International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 17 (2019) pp. 3616-3621
© Research India Publications. http://www.ripublication.com

Intrusion Detection using Rule Learning based Classifiers

Ashalata Panigrahi

Roland Institute of Technology, Berhampur, India.

Manas Ranjan Patra

Berhampur University, Berhampur, India.

Abstract

The tremendous improvements of technologies including various smart devices have made the use of computers easy for gathering and sharing information using the Internet. The various private and government organizations store valuable data over the network. The security of computer networks has become a crucial problem due to the importance and sensitivity of the information communicated. To defend against various cyber attacks and computer viruses, lots of computer security techniques such as data encryption, Virtual Private Network, proxies, avoiding programming errors, firewalls, antivirus have failed to protect networks and systems from increasingly sophisticated attacks and malwares. Intrusion Detection Systems (IDSs) are a major line of defense for protecting network resources from illegal penetrations. In this work, rule learning based techniques have been proposed for building an intrusion detection model using four classifiers, namely, Decision Table, Decision Table/Naïve Bayes (DTNB), JRip, and Ripple Down Rule Learner (RIDOR). Further, to improve the performance of the model, relevant features have been extracted from the dataset through preprocessing by applying three bio- inspired search techniques such as Ant search, Genetic search, and Particle Swarm Optimization search and two informed search techniques such as Best first search and Greedy stepwise search . The performance of the model has been evaluated on the NSL-KDD intrusion dataset in terms of accuracy, precision, detection rate or recall, false alarm rate and F-value.

Keywords: Genetic search, Ant search, Particle Swam Optimization, Decision table , RIDOR, DTNB.

1. INTRODUCTION

One of the biggest threats faced by internet based applications is different forms of cyber attacks. Traditional prevention techniques such as user authentication, data encryption, firewalls etc. are becoming less effective. The important goal of information security is to develop defensive information systems which are secure from unauthorized access, modification, or destruction. The goal of intrusion detection is to identify unauthorized use, misuse, and abuse of computer systems by both system insiders and external penetrators. Generally, IDSs are broadly classified into two categories namely misuse and anomaly detection systems. Misuse detection system detect known attacks. These methods are pre-coded with signatures of attacks and perform rule matching to detect intrusions. But these kinds of systems are not sufficient to detect new or unknown attacks. The anomaly based intrusion detection has the capacity to detect new types of intrusions without having any prior knowledge of them. Kavitha et al. [1] have proposed the Emerging Intuitionistic Fuzzy Classifier model for Intrusion Detection System and tested its performance on KDD-99 Intrusion dataset. The model generates the detection rules based on the intuitionistic fuzzy classifier. The performance of the Emerging Intuitionistic Fuzzy Rule for Intrusion Detection (EIFRID) outperforms the algorithms such as CTree, and FRIDS (Fuzzy Rules of Intrusion Detection System). Jain et al. [2] have proposed an intrusion detection method using information gain, NB and Bayes Net. They reduced the features of the dataset using information gain of the attributes. After feature reduction the data was analyzed using two learning algorithms, NB and Bayes Net. Bayes Net with an accuracy rate of approximately 99% was found to perform much better than NB in detecting intrusions. Bhavsar et al. [3] have proposed an Intrusion detection system using data mining techniques. Here classification was done by SVM. Results show that SVM can reduce the time required to build a model for classification and increase the intrusion detection accuracy when Gaussian RBF kernel is used. The experiment was conducted using 10-fold cross-validation and Gaussian RBF kernel of SVM. The attack detection accuracy achieved was 94.1857%. Atefi et al. [4] have proposed a hybrid model using SVM and GA(Genetic Algorithm). They compared true negative and false positive rates obtained by SVM and hybrid model SVM+GA. Hybrid model recorded low false negative rate of 0.5013% and high true negative value of 98.2194%. The result shows high accuracy of the hybrid model. Singh and Bansa (5) have proposed an intrusion detection method using different artificial neural network based classifier algorithms. They reduce the features of NSL-KDD dataset using feature selection methods. After feature reduction the data was analysed using four algorithms, namely, Multilayer Perception, Radial Base Function, Logistic Regression and Voted Perception. Results shown that Multilayer Perception neural network algorithm is providing more accurate results than other algorithms.

2. METHODOLOGY

2.1 Decision Table

Decision table classifiers [6] are of two types, viz., DTM (Decision Table Majority) and DTL (Decision Table local). DTM has two components, viz., a schema which is a set of features and a body that consists of a set of labeled instances. Given a set of unlabeled instances, a decision table classifier looks for exact matches in the decision table using only those features available in the schema. There could be many matching instances in the table where each instance consists of a value for each of the features in the schema and a value for the label. If no instances are found, the majority class of the DTM is identified else the majority class of all matching instances is identified. In order to build a DTM, the induction algorithm must decide which features are to be included in the schema and which instances need to be stored in the body.

2.2. Decision Table / Naïve Bayes (DTNB)

In case of Decision Table / Naïve Bayes classifier [7], at every point in the search, the algorithm estimates the value of dividing the features into two disjoint subsets: one for the decision table and the other for Naïve Bayes. Initially, all features are modeled by the decision table. Next, a forward selection search is used, and at each step the selected features are extracted by Naïve Bayes and rest of the features by the decision table. The algorithm iterates by dropping a feature from the model at each step. The overall class probability is generated by combining the individual class probability of DT and NB.

2.3. JRip

There are four phases in JRip algorithm [8], viz., Growth, Pruning, Optimization, and Selection. A sequence of individual rules is produced in the growth phase by adding predicates until a stopping criterion is satisfied by the rules. In the pruning phase, the rules that reduce the performance of the algorithm are pruned. In the third phase, each rule is optimized by adding attributes to the original rule or by generating new rules using the 1st and 2nd phases. In the final phase, the best rules are selected and the remaining rules are dropped from the model.

2.4. Ripple Down Rule Learner (RIDOR)

RIDOR is a knowledge acquisition and representation methodology [9] which learns rules with exceptions by generating default rules, and exceptions are generated by incremental reduced error pruning with smallest error rate. Next, the exceptions with minimum support are picked up and the best exceptions for each exception are found. The process is iterated several times. A ripple down rule refers to a list of rules where each rule can be associated with another ripple down rule, specifying exceptions. An expert can frame a new rule with impacts on its parent rules in a given context and can insert the new rule into the list.

3. PROPOSED RULE LEARNING BASED CLASSIFICATION MODEL

In this section, we present our proposed model for classifying intrusion data in order to build an efficient intrusion detection system which can exhibit low false alarm rate and high detection rate. The model consists of two major phases as depicted in figure1. In the first phase irrelevant and redundant features are removed using three bio- inspired search techniques such as Ant search, Genetic search, and Particle Swarm Optimization search and two informed search techniques such as Best first search and Greedy stepwise search. In the next phase, the reduced data set is classified using four rule learning classifiers, namely, Decision Table, Decision Table/Naïve Bayes (DTNB), JRip, and Ripple Down Rule Learner (RIDOR). Further, 10-fold cross validation technique used for training and testing of the model and the performance of the model is evaluated using certain standard criteria.

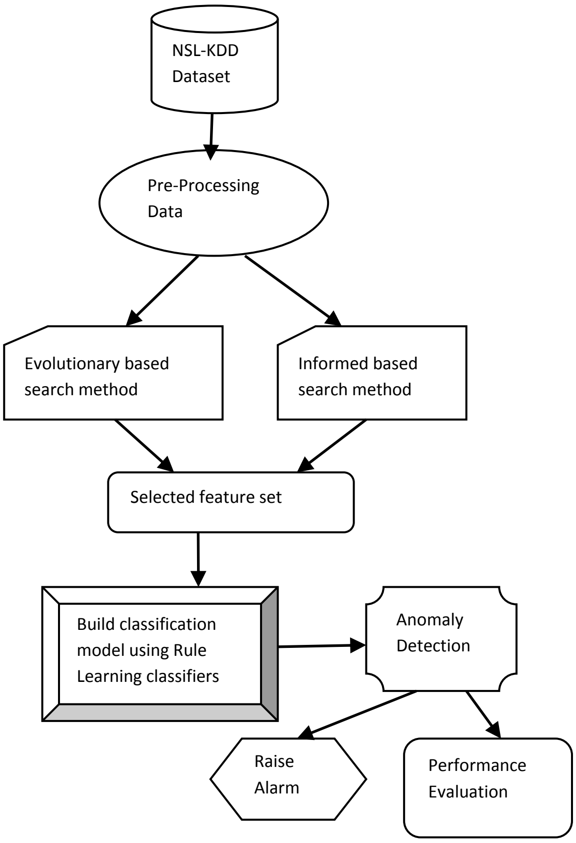
NSL-KDD Dataset
Pre-Processing Data
Evolutionary based search method
Informed based search method
Selected feature set
Build classification model using Rule Learning classifiers
Anomaly Detection
Raise Alarm
Performance Evaluation

Figure 1. Proposed Rule Learning based Classification Model

4. EXPERIMENTAL SETUP

4.1. NSL-KDD Dataset

NSL-KDD data set proposed by Tavallace et al. [10] is a reduced version of the original KDD 99 dataset and it has the same features as that of KDD 99. The number of records in the training and test sets is reasonable, which makes it affordable to run the experiments on the complete data set without the need to randomly select a subset of it. The data set consists of 41 feature attributes out of which 38 are numeric and 3 are symbolic. The total number of records in the data set is 125973 out of which 67343 are normal and 58630 are attacks. The dataset contains 24 different attack types which can be classified into four main categories, namely, Denial of Service (DOS), Remote to Local (R2L), User to Root (U2R), and Probing.

Table 1. Distribution of Records

Class	Number of Records	Percentage of Class Occurrences
Normal	67343	53.48%
DOS	45927	36.45%
U2R	52	0.04%
R2L	995	0.78%
Probes	11656	9.25%

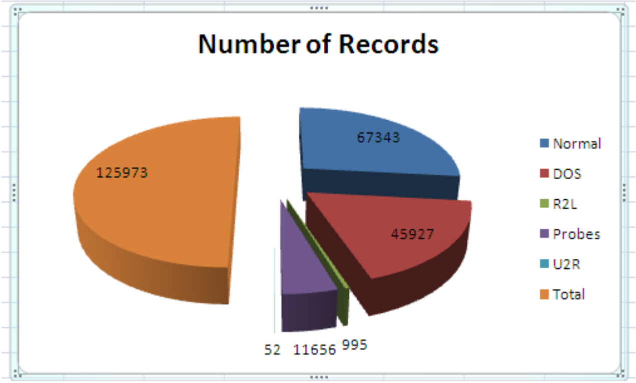

Figure 2. Distribution of Records

4.2. Feature Selection

Feature selection is the process of finding a subset of features from the original dataset. Redundant and irrelevant features of dataset may lead to low detection accuracy. Selection of significant features from intrusion dataset is a challenge for constructing high performance intrusion detection systems. The basic objective of feature selection is to remove noisy features and find important features which can represent data as a whole and maximize classification performance and at the same time reduce the computation time. In this study two categories of feature selection methods , namely, bio-inspired based feature selection methods such as Ant search, Genetic search, and Particle Swarm Optimization search algorithms and informed search methods such as best first search and greedy stepwise search methods have been applied to select relevant features. The number of features selected in each selection method is presented in Table 2

Table 2: Selected features using feature selection methods

Feature Selection Method	Name of the Method	Number of Features Selected	Feature Name
Bio-inspired Feature Selection Method	Ant Search	10	Flag, Src_bytes,Logg_in, R_shell, Se_se_rt, Sa_srv_rt, Di_srv_rt, Ds_Rate, Ds_d_h_rt, Ds_h_r
	Genetic Search	16	Service, Flag, Src_bytes. Dst_bytes, Land, Urgent,Logged_in, Srv_count, Serror_rate, Srv_serror_rate, Rerror_rate, Same_srv_rate, Diff_srv_rate, Dst_host_count, Dst_host_same_srv_rate, Dst_host_srv_serror_rate.
	PSO Search	12	Flag, Src_bytes, Dst_bytes, Urgent, Logged_in, Num_shells, Srv_serror_rate, Same_srv_rate, Diff_srv_rate, Dst_host_srv_diff_host_rate, Dst_host_serror_rate, Dst_host_srv_serror_rate.
Informed Search	Best First Search	10	Service,Flag, Src_bytes, Dst_bytes, Logged_in, Se_se_rt, Sa_srv_rt, Di_srv_rt, Ds_d_h_rt, D_h_sr
	Greedy Stepwise Search	11	Service,Flag, Src_bytes, Dst_bytes, Logged_in, Root_shell, Srv_serror_rate,Same_srv_rate, Diff_srv_rate, Dst_host_srv_diff_host_rate, Dst_host_serror_rate

4.3. Cross Validation

Cross validation calculates the accuracy of the model by separating the data into two different populations, a training set and a testing set. In 10-fold cross validation, a given dataset is partitioned into 10 subsets, of these 9 subsets constitute a training fold and a single subset is retained as the testing data. This cross-validation process is then repeated 10 times (the number of folds). The 10 sets of results are then aggregated via averaging to produce a single model estimation. The advantage of 10-fold cross validation over random sub-sampling is that the entire dataset is used for both training and testing, and each subset is used for testing only once per fold.

4.4. Confusion Matrix

An intrusion detection model can be evaluated by its ability to make accurate prediction of attacks. Intrusion detection systems mainly discriminate between two classes, attack class and normal class. The confusion matrix reports the number of False Positives (FP), False Negatives (FN), True Positives (TP), and True Negatives (TN).

TP (True Positive): An attack is detected successfully and an alarm is raised.

FP (False Positive): A normal connection is wrongly detected as an attack and a false alarm is raised.

TN (True Negative): A normal connection does not raise any alarm.

FN (False Negative): An attack is not detected and an alarm is not raised.

Based on these values the following performance measurements can be made:

Accuracy measure the probability that the algorithm can correctly predict positive and negative examples and is given by:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TN+TP+FN+FP}}$$

Precision is a measure of the accuracy provided that a specific class has been predicted and it is calculated as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$$

Recall / Detection rate measures the probability that the algorithm can correctly predict positive examples which is given by:

$$\text{Detection Rate or Recall} = \frac{\text{TP}}{\text{TP+FN}}$$

False Alarm Rate is defined as the number of normal instances incorrectly labelled as intrusion divided by the total number of normal instances and is given by:

$$\text{False Alarm Rate} = \frac{\text{FP}}{\text{TN+FP}}$$

F-Value is the harmonic mean of precision and recall which measures the quality of classification which is given by:

$$\text{F - Value} = 2 \times \frac{(\text{ Precision} * \text{Recall})}{(\text{Precision} + \text{Recall })}$$

5. RESULT ANALYSIS

Four rule learning based classifiers along with two categories of feature selection methods were applied on the dataset and their performance were measured in terms of accuracy, precision, recall, false alarm rate, and F-value. A comparative view of different combinations of classifiers and feature selection techniques are depicted in table 3 and 4. The result shows that JRip technique with Genetic Search feature selection gives the highest accuracy of 99.8246% and highest detection rate of 99.7834%. Decision table with Greedy stepwise feature selection method gives the lowest false alarm rate of 0.1084%. Recall and False Alarm Rate of different combinations of classifiers with bio-inspired search based feature selection methods and informed search based feature selection methods are presented in figures 3, 4, 5, and 6 respectively.

Table 3. Comparison of four Rule Learning classifiers using Bio-inspired search based feature selection methods

Bio-inspired Search Method	Classifier Techniques	Evaluation Criteria				
		Accuracy in %	Precision in %	Recall / Detection Rate in %	False Alarm Rate in %	F-Value in %
Ant Search	Decision Table	99.2625	99.5109	98.9016	0.4232	99.2053
	DTNB	99.2768	99.5399	98.9033	0.398	99.2206
	JRip	99.4459	99.6214	99.1864	0.3282	99.4034
	RIDOR	99.4538	99.5095	99.316	0.4262	99.4127
Genetic Search	Decision Table	99.0712	99.8715	98.1306	0.1099	98.9934
	DTNB	99.3864	99.1088	99.577	0.7796	99.3423
	JRip	**99.8246**	99.8396	**99.7834**	0.1396	**99.8115**
	RIDOR	99.7333	99.7406	99.6862	0.2257	99.7134
PSO Search	Decision Table	99.3594	99.8362	98.7856	0.1411	99.3081
	DTNB	99.1228	99.1709	98.9425	0.7202	99.0566
	JRip	99.488	99.7785	99.1199	0.1915	99.4481
	RIDOR	99.5944	99.7228	99.4047	0.2405	99.1656

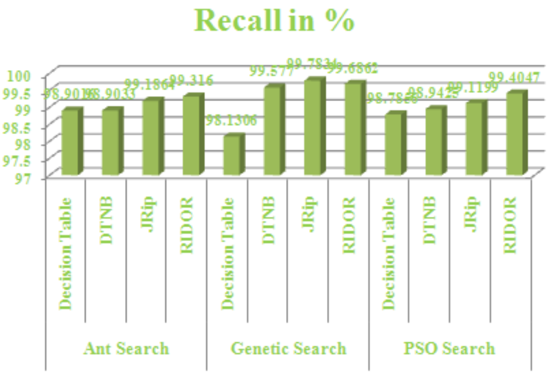

Figure 3. Comparison of Recall

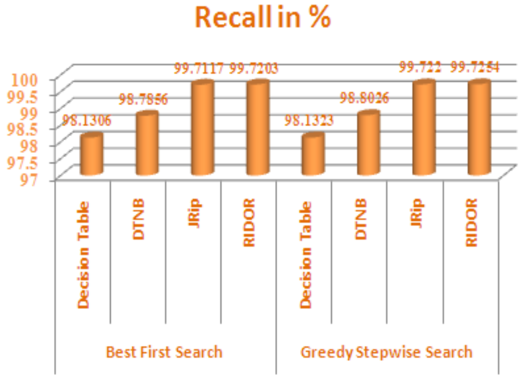

Figure 5. Comparison of Recall

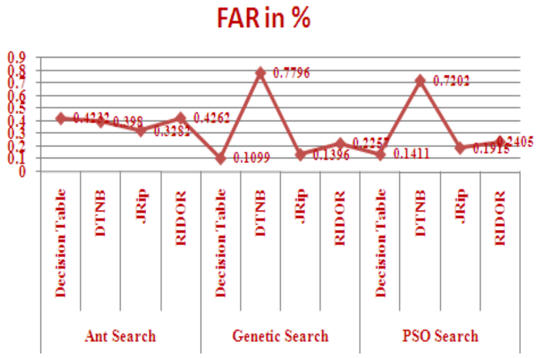

Figure 4. Comparison of False Alarm Rate

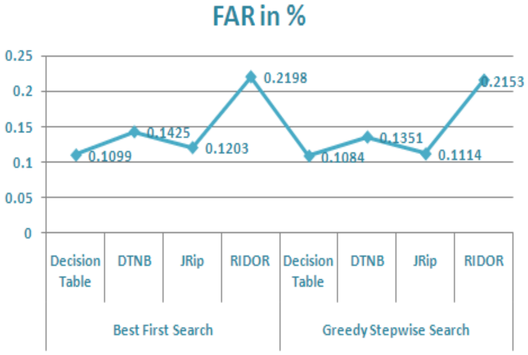

Figure 6. Comparison of False Alarm Rate

Table 4. Comparison of four Rule Learning classifiers using Heuristic search based feature selection methods

Heuristic / Informed Search Method	Classifier Techniques	Evaluation Criteria				
		Accuracy in %	Precision in %	Recall / Detection Rate in %	False Alarm Rate in %	F-Value in %
Best First Search	Decision Table	99.0712	99.8715	98.1306	0.1099	98.9934
	DTNB	99.3586	99.8345	98.7856	0.1425	99.3073
	JRip	99.8015	99.8616	99.7117	0.1203	99.7866
	RIDOR	99.7523	99.7475	99.7203	0.2198	99.7339
Greedy Stepwise Search	Decision Table	99.0728	99.8733	98.1323	**0.1084**	98.9952
	DTNB	99.3705	99.8431	98.8026	0.1351	99.3202
	JRip	99.8111	99.8719	99.722	0.1114	99.7969
	RIDOR	99.7571	99.7526	99.7254	0.2153	99.739

6. CONCLUSION

In this paper, we have proposed an intrusion detection model based on four rule learning based classifiers and two different categories of feature selection methods. The performance of the model was analyzed along different evaluation criteria on the intrusion dataset. It was observed that JRip technique with Genetic Search feature selection gives the highest accuracy of 99.8246% and highest detection rate of 99.7834%. Decision table with Greedy stepwise feature selection method gives the lowest false alarm rate of 0.1084%.

REFERENCES

[1] B. Kavitha, S. Karthikeyan, and P. Sheeba Maybell, Emerging Intuitionistic Fuzzy Classifiers for Intrusion Intrusion Detection System, , Journal of Advances in Information Technology, Vol. 2(2), 2011.

[2] Y.K.Jain and Upendra, Intrusion Detection using Supervised Learning with Feature Set Reduction, International Journal of Computer Applications, Vol. 33, 2011, pp. 22-31.

[3] Y. B.Bhaysar and K.C.Waghmare, Intrusion Detection System Using Data Mining Technique:Support Vector Machine, International Journal of Emerging Technology and Advanced Engineering, Vol.3, 2013, pp. 581-586.

[4] K.Atefi, S. Yahya, A.Y. Dak, and A. Atefi, A Hybrid Intrusion Detection System Based on Different Machine Learning Algorithms, Proceedings of the 4th International Conference on Computing and Informatics, ICOCI 2013, Malaysia, pp. 312-320.

[5] Sahilpreet Singh and Meenakshi Bansa , Improvement of Intrusion Detection System in Data Mining using Neural Network, International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, Issue 9, 2013 , pp. 1124-1130.

[6] R. Kohavi, The Power of Decision Tables, Proceedings of the European Conference on Machine Learning (ECML), Lecture Notes in Artificial Intelligence, Heraclion, Crete, Greece, Springer Verlag, 1995, pp. 174-189.

[7] M. Hall and E. Frank, Combining Naïve Bayes and Decision Table, Proceedings of the 21st Florida Artificial Intelligence Society Conference (FLAIRS), Florida, USA, May 15-17, 2008, pp. 318-319.

[8] William W. Cohen, Fast Effective Rule Induction, Proceedings of the Twelfth International Conference on Machine Learning, 1995, pp. 115-123.

[9] P. Sharma, “Ripple-Down rules for knowledge acquisition in intelligent system”, Journal of Technology and Engineering Sciences Vol. 1, pp. 52-56, 2009.

[10] M. Tavallaee, E. Bagheri, Wei Lu, and A. Ghorbani, A Detailed Analysis of the KDD CUP 99 data set”, Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Security and Defense Applications (CISDA 2009), pp. 1-6, 2009, pp. 1-6.