ANALYZING AND CLASSIFYING WEB APPLICATION ATTACKS

EIEI HAN

University of Computer Studies (Yangon)
E-mail: eieihan.ucsy@gmail.com

Abstract- This paper compares two approaches of random forests and K-Means+ID3 algorithms for classifying web application attacks. In the first method, Random Forest algorithm is used for the classification of web attacks. In the second method, the k-Means clustering method first partitions the training instances into k clusters using Euclidean distance similarity. On each cluster, an ID3 decision tree is built. To obtain a final decision on classification, the decision of the k-Means and ID3 methods are combined using two rules: (1) the Nearest-neighbor rule and (2) the Nearest-consensus rule. This paper also describes comparison results of these two approaches.

Keywords- K-Means+ID3, Random Forests Algorithm, Classification.

I. INTRODUCTION

Web applications are becoming increasingly popular and complex in all sorts of environments, ranging from e-commerce applications to banking. The security of web applications has become increasingly important and a secure web environment has become a high priority for e-business communities. They are subject to all sorts of attacks. In today's times, the most critical issue for any web application is security. Web servers and web-based applications are popular attack targets. To detect web-based attacks, intrusion detection systems are configured with a number of signatures that support the detection of known attacks. This paper describes how to differentiate normal access patterns from malicious access patterns. This system will be effective for the security of web application systems. Web application security is one of the research areas in these years.

Malicious users try to attack a web site or web server by using various attack patterns. Web application attacks are occurred by performing Web application queries. They take the form of well-defined strings and parameters. These are recorded in the web server log file. By analyzing each record of server log file, malicious patterns can be detected.

This paper is organized as follows. Section 2 describes about related works of the system. In section 3, background theories and random forests algorithm are discussed. Section 4 is about anomaly detection with K-Means clustering and ID3 decision tree algorithms. Section 5 illustrates experimental results of the system and Section 6 concludes the paper.

II. RELATED WORKS

Jiong Zhang and Mohammad Zulkernine employ random forests for intrusion detection system. Random forests algorithm is more accurate and efficient on large dataset like web attack dataset. Intrusion Detection System for protecting a Web application uses ID3, a well-known classifier that builds a decision tree from a fixed set of examples. Each input example is a Web application query; it has several attributes and belongs to a class, either attack or normal.

K-Means clustering and the Id3 decision tree learning methods to classifying anomalous and normal activities can be used in computer networks. The K-Means clustering method first partitions the training instances into two clusters using Euclidean distance similarity. On each cluster, representing a density region of normal or anomaly instances, we build an ID3 decision tree. The decision tree on each cluster refines the decision boundaries by learning the subgroups within the cluster.

III. BACKGROUND THEORY

The broad categories of web attack detection techniques exist. Supervised attack detection techniques learn a classifier using labeled instances belonging to normal and abnormal class and then assign a normal or anomalous label to a test instance. Data Mining interfaces support the following supervised functions:

A classification task begins with build data for which the target values (or class assignments) are known. Different classification algorithms use different techniques for finding relations between the predictor attributes' values and the target attribute's values in the build data.

Decision tree rules provide model transparency so that a business user, marketing analyst, or business analyst can understand the basis of the model's predictions, and therefore, be comfortable acting on them and explaining them to others.

Unsupervised anomaly detection techniques detect anomalies in an unlabeled test data set under the assumption that majority of the instances in the data set are normal.

Clustering is a data mining technique used to place data elements into related groups without advance knowledge of the group definitions.

As data mining techniques make it possible to search large amount of data for characteristics, rules and patterns, it can be applied to web application attacks detection.

Decision Tree algorithm ID3 can be used for Web attack detection. It is sufficient to put apart a number of Web attacks, including a large proportion of their variants. The ID3 output classification rules are easy to read. Decision tree can be easily explored to find out the rules applied by the classifier. ID3 is able to successfully classify unseen Web application queries as an attack. It suffices to generalize on specifying a wide range of Web attacks. ID3 is a simple inductive, non-incremental, classification algorithm.

A. Random Forests Algorithm

The random forests algorithm is a classification algorithm consisting of a collection of tree structured classifiers, where each tree casts a unit vote for the most popular class at each input. Random Forest can run efficiently on large databases. It is an effective method for estimating missing data and maintains accuracy when a large proportion of data are missing. Random forests are examples of ensemble methods, which combine predictions of weak classifiers. Steps of random forests algorithm are as follows:

1. For b=1 to B:

(a) Draw a bootstrap sample Z^* of size N from the training data.

(b) Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.

i. Select m variables at random from the p variables.
ii. Pick the best variable/split-point among the m.
iii. Split the node into two child nodes.

2. Output the ensemble of trees $\{T_b\}^B_1$.

To make a prediction at a new point x:

Regression: $f^{\wedge B}_{rf}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$.

Classification: Let $C^{\wedge}_b(x)$ be the class prediction of the b^{th} random-forest tree. Then $C^{\wedge B}_{rf}(x)$=majority vote $\{C^{\wedge}_b(x)\}^B_1$.

IV. ANOMALY DETECTION WITH K-MEANS CLUSTERING AND ID3 DECISION TREE

In the first stage, K-Means clustering is performed on training instances to obtain k disjoint clusters. Each K-Means cluster represents a region of similar instances, "similar" in terms of Euclidean distances between the instances and their cluster centroids. In the second stage, an ID3 decision tree is built using the instances in each K-Means cluster. This combined approach alleviates two problems (1) the Forced Assignment problem and (2) the Class Dominance problem.

A. Anomaly Detection with K-Means Clustering

K-means clustering is a clustering analysis algorithm that groups objects based on their feature values into K disjoint clusters. Steps in K-means clustering are as follows:

1. Define the number of clusters K.
2. Initialize the K cluster centroids.
3. Iterate all objects & compute the distances to the centroids of all the clusters.
4. Recalculate the centroids of all the modified clusters.
5. Repeat step 3 while the centroids do not change any more.

A distance function is Euclidean function,

$$d(x,y) = \sqrt{\sum_{i=1}^{m}(xi - yi)2} \quad (1)$$

where, $x=(x_1,\ldots\ldots,x_m)$ and $y=(y_1,\ldots\ldots,y_m)$ are the two input vectors with m quantitative features. This paper applies the K-Means clustering algorithm to training dataset that contain normal as well as anomalous traffic.

B. Anomaly Detection with ID3 Decision Trees

Steps of ID3 algorithm are as follows:

Create a root node for the tree
IF all examples are positive, Return the single-node tree Root, with label = +
If all examples are negative, Return the single-node tree Root, with label = -
If number of predicting attributes is empty, then Return the single node tree Root, with label = most common value of the target attribute in the examples
Otherwise Begin
A ← The Attribute that best classifies examples
Decision Tree attribute for Root ← A
For each positive value, v_i, of A,
Add a new tree branch below Root, corresponding to the test A = v_i
Let Examples(v_i), be the subset of examples that have the value v_i for A
If Examples(v_i) is empty
• Then below this new branch add a leaf node with label = most common target value in the examples
• Else below this new branch add the subtree ID3 (Examples(v_i), Target_Attribute, Attributes – {A})
End
Return Root

The ID3 decision tree learning algorithm computes the Information Gain G on each attribute A, as:

$$G(S,A)=Entropy(S)-\sum_{v\epsilon Values(A)}\frac{|S_v|}{|S|}Entropy(S_v) \quad (2)$$

where S is the total input space and S_v is the subset of S for which attribute A has a value v. The Entropy(S) over c classes is given by:

$$\text{Entorpy (S)} = \sum_{i=1}^{c} -p_i \log_2(p_i) \quad (3)$$

The attribute with the highest information gain is chosen as the root node of the tree. Aleaf-node is formed when all the instances within the available training subspace are from the same class. The algorithm constructs the ID3 decision tree with the normal purified traffic. Anomaly detection is performed using this tree by traversing the tree with features of test instance. If the traverse reaches a leaf node, the test instance will be detected as normal; else it will be detected as abnormal.

C. Combination of K-Means+ID3

We are provided with a normal (purified) training data set $X_i$where each instance represents an n dimensional vector. The proposed method has two phases: training and testing. During training phase, K-Means-based anomaly detection method is first applied to partition the training space into k disjoint clusters $C_1,C_2,\ldots,C_k$. Then ID3 decision tree is trained with instances in each K-Means cluster. The K-Means clustering method ensures that each training instance is associated with only one cluster. However, if there are any subgroups or overlaps within a cluster, the ID3 decision tree trained on that cluster refines the decision boundaries by partitioning the instances with a set of if-then rules over the feature space.

The combined application of the two algorithms overcomes some limitations of each algorithm when applied individually. For example, selection of a right value for parameter k in the K-Means clustering algorithm can effect on the overall accuracy of the algorithm. Considerably little values of k, compared to inherent number of natural sub groupings within the training data will lead to overlapping subgroups within clusters. This problem is compensated by ID3 decision tree constructed in each cluster.

The results of K-Means clustering and ID3 decision tree are combined by nearest-consensus rule and nearest-neighbor rules.

1) Nearest-Consensus Rule

In the Nearest-consensus rule, the decision is selected in which there is consensus between the decisions of the K-Means and the ID3 decision tree methods. In Figure 1, '1' means 'anomaly' and '0' means normal.

K-Means	1	1	0	…	1
ID3	0	1	0	…	0

Consensus

Figure 1: Nearest Consensus Rule

2) NEAREST-NEIGHBOR RULE

The Nearest-neighbor rule chooses the decision of the ID3 decision tree that is associated with the nearest candidate cluster.

The K-Means+ID3 methods also have high detection accuracy than the individual K-Means and the ID3. Figure2 describes about system architecture.

V. PROPOSED SYSTEM ARCHITECTURE

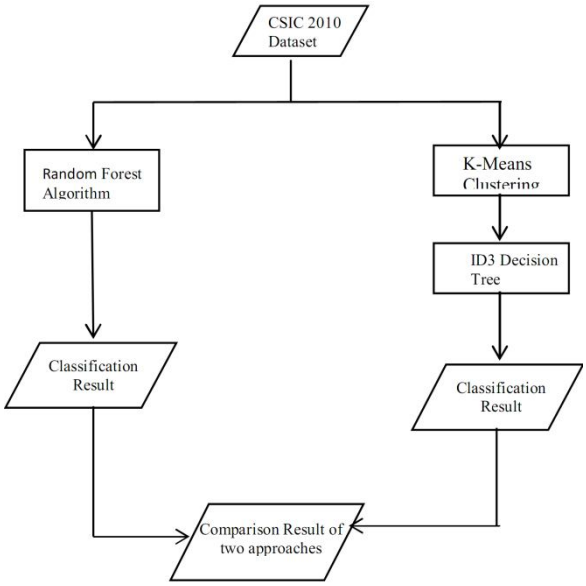


Figure 2: System Architecture

VI. EXPERIMENTAL RESULT

This system uses CSIC 2010 dataset in .csv format that contains normal and attack web logs. This dataset contains the field names of index, method, url, protocol, userAgent, pragma, cacheControl, accept, acceptEncoding, acceptCharset, acceptLanguage, host, connection, contentLength, contentType, cookie, payload and label. There are 223585 records in the dataset.

Figure3 illustrates the performance of random forest algorithm, K-means+ID3 algorithm and individual K-means and ID3 algorithms on the CSIC 2010 dataset. Results are tested by WEKA tool. By analyzing the results, Random Forest and K-means+ID3 methods have high accuracy and low error rate.

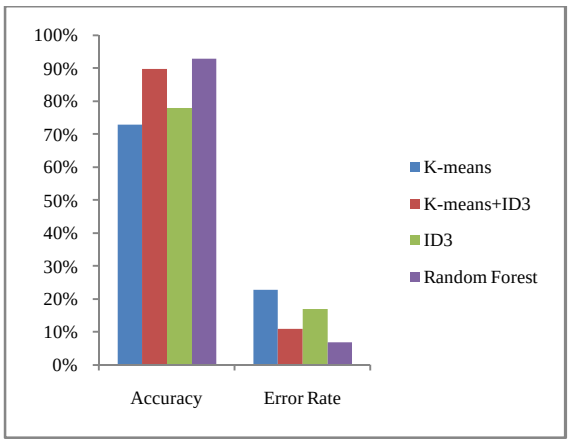


Figure 3: Performance of algorithms

CONCLUSION

This paper describes about anomaly detection of web based attacks. Normal and abnormal attack records are classified using random forest algorithm and K-means+ID3 algorithms. Comparison results of these two approaches are illustrated in this paper. By analyzing web attack log dataset with different data mining algorithms, one can know effectiveness of these methods.

REFERENCES

[1] AnandBahety, "Extension and Evaluation of ID3 – Decision Tree Algorithm", Department of Computer Science, University of Maryland, College Park.

[2] Breiman L, Cutler A. Random forests; 2006. <http://www.stat.berkeley.edu/~breiman/RandomForests/>.

[3] Christopher Kruegel, Giovanni Vigna, William Robertson, "A multi-model approach to the detection of web-based attacks", Reliable Software Group, University of California, Santa Barbara, USA, 2005.

[4] GhatgeDipali D, "Network Traffic Intrusion Detection System Using Decision Tree & K-Means Clustering Algorithm", Computer Engineering Department, KarmaveerBhauraoPatil College of Engg. & Poly.,Satara.

[5] J. Zhang and M. Zulkernine, "Anomaly Based Network Intrusion Detection with Unsupervised Outlier Detection" , Symposium on Network Security and Information Assurance Proc. of the IEEE International Conference on Communications (ICC), 6 pages, Istanbul, Turkey, June 2006.

[6] K. HanumanthaRao, G. Srinivas, AnkamDamodhar and M. Vikas Krishna, "Implementation of Anomaly Detection Technique Using Machine Learning Algorithms", Sri Indu College of Engineering and Technology, Hyderabad, India.

[7] Reda M. Elbasiony, Elsayed A. Sallam, Tarek E. Eltobely, Mahmoud M. Fahmy, "A hybrid network intrusion detection framework based on random forests and weighted K-Means", Tanta University, Faculty of Engineering, Tanta, Gharbia, Egypt, 2013.

[8] Shekhar R. Gaddam, Vir V. Phoha, Senior Member, IEEE, and Kiran S. Balagani , "K-Means+ID3: A Novel Method for Supervised Anomaly Detection by Cascading K-Means Clustering and ID3 Decision Tree Learning Methods", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 2007.

[9] Talasila Vamsidhar, Reddyboina Ashok and RayalaVenkat ,"Intrusion Detection System For Web Applications With Attack Classification", Journal of Global Research in Computer Science, 2012 .

[10] Vıctor H. Garcıa, Raul Monroy, and Maricela Quintana, "Web Attack Detection Using ID3", Computer Science Department, Mexico.

★★★