the reflection and the visibility of pattern.

Step 3:Use iridophores cells operators to calculate new solutions based on the reflected light from the best solution and the visibility of matching pattern.

Step 4:Use leucophores cells operator to produce new solution by reflecting light from the area around the best solution and visibility of the pattern.

Step 5:Use leucophores cells operator in case 6 for random solution by reflecting incoming light.

Step 7:Find the Reflection subset from randomly selected set using ICRF.

Step 8:Find the Visibility set for removing the elements of R using ICRF.

Step 9:New subset is created by using the features of visibility and reflection.

Step 10:Evaluate the new subset using EMSVM.

Step 11:If the new subset is better than the set B then the current new subset is considered as B.

ICRFCFA selects the best subset from resulted features which are finalized by the evaluation process using fitness function. Ascending the feature subsets and selecting features using.

V.EXPERIMENTAL RESULTS

A. RESULTS AND DISCUSSION

The Benchmark KDD' 99 intrusion data set is used for experiments [3]. We use 10 percent of the total training data and 10 percent of the test data (with corrected labels), which are provided separately for system. For our results, we give the Precision, Recall, and F-Value. They are defined as follows:

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

$$Recall = number\ of\ True\ Positives\ \ /\ number\ of\ True\ Positives + numberof\ False\ Negative$$

where TP, FP, and FN are the number of True Positives, False Positives, and False Negatives, 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 perform experiments separately for all the five attack classes by randomly selecting data corresponding to that particular attack class and normal data only.

B. Detecting Probe Attacks with Feature Selection

For detecting probe attack 5 significant features are selected out of 41 features shown in appendix. After selecting these 5 features, we have formed the probe patterns by using CRF coding in Java programming language. For this purpose, we used the records from 10 percent KDD train data set which is of type 'Normal + Probe'.

C. Detecting DOS Attacks with Feature Selection

For detecting DoS attack 9 significant features are selected from appendix and formed the DoS patterns.After that, we tested it with 10 percent corrected KDD test data and old test data. Figure shows the DoS attack result.

D. Detecting R2L Attacks with Feature Selection

For detecting R2L attack 14 significant features are selected out of 41 features shown in appendix. After selecting these 14 features, we have formed the R2L patterns. For this purpose, we used the records from 10 percent KDD train data which is of type 'Normal +R2L'. After that, we tested it with 10 percent corrected KDD test data and old test data. Figure shows the R2L attack result.

E. Detecting U2R Attacks with Feature Selection

For detecting U2R attack we have selected 8 significant features out of 41 features shown in appendix. After selecting these 8 features, we have formed the U2R patterns.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

F. Detecting Other Attacks

For Other attacks, we selected features such as 'duration', 'protocol' and 'service requested', while we ignored features such as 'number of file creations'.

TABLE I SARSA

	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 II SARSA- RBF

	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

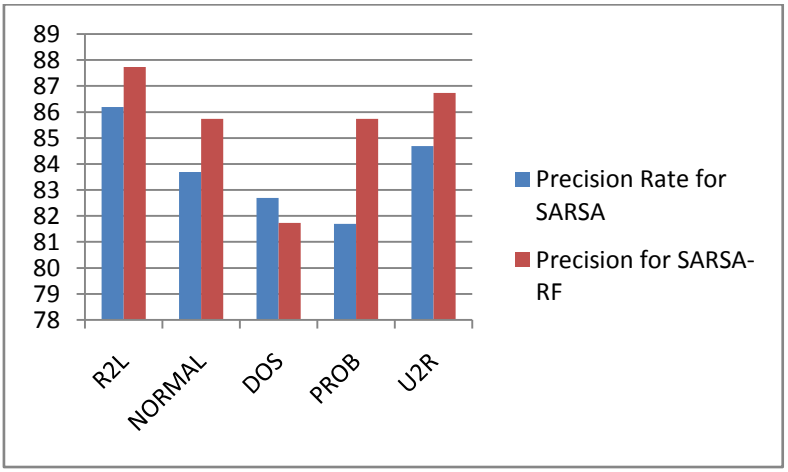
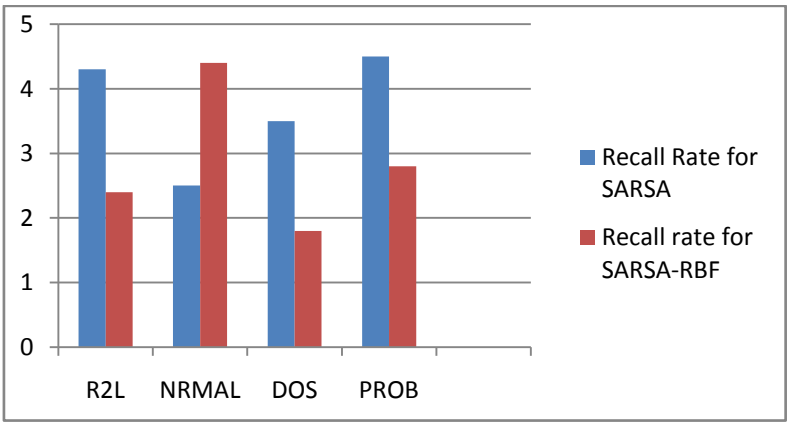


Figure II Precision Rate



VI. CONCLUSION

New intrusion detection system that improves the detection accuracy and time efficiency for building the intrusion detection systems.A intrusion detection system is proposed in this paper for detecting novel internet attacks. Moreover,

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

a new Incremental Feature Selection Algorithm (IFSA) is also proposed and implemented for effective feature selection. The proposed feature selection algorithm is the combination of Cuttle Fish Feature Selection algorithm and the Extended Chi-square algorithm. The experimental result shows the performance of the proposed 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", 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.
[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.
[17] Sannasi Ganapathy, Kanagasabai Kulothungan, Sannasy Muthurajkumar, Muthusamy Vijayalakshmi, Palanichamy Yogesh, Arputharaj Kannan, "Intelligent feature selection and classification techniques for intrusion detection in networks: a survey", EURASIP Journal on Wireless Communications and Networking, Vol. 2013, No.1, pp. 1-16, 2013.
[18] Wei-Chao Lin, Shih-Wen Ke and Shih-Fong Tsai, "CANN: An intrusion detection system based on combining cluster centers and nearest neighbors", Knowledge-Based Systems, Elsevier, vol. 78, pg. 13-21, 2015.
[19] Sindhu S., Geetha S., and Kannan A., "Decision Tree Based Light Weight Intrusion Detection Using a Wrapper Approach," Expert Systems with Applications, vol. 39, no. 1, pp. 129-141,2012.
[20] Wilk T. and Michal K., "Soft Computing Methods Applied to Combination of One-class Classifiers," Neuro Computing, vol. 75, no. 1, pp. 185-193, 2012.