K-Nearest Neighbor and Boundary Cutting Algorithm for Intrusion Detection System

Punam Mulak, D.P. Gaikwad and N.R. Talhar

Abstract Intrusion detection system is used for securing computer networks. Different data mining techniques are used for intrusion detection system with low accuracy and high false positive rate. Hicuts, HyperCuts, and EffiCuts are decision tree based packet classification algorithm which performs excellent search in classifier but requires high amount of memory. So in order to overcome these disadvantages, new approach is provided. In this, we present a hybrid approach for intrusion detection system. Boundary Cutting Algorithm and K-Nearest Neighbor using Manhattan and Jaccard coefficient similarity distance is used for high detection rate, low false alarm and less memory requirement. KDD Cup 99 dataset is used for evaluation of these algorithms. Result is evaluated using KDD CUP 99 dataset in term of accuracy, false alarm rate. Majority voting is done. This approach provides high accuracy and low memory requirements as compare to other algorithm.

Keywords Intrusion detection system · Supervised learning · Unsupervised leaning · Data mining

1 Introduction

Network security has become more and more important because of the rapid development of network services and sensitive information on the network. Although different security technologies are available for securing computer network, there are many undetected attacks. IDS are used to find any malicious attack on network. Therefore, intrusion detection system has become a research area.

P. Mulak (✉) · D.P. Gaikwad · N.R. Talhar
Computer Department, AISSMS College of Engineering, Kennedy Road, Pune, India
e-mail: punammulak@gmail.com

D.P. Gaikwad
e-mail: dp.g@rediffmail.com

N.R. Talhar
e-mail: nrtalhar@gmail.com

S.C. Satapathy et al. (eds.), *Information Systems Design and Intelligent Applications*, Advances in Intelligent Systems and Computing 434,
DOI 10.1007/978-81-322-2752-6_26

Intrusion detection system is mainly divided into three types [1]: Host based; Network based and Hybrid Intrusion detection system. Host based IDS is deployed at host level. All the parts of dynamic behavior and state of computer system are analyzed by a host based IDS. Network based IDS observe all the packets into the network and find the threats which affects network. Hybrid IDS is combination of both network and host based IDS. Supervised and unsupervised learning approach is used in intrusion detection system. Supervised learning predicts unknown tuples from training dataset. Hidden structure from unlabeled data can be found using unsupervised learning.

Different data mining and packet classification techniques are provided for intrusion detection system. Most of data mining techniques give high misclassification cost and false positive rate. So due to this reason quality of product may degrade. Hicuts, HyperCuts, and EffiCuts are decision tree based packet classification algorithm which performs excellent search in classifier but requires high amount of memory.

In this paper, a new method is introduced by combining data mining and packet classification approach to reduce memory requirement for rule matching and increase intrusion detection rate. Two proximity measures: Jaccard coefficient and Manhattan Distance and boundary cutting packet classification algorithm is used in proposed framework for high detection rate, low false alarm and less memory requirement. Jaccard coefficient requires less computation and provides high accuracy. Jaccard Coefficient and Manhattan distance helps to improve quality of classification algorithm. KDD Cup 99 dataset is used for evaluation of these algorithms. Classification is done based on voting criteria. Result is evaluated using KDD CUP 99 dataset in term of accuracy, false alarm rate. This combined approach provides high accuracy and low memory requirements than other algorithm.

Rest of the paper is organized as follows: Sect. 2 presents related work for intrusion detection system. Section 3 describes theoretical background of used algorithm. Section 4 provides proposed system. Section 5 gives experimental result of the system. And Sect. 5 present conclusion of paper.

2 Related Work

Different mechanism is used for intrusion detection system. In paper [2], author used random forest to reduce tedious rule formation for network traffic. Active routers are identified by using anomaly based detection. Packets are checked by using active router and random forest is used to detect attacks. Misuse and anomaly detection are performed by using random forest. In paper [3] author used decision tree data mining algorithm for intrusion detection and then compared performance with support vector machine. Both algorithms are evaluated using KDD cup 99 dataset. Decision tree algorithm gives better performance than SVM. Neural network is used for intrusion detection system and classification of attacks [4]. Author used multilayer Perceptron for intrusion detection system in offline approach.

In order to improve the detection and classification rate accuracy and achieve high efficiency, author proposed a multi-layer hybrid approach [5]. Three layers are implemented. In first layer, feature selection is done using principal component analysis algorithm. Anomaly detector is generated by using genetic algorithm which differentiates normal and abnormal behaviors. Classification is done using Different classifiers such as naive Bayes, multilayer perceptron neural. Bhattacharyya introduced clustering based classification method for Network intrusion detection system [6]. Similarity function is used to divide a set of labeled training data into clusters. Methods are provided for clustering, training and prediction. TUIDS Intrusion data set is used to evaluate proposed system.

Yanya and Yongzhong present multi-label k-Nearest Neighbor algorithm with semi supervised and multi-label learning approach [7]. K-nearest neighbor algorithm is used to indentify unlabeled data. Then Maximum a posterior principle is used to determine the label set for unlabeled data. Overall performance of intrusion detection system is increased. Hierarchical clustering is used for IDS and achieved 0.5 % false positive rate which is evaluated on KDDCup99 [8]. Te-Shun Chou introduced hybrid classifier systems [9] using fuzzy belief KNN for intrusion detection system. Overall detection rate is 93.65 %.

3 Theoretical Background

3.1 K-Nearest Neighbour Algorithm

K-nearest neighbor is instance based learning. It is the lazy learner. Lazy learners do more work at the time of classification than training [10]. Firstly all training tuples are stored. K-NN algorithm calculates the distance between test tuple $z = (x', y')$ and all the training tuples $(x, y) \in D$ which helps to decide its nearest neighbor list, D_z. Test tuple is classified based using majority voting class.

Algorithm

1. Determine k as number of nearest neighbors. D is the set of training tuples.
2. **for** each test example $z = (x', y')$ do
3. Compute $d(x', x)$, calculate the distance between z and every training tuple, $(x,y) \in D$
4. Select D_z, the set of k closest training example to z.
5. Classify test example based on majority voting
$$y' = argmax \sum_{(xi,yi) \in D} I(v = yi)$$
6. **end for**

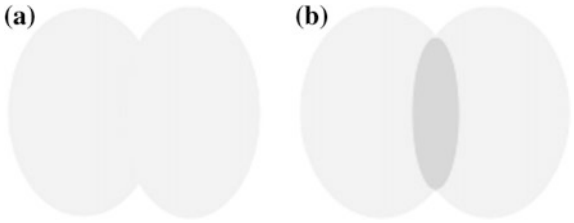

Fig. 1 The region of union and intersection between two sets P and Q. **a** PUQ. **b** PnQ

Different proximity measures are used to calculate the distance between training tuple and test tuple. It is describe as follows:

A. **Jaccard Coefficient Distance**

The similarity between sample sets is calculated by using Jaccard coefficient [11]. As shown in figure, the region of Intersection (P ∩ Q) and Union (P ∪ Q) between these two sets can be measured according to set theory (Fig. 1).
The Jaccard coefficient can be given as below:

$$JC(P,Q) = \left|\frac{P \cap Q}{P \cup Q}\right| \tag{1}$$

where P and Q are two different sets. Jaccard distance finds dissimilarity between two different sets. It is complementary of Jaccard coefficient (JC). The Jaccard distance can be given as below:

$$JC(P,Q) = 1 - JC(P,Q) = \frac{|P \cup Q| - |P \cap Q|}{|P \cap Q|} \tag{2}$$

B. **Manhattan Distance**

It is the distance between two points measured along axes. The distance between each data point and training tuples is calculated using the Manhattan distance metric as follows [12]:

$$Dist_{xy} = \left|X_{ik} - X_{jk}\right| \tag{3}$$

3.2 Boundary Cutting

Packet classification is one of essential function in network security. Hicuts, HyperCuts, EffiCuts are packet classification algorithm which perform excellent search but requires huge storage requirements. Boundary cutting (BC) algorithm is

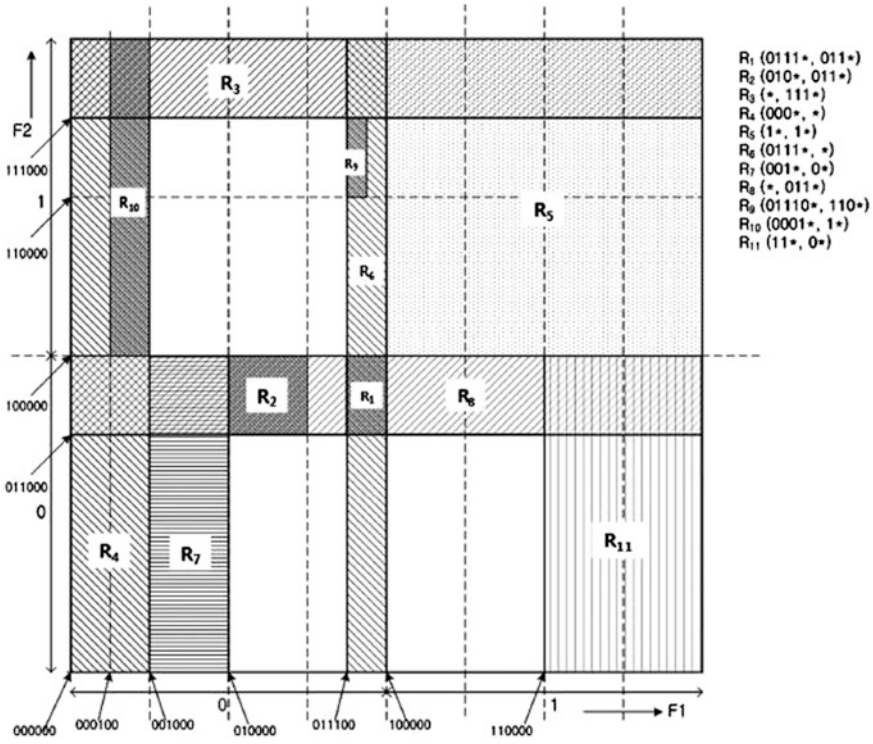

Fig. 2 Prefix plane for rules

used to overcome these disadvantages. BC finds the space that each rule covers. And then perform cutting according to the space boundary [13].

Consider two dimensional (2-D) planes which consist of first two prefix fields. An area is used to present the rule. Area that each rules cover in prefix plane are given in following Fig. 2.

In case of Hicuts algorithm, it cuts the spaces into subspaces using one dimension per step. Space and binth are used. A space factor (spfac) is space measure function used to give the number of cuts for selected field. The binth is a function which predetermined number of rules. Process starts from root node, and then bits of a header field are checked that is set at each interval node until a leaf node is reached to condition. But in this algorithm, cutting is based on a fixed interval and partitioning is ineffective. In order to balanced required memory size and search. In boundary cutting algorithm, each rule of starting and ending boundaries can be used for cutting. Decision tree for boundary cutting algorithm is given in Fig. 3. Binth is set to 3. Fixed intervals are not used for cuts at each internal node of Boundary cutting decision tree. This decision tree algorithm finds for subspace in which packet belongs and compares the header of input packets for

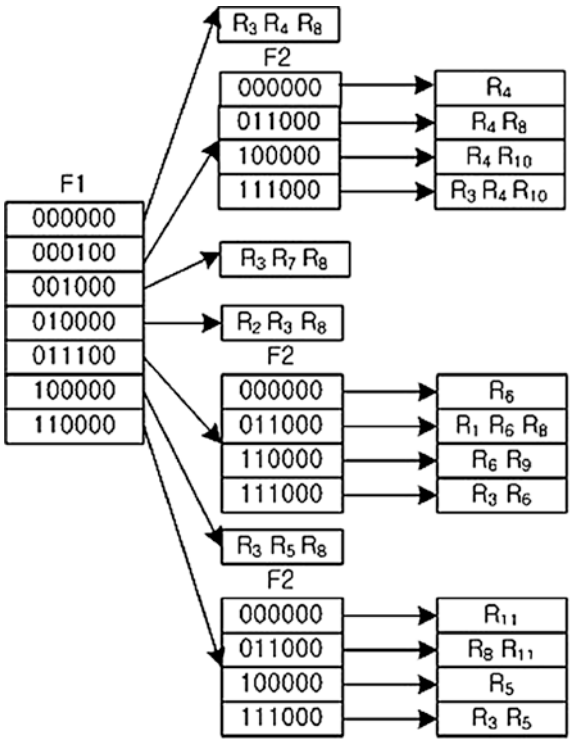

Fig. 3 Decision tree using boundary cutting algorithm

entire fields to rules belonging the subspace. Unnecessary cuttings are avoided. Rules replication is also reduced.

3.3 Normalization Techniques

The dataset available may contain some row data. Attributes of dataset have different data types. Range of these attributes may vary widely. The attributes are scaled to fit into a specific range. All the values are divided by the maximum value in the dataset. This leads to normalization. Different methods are provided for normalization [14].

1. Min-Max Normalization: This method performs linear transformation on original data. It maps a value of v_i in the range of new_min_A, new_max_A.

$$v_i' = \frac{v_i - \min_A}{\max_A - \min_A}\left(new\_\max_A - new\_\min_A\right) + new_{\min_A} \tag{4}$$

2. Z-score normalization: The value for attribute A is normalized by using mean and standard deviation of A. Equation for Z-score normalization is given below:

$$v_i' = \frac{v_i - A}{\sigma_A} \tag{5}$$

3. Decimal Scaling: Normalization is performed by moving the decimal point of value of attribute A. Result of this method is in between −1 and 1.

$$v_i' = \frac{v_j}{10^j} \tag{6}$$

4 Proposed System

The proposed system is a intrusion detection system which is developed using two proximity measures and packet classification algorithm i.e. Jaccard Coefficient based classification, Manhattan distance and Boundary cutting algorithm respectively as shown in Fig. 4. Majority voting has been done to increase the overall accuracy.

System is presented with authentication. User need to first login to the system.

Here, KDDCup99 dataset is used in system. This dataset is given as input to training module. KDD dataset is divided into different files depending on attack type like training3attack, training4attack, training5attack, training6attack and so on. Attributes of KDD dataset have different data types. Range of these attributes may vary widely. So there is need to perform normalization. Different methods are provided for normalization. Among them min-max normalization method is more effective. So here min-max normalization is performed to fit the attributes of dataset into specific range. KDD dataset contain some string values like services, flags, etc. which is not understand by classifiers. So these string values are converted into numeric value. After normalization, system is trained using three classifier i.e. K-NN using two proximity measures and boundary cutting algorithm. All the training tuples are stored in nearest neighbor classifier and value of k is defined. When new test tuple is given then distance between all training tuples and test tuple is calculated using Jaccard and Manhattan distance separately. All the tuples are sorted in ascending order. And majority voting is done to assign class label to test tuple. This majority

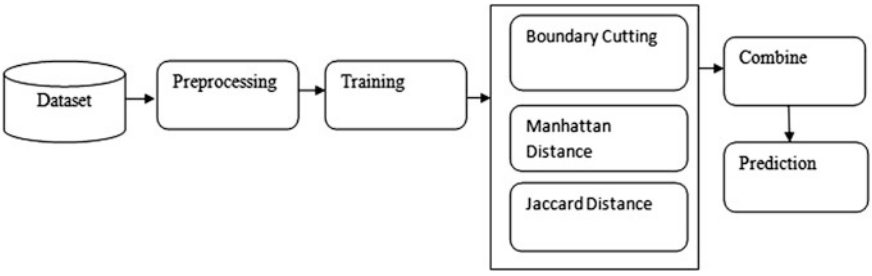

Fig. 4 The architecture of the proposed intrusion detection system

voting is depends on closeness of test tuple and training tuples. In case of boundary cutting algorithm, decision tree is build using boundary cutting algorithm. K-dimension header fields are considered for BC. Here, k = 10 means first 10 attributes of dataset are used for cutting. Algorithm cuts the space into smaller sub-regions, one dimension per step. Bucket size is taken as 16. Space factor is 2.0. Decision tree is constructed according cuts to header fields. Cutting is stopped when no space found in header fields. Depth of this tree is 5. Classification of dataset is done using decision tree. Cross Validation is used for testing. Main goal of cross validation is to check how accurately a predictive model will predict the unknown data. In cross validation, dataset is partition into subsets. One subset is given to training and other for testing. This partition of dataset is called folds. Maximum 10 folds cross validation can be done. Here, 3-fold cross validation is used because it provides more accuracy. KDD dataset is divided into three subsets. This division of dataset is done randomly. First two parts is provided for training and other part is applied for testing against trained algorithms. Combined approach is also provided. In combine approach majority voting is done. It is the process by which multiple classifiers are combined to solve a particular problem intelligently to improve the performance of a model. In system, predictions of these three algorithms are compared and majority of predictions are return. This method improves the accuracy of system.

5 Experimental Result

The proposed intrusion detection system using majority voting of BC algorithm and K-NN is evaluated and tested using KDDCup99 dataset. The experiments are performed on Laptop with 2 GB RAM and Pentium Dual Core processor. The performance of the proposed system is evaluated in term of classification accuracy. It is found that the classification accuracy of the proposed combined approach is increased up to 99 %. Following graph shows attack wise classification accuracy of the system (Fig. 5).

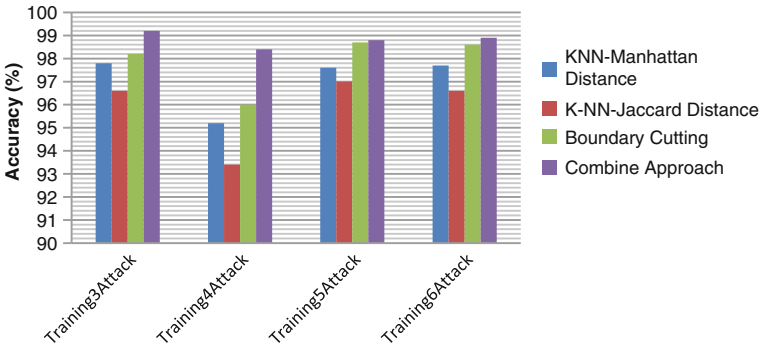

Fig. 5 Accuracy graph

6 Conclusion

In this paper, the architecture of general intrusion detection system have been studied and presented. The Hicuts algorithm is mostly used for packet classification in router. This algorithm can give better search performance, but it takes more memory for large rules. To overcome this problem, the BC algorithm is used with K-NN approach. The intrusion detection system has implemented using the combination method of BC algorithm and K-NN. The experiments have performed on separate algorithms and combined approach. It is found that the combined method provides better classification accuracy as compared to individual approaches. KNN using Jaccard and Manhattan distance gives more accuracy, sensitivity and specificity than other distance measures. The combined approach exhibits low false positive rate as compared to individual approaches. It is found that overall performance of the proposed classifier is increased due to the combination of these two approaches. The main disadvantage of the system is that it takes more time to generate rule from training dataset.

References

1. Punam Mulak, Nitin Talhar.: Novel Intrusion Detection System Using Hybrid Aprroach, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 11, November (2014).
2. Prashanth, V. Prashanth, P. Jayashree, N. Srinivasan.: Using Random Forests for Network-based Anomaly detection at Active routers, IEEE International Conference on Signal processing, Communication and networking, Madras Institute of Technology, Anna University Chennai India, and Jan 4–6, (2008).
3. Sandhya Peddabachigari, Ajith Abraham, Johnson Thomas.: Intrusion Detection Systems Using Decision Trees and Support Vector Machines, IEEE.
4. Mehdi MORADI and Mohammad ZULKERNINE.: A Neural Network Based System for Intrusion Detection and Classification of Attacks.
5. Amira Sayed A. Aziz Aboul Ella Hassanien Sanaa El-Ola Hanafy M.F. Tolba.: Multi-layer hybrid machine learning techniques for anomalies detection and classification approach, IEEE, (2013).
6. Prasanta Gogoi, B. Borah and D. K. Bhattacharyya.: Network Anomaly Identification using Supervised Classifier, Informatica 37 93–7 (2013).
7. Yanyan Qian, Yongzhong Li.: An Intrusion Detection Algorithm Based on Multi-label Learning, Workshop on Electronics, Computer and Applications, IEEE, (2014).
8. R. W.-w Hu Liang and R. Fei.: An adaptive anomaly detection based on hierarchical clustering, Information Science and Engineering (ICISE), 2009 1st International Conference on, Changchun, China, Dec. 2009, pp. 1626–1629.
9. Zhao Ruan, Xianfeng Li, Wenjun Li.: An Energy-efficient TCAM-based Packet Classification with Decision-tree Mapping, IEEE, (2013).
10. Pang-Nang Tan, Michael Steinbach, Vipin Kumar.: Data Mining.
11. Rajendra Prasad Palnatya, Rajendra Prasad Palnaty.: JCADS: Semi-Supervised Clustering Algorithm for Network Anomaly Intrusion Detection Systems, IEEE, (2013).
12. Archana Singh, Avantika Yadav, Ajay Rana.: K-means with Three different Distance Metrics International Journal of Computer Applications, Volume 67–No.10, April (2013).

13. Hyesook Lim, Nara Lee, Geumdan Jin, Jungwon Lee, Youngju Choi, Changhoon Yim.: Boundary Cutting for Packet Classification, IEEE/ACM TRANSACTIONS ON NETWORKING, VOL. 22, NO. 2, APRIL, (2014).
14. Jiawei Han, Micheline Kamber, Jian Pei.: Data Mining concepts Technologies, Third Edition Elsevier.