ANOMALY-BASED INTRUSION DETECTION THROUGH K-MEANS CLUSTERING AND NAIVES BAYES CLASSIFICATION

Warusia Yassin, Nur Izura Udzir[1], Zaiton Muda, and Md. Nasir Sulaiman

[1]Faculty of Computer Science and Information Technology, Universiti Putra Malaysia,
43400 UPM Serdang, Selangor, Malaysia, izura@fsktm.upm.edu.my

ABSTRACT. Intrusion detection systems (IDSs) effectively balance extra security appliance by identifying intrusive activities on a computer system, and their enhancement is emerging at an unexpected rate. Anomaly-based intrusion detection methods, which employ machine learning algorithms, are able to identify unforeseen attacks. Regrettably, the foremost challenge of this method is to minimize false alarm while maximizing detection and accuracy rate. We propose an integrated machine learning algorithm across K-Means clustering and Naïve Bayes Classifier called KMC+NBC to overcome the aforesaid drawbacks. K-Means clustering is applied to labeling and gathers the entire data into corresponding cluster sets based on the data behavior, *i.e.* , *i.e.* normal and attack, while Naïve Bayes Classifier (NBC) is applied to reorder the misclassified clustered data into correct categories. Experiments have been performed to evaluate the performance of KMC+NBC and NBC against ISCX 2012 Intrusion Detection Evaluation Dataset. The result shows that KMC+NBC significantly improves the accuracy, detection rate up to 99% and 98.8%, respectively, while decreasing the false alarm to 2.2%.

Keywords: Intrusion Detection System, Anomaly-based Intrusion Detection, Machine Learning, Clustering, Classifier

INTRODUCTION

The emergence of the Internet as a communication tool and the growth of network technologies in accessing critical information in real-time manner raise cyber offense cases as the attackers' interest on this valuable information radically increased the amount of cyber-attacks. Intrusion detection system (IDS) is used to identify such activities that breach the security and privacy either in network or in computer system surroundings. IDS can be classified into two categories, namely signature-based detection (SBD) and anomaly-based detection (ABD) (Lee, Stolfo & Mok, 1998).SBD inspects patterns of incoming packets against the record of predefined signatures to detect known attacks. Unfortunately, the signature database requires frequent updating, otherwise, uncovered known attacks will be missed due to the absence of the signature. ABD on the other hand was designed to distinguish normal and abnormal activities, thus helps in detecting unknown attacks more precisely. However, this detection method is likely to generate a high rate of false alarms: high false positives may lead to an actual intrusion being detected but ignored by the users, while false negatives may give a misleading sense of security (Yingbing, 2012, Carlos, 2012).

Machine learning algorithms have been widely used by many researchers as ABD to discover attacks (Patcha & Park, 2007). However, the high false alarm rate still remains as a main limitation in structuring an efficient ABD in these works. (Carlos, 2012, Panda, Abraham & Patra, 2012). Therefore, as false alarms may render the IDS unreliable, there is a necessary requirement to detect attacks more accurately to reduce false alarms.

In this work, we propose an integrated machine learning algorithm based on K-Means clustering and Naïve Bayes Classifier called KMC+NBC to perform the challenging task of ABD in improving false alarm rate as well as maximizing detection and accuracy rate. The performance of KMC+NBC compared with Naïve Bayes individually using Intrusion Detection Evaluation dataset ISCX 2012.

The rest of the paper is organized as follows: Similar works are discussed in Section 2, while the proposed approach is described in Section 3. In Section 4, experimental results and performance comparison is discussed. Finally, we present the conclusion and future work in Section 5.

RELATED WORK

Machine learning algorithms were introduced in the field of anomaly detection to discover predictability of data behavior, either it is normal or abnormal. In addition, better accuracy rate can be achieved using the merged approach, in which two or more machine learning algorithms from different classification and clustering techniques reintegrated to perform anomaly detection (Tsang, Kwong & Wang, 2007, Tsai & Lin, 2010).However, reducing false alarms remains as a challenging task for researchers in the area.

In recent years, integrated approaches have been widely explored. For instance, a Triangle Area based Nearest Neighbor (TANN) approach (Tsai & Lin, 2010) merged K-Means and K-Nearest Neighbor (K-NN) as anomaly detection. K-Means was used to group data into several clusters that represents a set of specific type of attacks. Later, K-NN Classifier was utilized to group data based on the characteristics of the triangle area. TANN increased the detection rate for a particular type of attacks, but underperform in decreasing the false alarm rate.

Horng et al. (2011) recently improved the Support Vector Machine (SVM) based IDS when they applied BIRCH Hierarchical clustering on the preprocessing stage to cluster data and feature selection method to exclude unimportant attributes. The generated SVM model is good in classifying some attack data correctly. However, the detection percentage for normal data decreased and as a consequence the approach yields a medium rate of false positives.

An extended version of Naives Bayes called Hidden Naives Bayes (HNB) Classifier (Koc, Mazzuchi & Sarkani, 2012) has been introduced to classify network attacks correctly. The author integrates HNB with various discretization and feature selection methods to increase the accuracy rate while at the same time decreasing the error rate. In HNB, the entire features are considered as independent or unbiased to each other. Each attribute in the HNB model has a hidden parent, in which the parameter on the training dataset estimates to unite the attributes that have relation with each other throughout conditional mutual information. HNB significantly increased the accuracy rate of attacks compared to other leading methods, such as decision tree, neural network, and sequential minimal optimization.

Chung and Wahid (2012) has proposed an Hybrid Network Intrusion Detection System (NIDS) using Intelligent Dynamic Swarm based Rough Set (IDS-RS) for feature selection and Simplified Swarm Optimization (SSO) for attack classification. First, the most important features that symbolize network packet patterns are extracted using IDS-RS. Later, a novel Weighted Local Search (WLS) scheme is used to enhance the performance of SSO classifier. WLS is responsible for mining intrusion patterns to determine the appropriate solution based on the neighborhood of the present solution generated by SSO. This approach offers reasonable accuracy rate but it could be further improved.

An intrusion detection framework based on Random Forest (RF) and K-Means algorithm has been proposed by Elbasiony et al. (2013).RF was used as a misuse (signature) detection to construct an intrusion pattern from a training dataset. This pattern will be compared with incoming network connection to capture possible intrusion activities. To perform anomaly-based detection, the author applied the weighted K-Means algorithm. These algorithm partitions network packets into several *K*-clusters based on the relationship of data points wherein some of these clusters set will be considered as an anomaly cluster. This approach offers high detection rate with low false positive rate.

So far, various methods in anomaly detection domain has been employed; but interestingly most of them evaluate their approaches with the outdated KDD Cup 1999 dataset. In contrast, our proposed integrated based approach has been tested on the more current ISCX dataset to demonstrate that it is able to increase the attack and normal detection rate while keep maximizing the false alarm rate.

K-MEANS CLUSTERING AND NAÏVE BAYES CLASSIFIER (KMC+NBC)

Machine learning method used in anomaly-based detection in recent years promises high detection and accuracy rate. However, the rate for false alarms also increases accordingly. KMC+NBC is able to detect intrusive activities and focuses to achieve high detection and accuracy rate with lower false alarm.

There are two modules in KMC+NBC; namely the pre-classification module and the classification module. The first module, involving K-Means clustering iteration function where similar data are grouped into several clusters based on their behavior. The entire data are labeled with the *K*-th clusters set accordingly. Next, the labeled clustered data are classified into attack and normal classes using the NB classifier to recover the misclassified data from the first module. We find that KMC+NBC is able to classify the attack and normal data more accurately at the subsequent classification module.

K-Means Clustering (KMC)

The principle goal of employing the K-Means clustering scheme is to separate the collection of normal and attack data that behave similarly into several partitions which is known as *K*-th cluster centroids. In other words, K-Means estimates a fixed number of *K*, the best cluster centroid representing data with similar behavior. In our work, we predefined K=3, representing Cluster 1, Cluster 2, and Cluster 3. Thus, the iterative K-Means algorithm is designed as follows:

```
Initially: Randomly select K = 3 cluster centroid
Do
    Correspond data point to nearest clusters
    Update optimal cluster centroid based on corresponding data
    points and labeling the points
While no change remains
```

Certain activities or data are alike to either normal or abnormal behaviour. The K-Means algorithm is unable to differentiate this behaviour precisely. Thus, we applied NB classifier to re-classified clustered labelled data to improve the shortcoming.

Naïve Bayes Classifier (NB)

NBC derives a conditional probability for independent variable based on strong independence assumption that is useful for classification task. In the NB classification, a set of attributes or features assigned to a set of classes based on Bayes theorem as in Eq. (1). The goal of computing the probability for each class is to find the conditionally probability for a given set of observed attributes.

$$P(C|A) = P(A|C).P(C) / P(A) \qquad (1)$$

Let's determine *P(C|A)* equal to the probability of attributes for a given class *P(A|C)* multiplied with the probability of the class *P(C)* over the probability of attribute *P(A)*. The probability of class for a given observed attributes *P(C|A)*,also called posterior probability and is usually used to find the class for particular attributes. *P(C)* can be referred as prior probability. By Bayes rules, *P(C|A)* also can be referred as *P(A|C).P(C)/P(A)*. We simplified Bayes rule in Eq. (2). In traditional Bayesian classification, the *P(A)* is always identical and constant for each class. Thus, the denominator *P(A)* can be eliminated.

$$P(A_1, A_2, \ldots A_n|C) = P(A_1|C).\ P(A_2|C) \ldots .P(A_n|C) \qquad (2)$$

The *P(A|C)* represents the probability vector of the attribute given the class where the set of attributes equal to A_1, $A_2, \ldots A_n$.*For P(Ai|Ci)* we assume the attributes probability *Ai* are independent for the given class *Ci*. The joint probabilities of all set of attributes conditional on class *C* as the product of all bunches of independent probabilities. In this work, we classify the entire data into two classes (*C1* = Normal and *C2* = Attack).

EXPERIMENT AND RESULTS

Dataset Description

The ISCX 2012 intrusion evaluation dataset is used to perform the experiments and evaluate the performance of the proposed approach for anomaly detection. The entire ISCX labeled dataset comprises nearly 1512000 packets with 20 features and covered seven days of network activity (*i.e.* normal and intrusion). Further description on this dataset can be found in (Shiravi, Shiravi, Tavallaee & Ghorbani, 2012). Since the ready-made training and testing dataset is not available, and difficult to perform experiments on huge data, we decided to select incoming packets for a particular host and particular days to validate the proposed approach as presented in Table 1. The training data contains 75372 normal traces and 2154 attack traces while the testing data contains 19202 normal traces and 37159 additional attack traces.

Table 1. Distribution of training and testing data

(Host: 192.168.5.122)

Date	**Training Data**		**Testing Data**	
	Normal	Attack	Normal	Attack
11th	0	0	147	0
12th	22612	0	0	0
14th	16260	1973	0	0
15th	0	0	19115	37159
16th	22879	0	0	0
17th	13621	181	0	0
Total	**77526**		**56421**	

Evaluation Measurement

Generally, the performance of the anomaly detection can be evaluated using four major criteria; that is *true positive (tp)* when an intrusion activity is identified as intrusion, *true negative (tn)* when an intrusion activity is identified as legitimate, *false positive (fp)* when a legitimate activity is identified as an intrusion activity, and *false negative (fn)* when intrusion is mistakenly identified as legitimate activity. Subsequently, detection and accuracy rate as well as false alarm can be computed as follows; *accuracy=(tp+tn)/(tp+tn+fp+fn), detection rate=(tp)/(tp+fp),* and *false alarm=(fp)/(fp+tn).*

Result and Discussion

Tables 2 and 3 exhibit the outcome dimension across *tp*, *tn*, *fp* and *fn* achieved from NBC and the proposed KMC+NBC using the training and testing data. KMC+NBC outperforms the NBC in identifying attack and normal data more accurately.

Table 2. Outcome dimension using training dataset

Dataset	Method	Result			
Training	NBC	*tp*	*tn*	*fp*	*fn*
		2145	62068	13304	9
	KMC+NBC	*tp*	*tn*	*fp*	*fn*
		2145	75269	103	9

Table 3. Outcome dimension using testing dataset

Dataset	Method	Result			
Testing	NBC	*tp*	*tn*	*fp*	*fn*
		37047	12762	6500	112
	KMC+NBC	*tp*	*tn*	*fp*	*fn*
		37050	18826	436	109

NBC is less efficient in distinguishing some attacks (i.e. R2L and L2L) and normal data that are almost the same with one another. Thus, the NBC yields lower accuracy and detection rate with higher false alarms. In contrast, KMC+NBC achieved higher accuracy and detection rate as well as lower false alarm rate as shown in Table 4. The clustering function employed as pre-classification element for gathering comparable data into corresponding categories assists KMC+NBC to produce superior outcome as compared to NBC. Besides, KMC+NBC also permits disordered data throughout the first module to be re-ordered again, hence refining the accuracy and detection rate while minimizing false alarms. For example, in the testing environment KMC+NBC enhances the accuracy, detection rate and false alarm of NBC, which shows an increment of +10.8%, +13.8% and decrement of false alarm up to - %17.47, respectively. In short, KMC+NBC are proven to be more efficient than NBC.

Table 4. NBC Vs. KMC+NBC Using Training and Testing Dataset

Method	**Training Data**			**Testing Data**		
	accuracy	detection rate	false alarm	accuracy	detection rate	false alarm
NBC	82.8	13.8	17.6	88.2	85.0	33.7
KMC+NBC	99.8	95.4	0.13	99.0	98.8	2.2

CONCLUSION AND FUTURE WORK

In this work, we proposed an integrated machine learning method by merging K-Means clustering and Naïve Bayes classifier (KMC+NBC) for anomaly detection in IDS. KMC+NBC were assessed using the ISCX 2012 intrusion detection evaluation benchmark dataset. The essential resolution is to split the data among the conceivable attack and normal data into different clusters in earlier module. The labeled clustered data are later re-classified into specific classes, namely attack classes and normal classes. KMC+NBC substantially increase the accuracy and detection rate while keeping the false alarm reduced. The proposed method is capable to categorize data accurately, except for attack types L2L and R2L. Therefore, for future enhancement, we are considering the expansion of KMC+NBC by encompassing feature selection method which has been geared with different elements and purposes.

REFERENCES

Elbasiony, R.M., Sallam, E.A., Eltobely, T.E., & Fahmy, M.M. (2013). A hybrid network intrusion detection framework based on random forests and weighted k-means. *Ain Shams Engineering Journal.* Available online 7 March 2013, ISSN 2090-4479, 10.1016/j.asej.2013.01.003.

Horng, S-J., Su, M-Y., Chen, Y-H., Kao, T-W., Chen, R-J., Lai, J-L., & Perkasa, C. D. (2011). A novel intrusion detection system based on hierarchical clustering and support vector machines. *Expert Systems with Applications*, 38(1), 306-313.

Koc, L., Mazzuchi, T.A., & Sarkani, S. (2012). A network intrusion detection system based on a Hidden Naïve Bayes multiclass classifier, *Expert Systems with Applications*, 39(18), 13492-13500.

Lee, W., Stolfo, J. S., & Mok, W.K. (1998). A data mining framework for adaptive intrusion detection, *Proceedings of the lEEE Symposium on Security and Privacy*, I20-132.

Chung, Y.Y., & Wahid, N. (2012). A hybrid network intrusion detection system using simplified swarm optimization (SSO). *Applied Soft Computing*, 12(9), 3014-3022.

Patcha, A., & Park, J-M. (2007). An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Computer Networks*, 51(12), 3448-3470.

Shiravi, A., Shiravi, H., Tavallaee, M., & Ghorbani, A.A. (2012). Toward developing a systematic approach to generate benchmark datasets for intrusion detection. *Computers & Security*, 31(3), 357-374.

Tsai, C-F., & Lin, C-Y. (2010). A triangle area based nearest neighbors approach to intrusion detection. *Pattern Recognition*, 43(1), 222-229.

Tsang, C-H., Kwong, S., & Wang, H. (2007). Genetic-fuzzy rule mining approach and evaluation of feature selection techniques for anomaly intrusion detection. *Pattern Recognition*, 40(9), 2373-2391.

Yingbing, Y. (2012). A survey of anomaly intrusion detection techniques. *Journal of Computer Science*, 28(1), 9-17.

Carlos, A. C., Carlos, G. G. (2012). Automatic network intrusion detection: Current techniques and open issues. *Computers & Electrical Engineering*, 38(5), 1062-1072.

Panda, M., Abraham, A., Patra, .M. R. (2012). A Hybrid Intelligent Approach for Network Intrusion Detection, *Procedia Engineering*, 30, 1-9.