International Journal of Soft Computing 8 (6): 400-405, 2013
ISSN: 1816-9503

An Intrusion Detection System for MANET using CRF Based Feature Selection and Temporal Association Rules

R.M. Somasundaram and K. Lakshmanan
Anna University, 600 025 Chennai, Tamilnadu, India

Abstract: As the Internet services spread all over the world, many kinds of security threats are introduced by malicious users. For the secured usage of the internet, the intrusion detection system plays a main role. Intrusion is an unauthorized access of the network resource by either a person or any software program. The role of the IDS is to analyze the network traffic and gives alerts about the attacks. In this study, researchers propose a new intrusion detection system using the temporal association rules for effective classification. Moreover, a new feature selection algorithm based on the Conditional Random Field (CRF) is used to improve the detection accuracy. The experimental results of the proposed model show that this system detects anomalies with low false alarm rate and high detection rate when tested with KDD Cup'99 dataset.

Key words: Intrusion Detection System (IDS), temporal association rules, Conditional Random Field (CRF), time intervals, world

INTRODUCTION

Intrusion Detection System is a software that can analyze the network traffic within the computer or in the network in order to find any illegal activities in it. The intrusion happens due to the unauthorized access that violates the network system. The process of identifying the intruders is known as intrusion detection. There are two types of intrusion detection techniques namely misuse detection and anomaly detection. In misuse detection, the system finds specific pattern or a behavior that matches with the well known intrusions which are stored in tables of databases. On the other hand, anomaly detection finds the intrusions by analyzing the deviations from the normal behavior such as number of times login failed. Another classification of intrusion detection systems are Host based Intrusion Detection System (HIDS) and Network based Intrusion Detection System (NIDS). The HIDS performs within the local system. The NIDS performs between the system and within the network.

Temporal association rule mining is a popular data mining techniques that considers time attributes to produce quality misused detection. However, due to the large number of network traffics and its attribute, it causes the large number of rules to present the pattern. These large number of rules will reduce the performance of the IDS. Therefore, this study proposes techniques to apply post mining technique to reduce the association rules by retaining the quality pattern.

A CRF is a probabilistic model that is useful to model the posterior distribution of a label sequence of data items which are conditioned on the observed data. Unlike other statistics methods which attempts to model how the observed data is generated to select the most appropriate label, a CRF is a discriminative model that uses attributes of the observed data to constrain the probabilities of the various labels that the observed data can receive (Gupta *et al*., 2010). A CRF defines a posterior probability P(x, y) of a label sequence y for a given input sequence x. In a linear chain conditional random field, the label for a given frame depends jointly on the label of the previous frame, the label of the succeeding frame and the observed data. These dependencies are computed in terms of functions defined by pairs of labels and by label-observation pairs. The input sequence corresponds to a series of frames of speech data while the label sequence is the series of labels assigned to that observed frame sequence. Each frame is assigned exactly one label in y.

Feature selection is the most critical step in building intrusion detection models (Namik and Othman, 2011; Gupta *et al*., 2010; Mabu *et al*., 2011). During this step, the set of attributes or features deemed to be the most effective attributes is extracted in order to construct suitable detection algorithms (detectors). A key problem that many researchers face is how to choose the optimal set of features as not all features are relevant to the learning algorithm and in some cases, irrelevant and redundant features can introduce noisy data that distract the learning algorithm, severely degrading the accuracy of

Corresponding Author: R.M. Somasundaram, Anna University, 600 025 Chennai, Tamilnadu, India

the detector and causing slow training and testing processes. Feature selection was proven to have a significant impact on the performance of the classifiers. Experiments by Alcala-Fdez *et al.* (2011) show that feature selection can reduce the building and testing time of a classifier by 50%. In this study, a new IDS is proposed for effective intrusion detection. For this purpose, a new temporal association rule mining algorithm is proposed by extending the Apriori Algorithm with new temporal conditions. Moreover, a CRF based feature selection algorithm is needed to select valuable attributes from the KDD Cup (1999) dataset.

Literature review: Mabu *et al.* (2011) proposed a GNP-based fuzzy class-association-rule mining with sub attribute utilization and the classifiers based on the extracted rules have been proposed which can consistently use and combine discrete and continuous attributes in a rule and efficiently extract many good rules for classification.

Gupta *et al.* (2010) have addressed the dual problem of accuracy and efficiency for building robust and efficient Intrusion Detection Systems. Their experimental results show that CRFs are very effective in improving the attack detection rate and decreasing the FAR.

Namik and Othman (2011) presented the important of post mining techniques in network intrusion detection. The use of Chi-square pruning techniques has reduced the number of rules of a misused detection up to 98% with preserving the same quality of knowledge above 90% confident. This research has shown the beneficial of applying post-mining in building a quality performance of IDS. The less number of rules the better performance of IDS. Alcala-Fdez *et al.* (2011) have proposed a new Fuzzy Associative Classification Method for high-dimensional datasets, named FARCHD. They have made use of a pattern weighting scheme in order to reduce the number of candidate rules, pre selecting the rules with the best quality. A genetic rule selection and lateral tuning have been applied to select a small set of fuzzy association rules with a high classification accuracy.

Morris and Fosler-Lussier (2008) explored the application of CRFs to combined local posterior estimates provided by Multilayer Perceptions (MLPs) corresponding to the frame-level prediction of phone classes and phonological attribute classes. They compared phonetic recognition using CRFs to an HMM system trained on the same input features and show that the monophone label CRF is able to achieve superior performance to a monophone-based HMM and performance comparable to a 16 Gaussian mixture triphone-based HMM in both of these cases, the CRF obtains these results with far fewer free parameters.

Stewart *et al.* (2008) presented a model that is capable of learning such structures using a random field of parameterized features. These features can be functions of arbitrary combinations of observations, labels and auxiliary hidden variables. Wu *et al.* (2010) proposed framework for intrusion detection in distinctly composed of two phases: training and detection phases. In the training phase, the system train and analyze data provided firstly. In the detection phase using a security mode established in the training phase as the standard model, the system analyzes the actual testing data and alarms and records the data security models which do not meet the standard.

Changguo *et al.* (2009) enhanced detection speed and accuracy in wireless network. This study adopts Fuzzy Association Rules Mining Methods and apriori algorithm to conduct anomaly detection experiments of wireless network. In their experiment, they improved the support degree of the two algorithms and made comparisons and analysis of the results. The method of support degree had improved and results were analyzed. Su *et al.* (2011) proposed system has adopted incremental mining of fuzzy association rules with genetic optimization on the membership functions. Experiments were conducted by them which demonstrated the effectiveness and efficiency of this intrusion detection system in detecting DoS attacks by an anomaly detection approach. All the research present in the literature focused only on association rule mining. However, this study proposes a temporal association rule mining approach for effective intrusion detection.

MATERIALS AND METHODS

System architecture: The architecture of the system proposed in this research consists of four major modules namely, user interface module, data preprocessing module, Intrusion Detection System module and Prevention module as shown in Fig. 1.

User interface module: The user interface module collects the network data from the network layer. This data are sent to the preprocessing module for preprocessing the data.

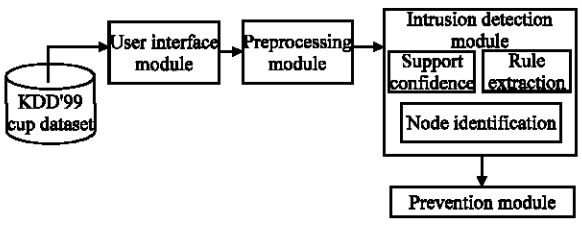

Fig. 1: System architecture

Preprocessing module: The data preprocessing module uses a preprocessing technique called feature selection for effective preprocessing. In this technique, the agent selects only the valuable attributes from the dataset using projection. Moreover, data cleaning, data integration and data transformation are carried out for performing effective preprocessing.

Intrusion detection module: This module detects the intruders from the given data using their minimum support, confidence and rule extraction. Malicious nodes are identified based on the support, confidence and the time duration between the records or nodes.

Prevention module: After identification of each and every malicious node is filtered from the data set as a separate group of data then it will be avoided for further transactions in the network.

Proposed research: In this study, researchers proposed a new intrusion detection model which is combination of temporal association rules. Moreover, researchers used an effective feature selection algorithm that is CRF based Feature selection for improving the detection accuracy.

Temporal association rule mining: Association rule discovery, the area being studied most actively (Han and Cho, 2006; Hu *et al*., 2008; Luo, 1999; Yu *et al*., 2006) is a technique to investigate the possibility of simultaneous occurrence of the data. Let $I = \{i_1, i_2, ..., i_m\}$ be a set of data items which are valid during the time interval $[t_1, t_2]$ where t_1 is the start time and t_2 is the end time.

Let D be a set of transactions where each transaction T is a set of items such that $T \in I$. A temporal association rule is an implication of the form $X \rightarrow Y$ where $X \in I$, $Y \in I$, $X \cap Y = \phi$ and it happens in $[t_1, t_2]$. Temporal support and confidence are used to describe the interestingness or goodness of a rule as follows:

$$\text{Temporal support } (X \rightarrow Y, TS[t_1, t_2]) = P(X \cup Y), TS[t_1, t_2]$$

$$\text{Temporal confidence } (X \rightarrow Y) = P(Y \mid X), TS[t_1, t_2]$$

Or:

$$\text{Confidence}(X \rightarrow Y, TS[t_1, t_2]) = \text{Support}(X \rightarrow Y)/P(x), TS[t_1, t_2]$$

where, TS stands for Time Stamp which uses interval stamping of tuples with t_1 as start time and t_2 as end time.

The temporal support of the rule $X \rightarrow Y$ is the proportion of transactions in D that contain $X \cup Y$ in the interval $[t_1, t_2]$. The temporal confidence of the association rule $X \rightarrow Y$ is the proportions of transactions including both X and Y to all the transactions that include X in $[t_1, t_2]$. The existing association rule discovery techniques have discovered the association that may happen only among the data that satisfy the minimum support and minimum confidence set by the users. However, they do not consider the time interval in which the rules are valid.

Temporal class association rule mining: The following is a statement of temporal association-rule mining (Shimada *et al*., 2006a, b). Let I = {A1, A2, ... , Al} be a set of attributes valid for an interval $[t_1, t_2]$. Let G be a set of tuples where each tuple T is a set of attributes such that $T \subseteq I$. Let TID be an ID number associated with each tuple. A tuple T contains X, a set of some attributes in I, if $X \subseteq T$ in $[t_1, t_2]$. A temporal association rule is an implication of the form $X \Rightarrow Y$ where $X \subset I$, $Y \subset I$ and $X \cap Y = \varnothing$ and happens in $[t_1, t_2]$. X is called antecedent and Y is called consequent of the rule. If the fraction of tuples containing X in G equals x, then researchers say that temporal support (X) = x with TS = $[t_1, t_2]$. The rule $X \Rightarrow Y$ has a measure of its strength called temporal confidence defined by support $(X \cup Y)$/support (X) during $[t_1, t]$. Based on these definitions, this study proposes the temporal apriori algorithm as follows:

Temporal class association rule mining algorithm: The pseudo code for the temporal apriori algorithm is given below for a transaction database T and a temporal support threshold of δ. Common set theoretic data is employed in this paper where T is a multiset. C_k is the candidate set for level k. Generate () algorithm is assumed to generate the selected features sets from KDD cup data set of the prior level, heeding the downward closure lemma. Count [c] accesses a field of the data structure that represents feature set C which is originally assumed to be zero. Many details are misplaced below, typically the most significant part of the implementation is the data structure used for storing the data sets and including their frequencies.

Temporal apriori algorithm:

```
Input          : Data set T
Output         : Temporal Association Rule
Apriori ( T, δ, Time stamp)
   L1 ← {First item sets}
   k ← 2
   t1 = Timestamp
   while Lk-1 ≠ ϕ and start time ≥t1 and end time ≤t2 do
      Ck ← {c|c = a U{b} and a∈Lk-1 and b∈ ULk-1 and b∈a}
        For transactions t∈T
        begin
        Ct ← {c|c∈Ck and t1≤t2}
        t1 = t1+1
         For candidates c∈Ct
         begin
            Count [c] ← count [c]+1
          end
            Lk ← {c|c∈Ck and count [c]≥δ}
            k ← k+1
         end
     return Lk
```

RESULTS AND DISCUSSION

Training and testing data: The data set used in the experiment was taken from the Third International Knowledge Discovery and Data Mining Tools Competition (KDD Cup, 1999; Shimada *et al*., 2006a). Each association record is described by 41 attributes. The list of attributes consists of both continuous and discrete type variables with statistical distributions varying significantly from each other which make the intrusion detection a very challenging task.

In this dataset, it has five million network connection records such as land attack, Neptune attack, password guess, port scan, etc. The 22 categories of attacks from the following four classes: DoS, R2L, U2R and Probe. These 41 features describe the basic information about the network packet, network traffic, host traffic and content information. Table 1 shows the name and serial number of the 41 features. Each record contains the five class labels such as normal, Probe, DoS, R2L and U2R. It has 391458 DoS attack records, 52 U2R attack records, 4107 Probe attack record, 1126 R2L attack records and 97278 normal records only in this 10% of this dataset.

Experimental results: This model has achieved the highest detection rates for old DoS, Probe and R2L attacks. The attribute selection algorithm provides the highest accuracy for all the known attacks namely DoS, Probe, U2R and R2L in preprocessing. Earlier algorithms achieved very low classification accuracy for U2R compared with other attacks. But the attribute selection algorithm achieves better detection accuracy for even U2R classes. This has been shown in the Table 2.

Table 1: The 41 features in KDD Cup'99 dataset

Features name	Features name
Duration [t_1, t_2]	is_guest_login
protocol_type	Count
service	serror_rate
src_byte	rerror_rate
dst_byte	same_srv_rate
flag	diff_srv_rate
land	srv_count
wrong_fragment	srv_serror_rate
urgent	srv_rerror_rate
hot	srv_diff_host_rate
num_failed_logins	dst_host_count
logged_in	dst_host_srv_count
num_compromised	dst_host_same_srv_count
root_shell	dst_host_diff_srv_count
su_attempted	dst_host_same_src_port_rate
num_root	dst_host_srv_diff_host_rate
num_file_creations	dst_host_serror_rate
num_shells	dst_host_srv_serror_rate
num_access_shells	dst_host_rerror_rate
num_outbound_cmds	dst_host_srv_rerror_rate
is_hot_login	

Table 2 shows that the selected features from the 41 features in this dataset. The agent based attribute selection algorithm selects these following 19 attributes from the dataset. Table 3 shows that the comparison of the detection accuracy of the Apriori algorithm and temporal association rule mining algorithm. From Table 3, it can be observed that the detection accuracy of temporal association rule mining algorithm is better than the existing apriori algorithm.

Table 4 shows that the support, confidence and χ^2-values are measured or calculated based on the different combinations of the features which are selected by the feature selection algorithm. By using these values it can be identified the effective feature set.

Table 5 shows that the comparison of the detection accuracy of few important algorithms and the current

Table 2: List of 19 selected features from 41 features

Features name	Features name
protocol_type	dst_host_count
src_byte	dst_host_srv_count
wrong_fragment hot	dst_host_same_srv_count
root_shell	dst_host_diff_srv_count
su_attempted	dst_host_same_src_port_rate
num_access_shells	dst_host_srv_diff_host_rate
rerror_rate, diff_srv_rate	dst_host_serror_rate
srv_serror_rate	dst_host_rerror_rate
srv_diff_host_rate	

Table 3: Comparison of the accuracy of the temporal association rule algorithms

	Similarity			
	Apriori algorithm		Temporal association rule	
Network traffic (sec)	Normal	Attacks	Normal	Attacks
300	0.890	0.000	0.12	0.00
600	0.570	0.000	0.87	0.00
1800	0.908	0.650	0.61	0.01
3600	0.801	0.115	0.95	0.08

Table 4: Support, confidence and χ^2-values for various features combinations during [t_1, t_2]

Combinations				
Feature 1	Feature 2	Support	Confidence	χ^2-value
Protocol type	Service	100.0	100.0	50.0
Protocol type	Flag	100.0	100.0	50.0
Protocol type	Scr bytes	12.0	12.0	54.0
Protocol type	Hot	8.0	8.0	56.5
Protocol type	Ishotlogin	40.0	40.0	50.7
Service	Flag	100.0	100.0	50.0
Service	Scr bytes	12.0	12.0	54.0
Service	Hot	8.0	8.0	56.5
Service	Ishotlogin	40.0	40.0	50.7
Flag	Scr bytes	12.0	12.0	54.0
Flag	Hot	8.0	8.0	56.5
Flag	Ishotlogin	40.0	40.0	50.7
Src_bytes	Hot	2.0	16.6	59.7
Src_bytes	Ishotlogin	8.0	66.6	54.0
Hot	Ishotlogin	8.0	100.0	55.7

Table 5: Measures of class association rules using machine learning techniques: DR (%) and FPR (%) for TS = [t_1, t_2]

Machine learning techniques	DR (%)	FPR (%)
C4.5	95.00	1.00
SVM	95.50	1.00
MLP	94.50	1.00
K-NN	92.00	1.00
K means clustering	65.00	1.00
Y means clustering	89.89	1.00
Genetic programming	91.00	0.43
Neural networks+PCA	92.22	-
C4.5+PCA	92.16	-
GA	97.47	0.69
C4.5+hybrid neural networks	93.28	0.20
Hidden Markov Model (HMM)	79.00	-
Temporal association rule + CRF based feature selection	96.71	3.29

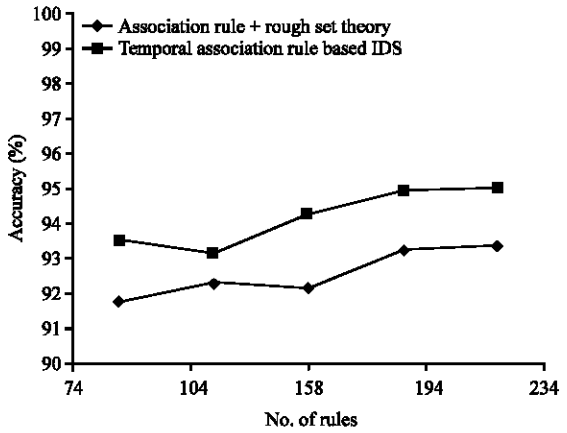

Fig. 2: Overall performance of the proposed IDS

intrusion detection system which is the combination of CRF based feature selection and Temporal Association Rule Mining algorithm.

Figure 2 shows that the comparison between the overall performance analysis of association rule mining intrusion detection system which is the combination of association rule using rough set theory and the proposed model which is the combination of CRF based feature selection and Temporal Association Rule Mining algorithm.

CONCLUSION

In this study, researchers proposed and implemented a new Intrusion Detection System using temporal association rules. Moreover, a new feature selection algorithm based on the Conditional Random Field (CRF) is used to improve the detection accuracy. The experimental results of the proposed model show that this system detects anomalies with low false alarm rate and high detection rate when tested with KDD Cup'99 dataset.

Further researches on this direction can be the proposal of a frequent pattern Tree based temporal association rule mining algorithm for improving the performance.

REFERENCES

Alcala-Fdez, J., R. Alcala and F. Herrera, 2011. A fuzzy association rule-based classification model for high-dimensional problems with genetic rule selection and lateral tuning. IEEE Trans. Fuzzy Syst., 19: 857-872.

Changguo, Y., Z. Qin, Z. Jingwei, W. Nianzhong, Z. Xiaorong and W. Tailei, 2009. Improvement of association rules mining algorithm in wireless network intrusion detection. Proceedings of the IEEE International Conference on Computational Intelligence and Natural Computing, June 6-7, 2009, Wuhan, China, pp: 413-416.

Gupta, K.K., B. Nath and R. Kotagiri, 2010. Layered approach using conditional random fields for intrusion detection. IEEE Trans. Dependable Secure Comput., 7: 35-49.

Han, S.J. and S.B. Cho, 2006. Evolutionary neural networks for anomaly detection based on the behavior of a program. IEEE Trans. Syst. Man Cybern. Part B: Cybern., 36: 559-570.

Hu, W., W. Hu and S. Maybank, 2008. AdaBoost-based algorithm for network intrusion detection. IEEE Trans. Syst. Man Cybern. Part B: Cybern., 38: 577-583.

KDD Cup, 1999. KDD cup'99: Computer network intrusion detection. http://www.kdd.org/kdd-cup-1999-computer-network-intrusion-detection.

Luo, J., 1999. Integrating fuzzy logic with data mining methods for intrusion detection. M.Sc. Thesis, Mississippi State University, Department of Computer Science.

Mabu, S., C. Chen, N. Lu, K. Shimada and K. Hirasawa, 2011. An intrusion-detection model based on fuzzy class-association-rule mining using genetic network programming. IEEE Trans. Syst. Man Cybern. Part C: Appl. Rev., 41: 130-139.

Morris, J. and E. Fosler-Lussier, 2008. Conditional random fields for integrating local discriminative classifiers. IEEE Trans. Audio Speech Lan. Process., 16: 617-628.

Namik, A.F. and Z.A. Othman, 2011. Reducing network intrusion detection association rules using Chi-squared pruning technique. Proceedings of the IEEE 3rd Conference on Data Mining and Optimization, June 28-29, 2011, Putrajaya, Malaysia, pp: 122-127.

Shimada, K., K. Hirasawa and J. Hu, 2006a. Genetic network programming with acquisition mechanisms of association rules. J. Adv. Comput. Intell. Inform., 10: 102-111.

Shimada, K., K. Hirasawa and J. Hu, 2006b. Class association rule mining with chi-squared test using genetic network programming. Proceedings of the IEEE International Conference on Systems Man and Cybernetics, October 8-11, 2006, Taipei, Taiwan, pp: 5338-5344.

Stewart, L., X. He and R.S. Zemel, 2008. Learning flexible features for conditional random fields. IEEE Trans. Pattern Anal. Mach. Intell., 30: 1415-1426.

Su, M.Y., C.Y. Lin, S.W. Chien and H.C. Hsu, 2011. Genetic-fuzzy association rules for network intrusion detection systems. Proceedings of the IEEE International Conference on Fuzzy Systems, June 27-30, 2011, Taipei, Taiwan, pp: 2046-2052.

Wu, K., J. Hao and C. Wang, 2010. Intrusion detection based on fuzzy association rules. Proceedings of the International Symposium on Intelligence Information Processing and Trusted Computing, October 28-29, 2010, Huanggang, China, pp: 200-203.

Yu, Z., J.J.P. Tsai and T. Weigert, 2006. An automatically tuning intrusion detection system. IEEE Trans. Syst. Man Cybern. Part B: Cybern., 37: 373-384.