BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **18**, No 3
Sofia • 2018 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2018-0042

Data Pre-Processing and Classification for Traffic Anomaly Intrusion Detection Using NSLKDD Dataset

L. Gnanaprasanambikai, Nagarajan Munusamy

C. M. S. College of Science and Commerce, Bharathier University, Coimbatore, India
E-mails: gnanaambikai@gmail.com mnaagarajan@gmail.com

Abstract**:** *Network security is essential in the Internet world. Intrusion Detection is one of the network security components. Anomaly Intrusion Detection is a type of intrusion detection that captures the intrinsic characteristics of normal data and uses it in the detection process. To improve the performance of specific anomaly detector selecting the essential features of data and generating a good decision rule is important. The paper we present proposes suitable feature extraction, feature selection and a classification algorithm for traffic anomaly intrusion detection in using NSLKDD dataset. The generated rules of classification process are initial rules of a genetic algorithm.*

Keywords**:** *Traffic anomaly ids, genetic algorithm, feature extraction, feature selection, classification.*

1. Introduction

With growth of Internet in the modern world, security threats to the computer systems and network has increased a lot. The security threats affect the network security services. To control security threats number of technologies have been developed and deployed in organizations, for example, firewall, anti-virus software, message encryption, secured software protocols, and so on. In addition, to this Intrusion Detection is an important technology that existed for long time [1].

Intrusion Detection is a method that monitors the actions occurring in a computer system or in network and analyzes the actions for the indication of intrusions. Intrusion is any action that affects the integrity, confidentiality or availability of any network resources. To detect intrusions either in a computer system or in network, intrusion detection techniques are used [20]. Based on the monitoring activity, Intrusion Detection System (IDS) is classified into two types: Host based IDS (HIDS) and Network Based IDS (NIDS). HIDS run on individual hosts or devices in a network and monitors the incoming and outgoing packets for any undesirable actions [2]. NIDS are placed at any strategic point within the network and monitors the traffic for any undesirable actions [2]. Further, IDS are classified into two categories based on the detecting method, they are: Signature based IDS (or

Misuse IDS) and Anomaly based IDS. Signature Based IDS detect intrusions based on observed data that matches pre-defined description about the intrusion action. The advantage of Signature based IDS is that it provides accurate intrusion detection with low false positive rate and the disadvantage is the lack of detecting new intrusions. Anomaly Based IDS detect intrusions by detecting anomaly from observed data which deviates from normal behaviour [29]. The advantage of Anomaly IDS is that it detects new intrusions and disadvantage – generation of false positive [3].

In this paper, we present decision rules which are initial population for Genetic Algorithm process. Data Pre-processing is prior done and decision rules are created. This paper discusses the data pre-processing and decision rule generation to Genetic Algorithm process on Intrusion Detection System. Our objective is to present decision rules for Traffic anomaly Intrusion Detection, so we are experimenting with NSLKDD dataset, which is de-facto dataset anomaly intrusion detection.

Paper Organization. We have presented a brief introduction of IDS. The rest of the paper is organized as follows: Section 2 – Genetic Algorithm and its importance; Section 3 – Related work relevant papers to this research paper; Section 4 presents Data pre-processing steps and its implementation; Section 5 presents Classification and Decision rule generation; Section 6 presents Data set used and experimental results; Section 7 presents Conclusion of the paper.

2. Soft Computing – GA

Soft Computing is a new research field which includes technologies for intelligent problem solving. The objective of Soft Computing is to exploit fault tolerance, partial truth and uncertainty problems by achieving tractability, robustness and low cost solutions [23, 33]. Genetic Algorithm is a constituent of Soft Computing and it is used to mimic biological evolution for finding optimal solution to a problem. In Genetic Algorithm, the individual solutions to the problem are represented as initial population. Each individual is evaluated through a fitness function. Based on fitness value, the individuals are selected as parents and a new offspring is produced; the offspring is to form next new generation. The steps above are repeated until a solution is found which satisfies the pre-defined termination condition. Some areas where GA can be applied are medicine, machine learning, engineering applications, networking and routing, wired and wireless communications [4]. Genetic Algorithm can be applied to a number of tasks in IDS which include optimization, classification, and automatic model structure design.

3. Related work

In this paper, we generate decision rules from the dataset. The decision rules are the initial population for the Genetic Algorithm task. In a prior work to generate decision rules data pre-processing work was presented, namely feature extraction and feature selection. In a number of related works to Data pre-processing, Genetic algorithm are presented. In all these works main objective is to reduce false values in intrusion detection system. The purpose and motivation of the proposed paper is to generate decision rules for genetic algorithm process which find optimized rules with less false positive and less false negative values for intrusion detection system. The proposed work serves as input for fitness function of genetic algorithm process finding the best rules for intrusion detection system.

In this section, the related works are categorized in two sections: Section 1 suggest papers related to Data pre-processing and classification on an Intrusion Detection Dataset and Database. Section 2 suggests papers related to Data pre-processing technique used in Genetic Algorithm Process in Network Intrusion Detection.

Section 1 [5] shows PCA as feature extraction method and KNN classifier on detecting category based attacks with reduced calculation time and accuracy. [6] proposes a combined Network IDS framework of PCA and Naïve Bayes classifiers for Intrusion Detection Dataset. [7] shows Genetic algorithm for feature selection and decision tree for classification for Signature based NIDS with experimental analysis and results. [8] proposes decision tree and CBA based classification model for misuse and anomaly detection with experimental results and analysis on KDD99 dataset, which is not effective dataset for anomaly intrusion detection. [9] surveys decision tree algorithms for classification with various application, and suggests tools for implementing the algorithms. [10] gives descriptive review on network features and data pre-processing techniques used in anomaly intrusion detection and suggests a good PCA technique in data dimensionality compared to feature selection. [11] proposes a comparative performance on Decision tree and Rule based classifiers on multiple relational databases and their applicability on the database with empirical results and observations. [12] proposes PCA with SVM for selecting a feature subset.

Section 2 [13] uses information Gain as Feature selection technique and uses linear structure to classify normal and abnormal data for genetic algorithm process in intrusion detection. [14] suggests a correlation technique for important feature selection in network connection for GA-approach on classifying normal or attack. [15] implements PCA technique for dimensionality reduction with GA-approach on Network Intrusion Detection.

4. Data Pre-processing

Data Pre-processing is one of the critical steps in data mining process which performs the preparation and transformation of the original dataset. The various steps are included in Data pre-processing, they are Data cleaning, Feature reduction, Feature construction [10]. Feature Reduction includes Feature extraction and Feature selection. Feature extraction, selection and construction are all independent methods in data pre-processing. They can be combined depending on the problem being analyzed, like feature extraction followed by feature selection, feature construction followed by feature selection [21, 32, 31]. In this paper we follow the combination of feature extraction followed by feature selection.

4.1. Feature extraction and selection

Feature Extraction transforms data from high dimensionality to low dimensionality. Feature extraction is a process that determines what evidence that can be taken from audit data is most useful for analysis [24]. In this paper, Principle Component Analysis (PCA) method is used for feature extraction. PCA is a linear method in dimensionality reduction for data analysis and compression [30]. It is based on transforming a relatively large number of uncorrelated features by finding a orthogonal linear combinations of the original features with the largest variance [6].

Steps in PCA Algorithm:

Step 1. Get the input data.

Step 2. Find the mean.

Step 3. Subtract the mean.

Step 4. Calculate the Covariance matrix.

Step 5. Calculate the Eigen vectors and Eigen values of the covariance matrix.

Step 6. Sort the Eigen values in decreasing order and form feature vector.

Step 7. Derive the new dataset with reduced features.

The input to the PCA program is the NSLKDD dataset. From the covariance matrix we will find the Eigen value and Eigen vector by using the equation

$$|A - \lambda I|x = 0. \tag{1}$$

Sort the Eigen vector corresponding to the Eigen value. The Eigen vector with the highest Eigen value represents the first principle component of the data. The *K* Eigen vectors with the highest Eigen values are selected for feature reduction by the equation

$$\frac{\sum_{i=1}^{k} li}{\sum_{i=1}^{N} li} = \text{threshold}. \tag{2}$$

The screen-plot plots the Eigen values on the graph. The *K* Eigen values are selected by detecting elbow on the screen plot graph [29].

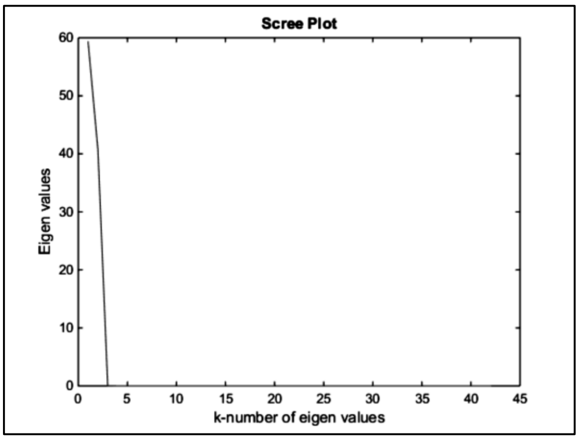

Fig. 1. Screen Plot of PCA

In the above screen plot the elbow decides the selection of eigen values. There is one elbow in the screeplot, where $k = 4$. Therefore, 41 features of NSLKDD dataset are reduced to four features [25].

4.2. Feature set for anomaly NIDS

In broad anomaly detection, separate feature sets are built for each of the anomaly detectors. For Traffic based anomaly NIDS a feature set of Multiple connection derivative features is included, which specifies the count of connections to a particular destination IP address and port [10]. In this paper a feature set of four features, namely, protocol, src_bytes, Dst_host_count, Dst_host_same_srv_rate, are constructed [10, 28]. The feature set thus constructed for anomaly detection is used to determine DDOS attack [16].

5. Classification

At this stage, we generate decision rules for the records of the dataset with the reduced feature set with two classes namely normal and anomaly [8]. The decision rules of the normal record set are considered for anomaly intrusion detection. There are two ways of generating decision rules for a dataset: Direct and Indirect. Direct way is to generate decision rules directly from data. Indirect way is to generate decision rules from other classification models. We have compared Ripper Algorithm, PART Algorithm and C4.5 Algorithm decision rule generating algorithms and analyzed C4.5 Algorithm is suitable for our problem solving.

5.1. Decision rule generators

5.1.1. RIPPER

RIPPER stands for Repeated Incremental Pruning to Produce Error Reduction. RIPPER is a direct way rule generator for two classes of problems: a rule is generated for one class and the other class is kept as Default class [17]. Rules are generated for the class having minimum records in a Dataset, and the other class is kept as Default class [17]. Ripper Algorithm is implemented as JRip in Weka3.6.

5.1.2. PART

PART stands for Projective Adaptive Resonance Theory. Part Algorithm is a Combination of C4.5 and Ripper Algorithm. Part Algorithm is indirect way rule generator, which employs a partial decision tree to generate decision rules [17] . Partial Decision tree is a regular tree that contain unexplored branches. In order to build this tree, subtree replacement is used as pruning startegy in building the tree. The Algorithm follows Divide-and-conquer strategy.

5.1.3. C4.5

C4.5 Decision Tree is a improvement to ID3 Decision Tree Algorithm. Decision tree is an indirect way of Decision rule generator. In C4.5 Decision Tree Algorithm, a decision tree is constructed using subtree replacement and subtree raising pruning strategies [22]. C4.5 uses Gain ratio criterion for choosing the best attribute for each decision node during construction of decision tree which gives larger information gain of an attribute[2], [22]. The process to generate Decision rules from Decision tree is straightforward. The algorithm generates decision rule for each leaf node of the decision tree. Conversion Algorithm of Decision tree to Decision rules generates the decision rules [22]. C4.5 is implemented as J48 in Weka 3.6.

Algorithm to generate decision rules from a decision tree

```
Input: Decision Tree T
Output: Decision Rules $R$
Algorithm:
$R$=0
For each path from root to a leaf in $T$ do
$a$=True
for each non-leaf node do
$a$=$a$^(label of node combined with label of incident outgoing arc)
$c$=label of leaf node
$R$=$R$ U $r$=⟨$a$, $c$⟩
```

5.2. Reasons for choosing C4.5 decision tree classifier

We used results from WEKA 3.6 for our comparison. The comparison is based on generation of maximum number of decision rules for normal class in minimum time. The table below gives the number of rules generated and time taken for it.

Table 1. Comparison of Algorithm Performance

Algorithm	No of rules for normal class	Time taken	Accuracy of correctly classified instances
RIPPER Algorithm	Normal Class becomes Default class	2.13 s	99.0632%
PART Algorithm	20 rules	0.52 s	99.0076%
C4.5 Algorithm	31 rules	0.36 s	99.0433%

From the table above, we could find that accuracy of correctly classified instances is same, for the above mentioned algorithms, the algorithms differ in number of rule generation. In case of RIPPER Algorithm, Decision rules are generated for anomaly class, making normal class as Default class which exploits the principle of anomaly intrusion detection [20, 8]. In case of PART and C4.5 greater number of normal class decision rules are generated in C4.5 as it follows Global optimization technique [27]. In our problem, we generate decision rules for initial population of Genetic Algorithm. Initial Population size of Genetic Algorithm should be large enough to explore the space of models more thoroughly and it helps in obtaining better optima [26]. The algorithm which follows Global optimization technique generates more decision rules. To our problem solving, we choose C4.5 as a suitable Algorithm for classification.

6. Experimental results

6.1. Dataset

NSLKDD dataset is de-facto a dataset for anomaly intrusion detection. [19] NSLKDD dataset is advanced version of KDDCUP99 dataset with no redundancy, no duplication of records and with less complexity level. NSLKDD dataset consists of 20% training dataset, full training dataset of 125,973 instances and testing dataset of 22,544 instances. In this paper, we create decision rules from 20%training dataset.

6.2. Generating decision rules

With PCA results and feature selection, Section 4, we select four features from 41 features in a dataset and form it - a reduced feature dataset. This reduced feature dataset serves as a dataset for generating decision rules. For our experiment we use WEKA3.6 for classification. C4.5 decision tree is implemented as J48 tree classifier in WEKA3.6. In this paper, we use only 20% of training dataset as input to J48 classifier and generate the decision tree below.

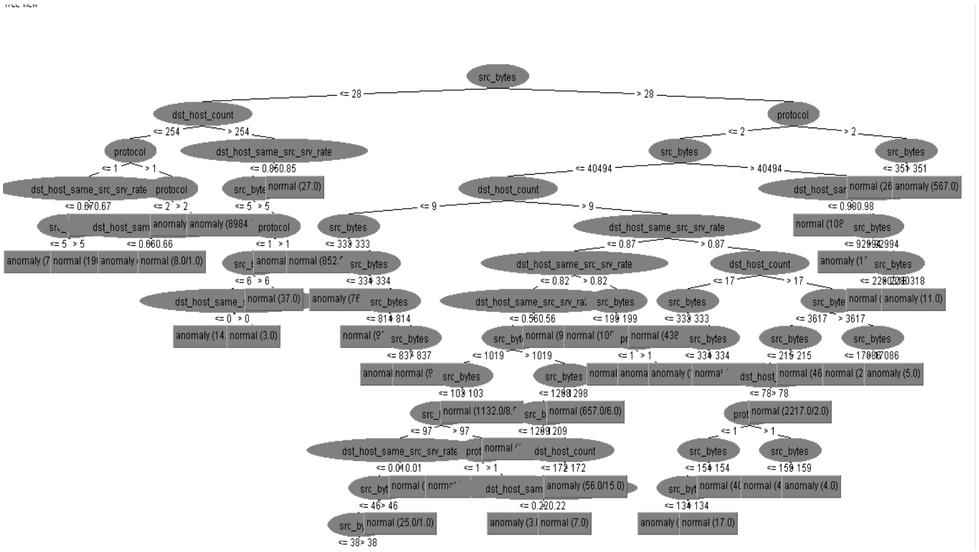

Fig. 2. Decision tree in WEKA3.6

By using the Conversion Algorithm from Section 5.1.3 on the decision tree above, we could generate 31 decision rules with normal class [8] for anomaly intrusion detection. Decision rules of simple if - then statement is used. The features are connected using a function in the if-then rule.

For example one decision rule could be

If protocol<=1 and src_bytes>5 and src_bytes<=28 and dst_host_count<=254 and dst_host_same_srv_rate <= 0.67 then normal

7. Conclusion

In this paper we generated decision rules for anomaly intrusion detection. To generate decision rules a prior work of Data pre-processing work is implemented with PCA as dataset dimensionality reduction. The features set used in decision rule serves for determining DDOS attack in turn identifying traffic volumes in the network. These Decision rules are generated for Traffic Anomaly Intrusion Detection and they serve as initial rules for Genetic Algorithm Process.

8. Future work

In this paper we generated decision rules from 20% of a training dataset, as a next step in future work we aim at finding the best fitness of the decision rules for Genetic Algorithm process using full training dataset and to generate fittest rules' set to apply on testing dataset.

References

1. Gong, R. H., M. Zulkernine, P. Abolmaesumi. A Software Implementation of a Genetic Algorithm Based Approach to Network Intrusion Detection. – In: Proc. of 6th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing and First ACIS International Workshop on Self-Assembling Wireless Networks, Towson, Maryland, USA, 23-25 May 2005, pp. 246-253.
2. Srinivasa, K. G., N. Pramod. gNIDS: Rule-Based Network Intrusion Detection Systems Using Genetic Algorithms. – International Journal of Intelligent Systems Technologies and Applications, Vol. **11**, 2012, Nos 3/4, pp. 252-266.
3. Wu, S. X., W. Banzhaf. The Use of Computational Intelligence in Intrusion Detection System – A Review. – Applied Soft Computing, Vol. **10**, 2010, Elseiver, pp. 1-35.
4. Sivanandam, S. N., S. N. Deepa. Introduction to Genetic Algorithms. Springer. ISBN 978-3-540-73189-4.
5. Zargar, G. R., T. Baghaie. Category Based Intrusion Detection Using PCA. – Journal of Information Security, Vol. **3**, 2012, pp. 259-271.
6. Neethu, B. Classification of Intrusion Detection Dataset Using Machine Learning Approaches. – International Journal of Electronics and Computer Science Engineering, Vol. **V1N3**, 2012, pp. 1044-1051.
7. Stein, G., B. Chen, A. S. Wu, K. A. Hua. Decision Tree Classifier for Network Intrusion Detection with GA-Based Feature Selection. – In: Proc. of 43rd Annual Southeast Regional Conference, ACM-SE 43, Vol. **2**, 2005, pp. 136-141.
8. Goel, R., A. Sardana, R. C. Joshi. Parallel Misuse and Anomaly Detection Model. – International Journal of Network Security, Vol. **14**, July 2012, No 4, pp. 211-222.
9. Patel, B. R., K. K. Rana. A Survey on Decision Tree Algorithm for Classification. – International Journal of Engineering Development and Research, Vol. **2**, 2014, Issue 1, pp. 1-5.
10. Davis, J. J., A. J. Clark. Data Preprocessing for Anomaly Based Network Intrusion Detection. – Computer & Security, 2011, Elseiver, pp. 353-375.
11. Thangaraj, M., C. R. Vijayalakshmi. Performance Study on Rule-Based Classification Techniques Across Multiple Database Relations. – International Journal of Applied Information Systems, Vol. **5**, March 2013, pp. 1-7. ISSN:2249-0868.
12. Eid, H. F., A. Darwish, A. E. Hassanien, A. Abraham. Principle Component Analysis and Support Vector Machine. – In: Proc. of 10th International Conference on Intelligent Systems Design and Applications, IEEE, 2010, pp. 363-367.

13. Abdullah, B., I. Abd-Alghafar, G. I. Salama, A. Abd-Alhafez. Performance Evaluation of a Genetic Algorithm Based Approach to Network Intrusion Detection. – In: Proc. of 13th International Conference on Aerospace Sciences & Aviation Technology, ASAT-13, 2009, pp. 1-17.
14. Kandeeban, S. S., R. S. Rajesh. A Mutual Construction for IDS Using GA. – International Journal of Advanced Science and Technology, Vol. **29**, April 2011, pp. 1-8.
15. Hashemi, V. M., Z. Muda, W. Yassin. Improving Intrusion Detection Using Genetic Algorithm. – Information Technology Journal, Vol. **12**, 2013, No 11, pp. 2167-2173.
16. Bhoria, P., D. K. Garg. Determining Feature Set of DOS Attacks. – International Journal of Advanced Research in Computer Science and Software Engineering, Vol. **3**, May 2013, Issue 5, pp. 875-878.
17. Vijayarani, S., M. Dhivya. An Efficient Algorithm for Generating Classification Rules. – International Journal of Computer Science and Technology, Vol. **2**, October-December 2011, Issue 4, pp. 512-515.
18. Kalyani, G., A. J. Lakshmi. Performance Assessment of Different Classification Techniques for Intrusion Detection. – IOSR Journal of Computer Engineering, Vol. **7**, November-December 2012, Issue 5, pp. 25-29.
19. Revathi, S., D. A. Malathi. A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection. – International Journal of Engineering Research & Technology (IJERT), Vol. **2**, December 2013, Issue 12, pp. 1848-1853.
20. Singh, B. Network Security and Management. PHI Learning Pvt Ltd. Second Edition. 2009.
21. Soman, K. P., S. Diwakar, V. Ajay. Insight into Data Mining Theory and Practice. PHI Learning Pvt Ltd. Third Edition. 2008.
22. Dunham, M. H. Data Mining Introductory and Advanced Topics. Pearson Education, Seventeeth, 2013.
23. Rajesekaran, S., G. A. Vijayalaksmi Pai. Neural Networks, Fuzzy Logic and Genetic Algorithms Synthesis and Applications. PHI, India, 2010.
24. Sumathi, S., S. N. Sivanandam. Data Mining in Security, Studies in Computational Intelligence (SCI). Springer, 2006, pp. 629 -648,
25. Janvier, 2013. **http://eric.univlyon2.fr/~ricco/tanagra/fichiers/en_Tanagra_Nb_Components_PCA.pdf**
26. Real, E., S. Moore, A. Selle, S. Sexana, Y. L. Suematsu, J. Tan, Q. V. Lie, A. Kurakin. Large-Scale Evolution of Image Classifier. – In: Proc. of International Conference on Machine Learning, 2017.
27. Eibe, F., I. H. Written. Generating Accurate Rulesets without Global Optimization. – In: Proc. of 15 International Conference on Machine Learning, 1998.
28. Rizwan, A., et al. Architecture of Hybrid Mobile Social Networks for Efficient Content Delivery. – Wireless Personal Communications, Vol. **80**, 2015, No 1, pp. 85-96.
29. Imran, M., et al. Pseudonym Changing Strategy with Multiple Mix Zones for Trajectory Privacy Protection in Road Networks. – International Journal of Communication Systems, Vol. **31**, 2018, No 1, pp. 34-37.
30. Zhao, X., et al. Dimension Reduction of Channel Correlation Matrix Using CUR-Decomposition Technique for 3-D Massive Antenna System. IEEE, Access 6, 2018, pp. 3031-3039.
31. Ezhilarasi, M., V. Krishnaveni. A Survey on Wireless Sensor Network: Energy and Lifetime Perspective. – Taga Journal of Graphic Technology, Vol. **14**, 2018.
32. Nagarajan, M., S. Karthikeyan. A New Approach to Increase the Life Time and Efficiency of Wireless Sensor Network. IEEE, 2012.
33. Ezhilarasi, M., V. Krishnaveni. An Optimal Solution to Minimize the Energy Consumption in Wireless Sensor Networks. – International Journal of Pure and Applied Mathematics, Vol. **119**, 2018, Issue 10.

Received 22.01.2018; Second Version 21.05.2018; Accepted 30.07.2018