Journal of Theoretical and Applied Information Technology
20th October 2014. Vol. 68 No.2
© 2005 - 2014 JATIT & LLS. All rights reserved·
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

REDUCING FALSE ALARM USING HYBRID INTRUSION DETECTION BASED ON X-MEANS CLUSTERING AND RANDOM FOREST CLASSIFICATION

[1]SUNDUS JUMA, [1]ZAITON MUDA, [2]WARUSIA YASSIN
[1]Faculty of Computer Science and Information Technology, University Putra Malaysia,
43400 UPM Serdang, Selangor, Malaysia
[2]Faculty of Information and Communication Technology, University Technical Malaysia Melaka,
Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia
E-mail: [1]p.sundus@hotmail.com, [1]zaitonm@upm.edu.my, [2]warusia@gmail.com

ABSTRACT

In recent times, Intrusion Detection systems (IDSs) incarnate the high network security. Anomaly-based intrusion detection techniques, that utilize algorithms of machine learning, have the capability to recognize unpredicted malicious. Unluckily, an essential provocation of this method is to maximize accuracy, detection whereas minimize false alarm rate. This paper proposed a hybrid machine learning approach based on X-Means clustering and Random Forest classification called XM-RF in order to aforementioned drawbacks. X-Means clustering is utilized to gather whole data into congruent cluster based on their behaviour whereas Random Forest classifier is utilized to rearrange the misclassified clustered data to apropos group. The ISCX 2012 Intrusion Detection Evaluation is used as a model dataset. The experimental result pose that the proposed approach obtains better than other techniques, with the accuracy, detection and false alarm rates of 99.96%, 99.99%, and 0.2%, respectively.

Keywords: *Intrusion Detection System, Anomaly-based Intrusion Detection, Machine Learning, X-Means, Random Forest.*

1. INTRODUCTION

Securing information in various areas of applications has become a fundamental issue. System vulnerabilities and precious information entice attackers' consideration. Cyber attacks have significantly risen lately. Proficient intrusion detection is required to prevent these vexatious activities. Consequently, intrusion detection system (IDS) has been introduced and become one of fundamental component in computer security to detect these malicious with the aim of preserving systems from common harms and grouping vulnerabilities of the intruded system [1].

IDSs can be categorized into signature-based detection (SBD) and anomaly-based detection (ABD) [2]. SDB can detect recognized attacks, similar to antivirus applications in detecting viruses. Unfortunately, the drawbacks of this technique are incapacity to detect a novel attacks and the need of continuous updating pattern signature of attack when there are enteric attacks. Alternatively, ABD determines a variation from the normal usage patterns as intrusion. Regrettably, most of this technique suffers from high rate of false alarms [1][3].

In order to overcome the aforesaid drawbacks, many ABD are promoted based machine learning algorithms [4]. Nevertheless, the high rate of false alarm stays as a major restriction in building a proficient ABD in these works [5]. Consequently, since false alarm may make IDS undependable, there is a need to detect attacks more perfectly to decrease the rate false alarm.

In this paper, a combination of machine learning algorithms based on X-Means clustering and Random Forest classification called XM-RF has been proposed in order to execute the defiance task of ABD in decreasing false alarm and increasing detection rate and accuracy. This approach is more proficient compared to prior techniques which correlated with high false alarm. XM-RF performance has been compared with Random Forest separately and other hybrid approaches that are testing using the same dataset, ISCX 2012.

The rest of the paper is arranged as follows: Section 2 describes the related work, Section 3 explained the proposed approach, and Section 4 presents the experimental results while Section 5 conclusion and future work.

© 2005 - 2014 JATIT & LLS. All rights reserved·
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2. RELATED WORK

Machine learning methods have been employed in the field of anomaly detection to identify whether the behaviour of data is normal or abnormal. Moreover, the reasonable accuracy and detection rate can be earned by employing the combinational approach, when as a minimum two algorithms of machine learning various clustering and classification procedures are gathered to perform anomaly detection [1][6]. Though, the main challenge in this area for researchers is decreasing false alarms.

In recent year, number of hybrid approaches has been vastly investigated. The accuracy, true positive, true negative, false positive, false negative, false alarm, detection rate have been explored too.

A combination of decision tree and support vector machine (DT-SVM) was used as a hierarchical hybrid intelligent system to build efficient IDS [7]. Despite the fact that DT-SVM can produce high percentage of detection rate, it is incapable to differentiate normal behaviour from attacks.

Random forest algorithm was used to design an anomaly detection structure to detect uninhabitable intrusion. As contrast to other stated unsupervised anomaly detection approaches, the proposed anomaly approach obtained high percentage of detection rate low rate of false positive. When the number of attack grows, the performance of detection also decreased [8].

In [1], hybrid machine learning was proposed through triangle area-based nearest neighbours (TANN). The approach was combined K-Means clustering and K-NN classifier to detect attacks efficaciously. Particularly, at the first stage K-Means was utilized to group a number of cluster centres that represents one definite kind of attacks. Secondly, the K-NN classifier was applied based on features of the triangle areas. TANN raised detection rate to 98.95% while came with high false alarm rate at 3.83%.

Author [3] utilized K-Means clustering and OneR classifier to generate hybrid machine learning called (KM-1R). The essential resolution is breaking up instances into attack and the normal instance through a first stage to various clusters. Afterwards, the clusters are assorted into Probe, R2L, U2R, DoS and Normal classes. The proposed approach attained low false alarm rate at 2.73% and elevated accuracy and detection rate at 99.26% and 99.33% respectively.

The hybrid approach of random forests classifier and Synthetic Minority Oversampling Technique (SMOTE) proposed by [9]. Random forests was used to developing proficient and effectual IDS while SMOTE was applied to enhance the detection rate of R2L and U2R classes in imponderables training dataset then single out the whole of the fundamental features of the minority classes by using R2L and U2R classes attack mode. By using this approach, the time required to build the model was decreased and the detection rate of R2L and U2R was increased to 0.963and 0.962 respectively but the detection rate could be promote enhanced.

Author [10] proposed a combination of K-Means clustering and Naïve Bayes Classifier (KM-NB) for anomaly detection in IDS. KM-NB approach considerably enhanced the accuracy to 99% and detection rate to 98.8%, whereas declined false alarm until 2.2%.

Author [11] used the hybrid model of Support Vector Machine (SVM) and genetic algorithms (GA) to evaluate IDS's performance by examining the most representative parameters which they are accuracy and false/true alarms. The result registered high accuracy rate to detect intrusion and good quality percentage of false positive, true positive, false negative and true negative compared to using SVM alone.

To sum up, a variety of hybrid techniques in intrusion detection based anomaly detection has been proposed; nevertheless, there is a need to boosting the detection rate and accuracy whereas decreasing the rate of false alarm.

3. PROPOSED HYBRID LEARNING APPROACH

Recently, machine learning techniques have commonly utilized in anomaly based detection (ABD) due to providing high average of accuracy and detection rate. Nonetheless, the rate of false alarm is compounding regularly. XM-RF is capable to detect invasive activities and obtaining high accuracy and detection rate with low false alarm rate.

Figure 1 shows the general steps for the proposed approach. Firstly, the hybrid approach analogous data instances based on their behaviours are grouped using X-Means clustering. Secondly, the consequential clusters are categorized into attack and classes as the last classification task by Random Forest classifier. Interestingly, the misclassified data from the first stage may be

Journal of Theoretical and Applied Information Technology
20th October 2014. Vol. 68 No.2
© 2005 - 2014 JATIT & LLS. All rights reserved·
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

accurately categorized in the following classification stage.

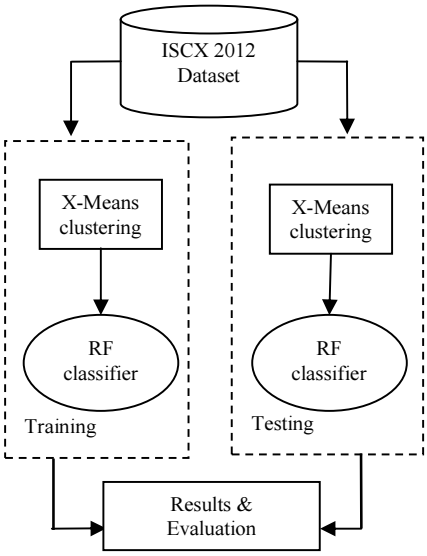

Figure 1: Steps of Proposed Approach

3.1 X-Means Clustering

X-Means clustering is the extended version of K-means clustering. X-Means was proposed in order to resolve K-Means's shortcomings; it scaling weakly computationally, the volume of clusters K has to be provide by the user and the explore inclined to local minima [12].

X-Means utilizes a Bayesian Information Criterion (BIC) in order to verify the clusters number centroids that can model the data effectively. Furthermore, X-Means uses KD-trees data structure for speed. Lastly, the user can determined the distance function to apply, the minimum and maximum volume of clusters to weigh, and the maximum volume of iterations to conduct.

Initially, X-Means algorithm starts with K. In this work, K = 3 to cluster the data into three clusters (C1, C2, C3). First, K has to be equivalent to the lower bound of the specified assortment and maintaining to add centroids where they are mandatory till the upper bound is reached. During this process, the centroid set that attain the finest score is listed, and this is one that is finally output. The general steps in the X-Means algorithm are shown in Figure 2.

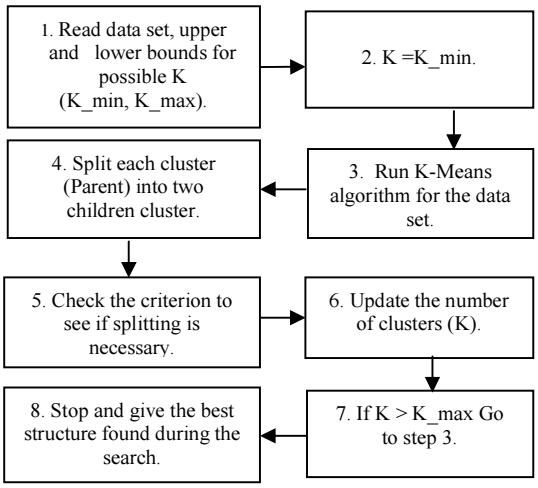

Figure 2: General steps in the X-Means algorithm

3.2 Random Forest Classifier

The random forests algorithm is a classification algorithm containing a assortment of tree structure formed classifiers, where each tree casts a unit vote for the most popular class at each input [13].

Each tree is grown as follows:

1. If the volume of cases in the training set is N, a sample of N cases is reserved at random from the data. This sample will be the training set for growing the tree.
2. If there are M input variables, a number m M is specified such that at each node, m variables are selected at random out of the M input variables, then, the best split on these m is used to split the node. The value of m is held constant during the forest growing.

Each tree is grown to the largest extent possible. There is no pruning [13].

4. EXPERIMENT AND RESULTS

4.1 Dataset Description

In this research, the hybrid method has evaluated using ISCX 2012 benchmark dataset for anomaly detection. The entire ISCX dataset contain nearly 1512000 packets with 20 features with seven days of network activity (normal and intrusion). Additional explanation regards the dataset can found in [14]. Using this dataset, there is no ready testing and training dataset provided, thus we select the incoming packet from the main server among the node on specific days to validate the proposed approach as in Table 1. The services running on the main server are Web, Mail, DNS and SSH. The training data divided into 75372 and 2154 normal and attack traces respectively. The testing data

© 2005 - 2014 JATIT & LLS. All rights reserved·
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

divided into 19202 and 37159 normal and attack traces respectively.

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	

4.2 Evaluation Measurement

An effective intrusion detection system (IDS) always required with high proportion of accuracy and detection rate with minimum false alarm rate. Normally, IDSs performance is evaluate in term of accuracy, detection rate, and false alarm rate as follow:

Accuracy = (TP+TN) / (TP+TN+FP+FN)
Detection Rate = (TP) / (TP+FP)
False Alarm = (FP) / (FP+TN)

The dimension of data behaviour in IDS for normal and attack classes are shown in Table 2.

Table 2: General Behaviour of Intrusion Detection Data

Actual	Predicted Normal	Predicted Attack
Normal	TN	FP
Intrusions (attacks)	FN	TP

- True positive (TP): an attack data identified as an attack.
- True negative (TN): a normal data identified as normal.
- False positive (FP): a normal data identified as an attack.
- False negative (FN): an attack data identified as normal.

4.3 Result and Discussion

Table 3: Classification Result for RF Using Training Dataset

Actual	Predicted Normal	Predicted Attack
Normal	75368	4
Intrusions (attacks)	7	2147

Table 4: Classification Result for XM-RF Using Training Dataset

Actual	Predicted Normal	Predicted Attack
Normal	75371	1
Intrusions (attacks)	6	2149

Table 3 and 4 represent results across binary dimension classes obtained from RF and XM-RF against training dataset. RF is less precise when the algorithm falsely detected 4 instances as attacks and 7 instances as normal as contrast to XM-RF with merely 1 instance and 6 instances, respectively.

Table 5: Classification Result for RF Using Testing Dataset

Actual	Predicted Normal	Predicted Attack
Normal	19241	21
Intrusions (attacks)	27	37132

Table 6: Classification Result for (XM-RF Using Testing Dataset

Actual	Predicted Normal	Predicted Attack
Normal	19258	4
Intrusions (attacks)	16	37143

In the case of binary dimension class detection for testing dataset, XM-RF performance is much better than RF, where merely 4 normal instances indentified as attack and merely 16 attacks instances were detected as normal. Thus, RF resulted in 21 false positive and 27 false negative as illustrated in Table 5. In summary, single classifier contributes in increasing false alarm rate compared to hybrid approach.

Table 7 and 8 illustrate the experimental results to evaluate the proposed hybrid approach against single classifier (XM-RF versus RF). Figure 1, Figure 2, and Figure 3 report performance measurement in term of accuracy, detection rate, and false alarm against the training and testing dataset. By using training dataset, RF produced almost the same accuracy with XM-RF but with lower detection rate and higher false alarm. In testing environment, hybrid approach increased the accuracy and detection rate by +0.05% whereas lowering down false alarm rate up to -0.09%. In contract, single classifier obtained 99.91%, 99.94% and 0.11% respectively. To sum up, RF undergoes in high false alarm compared to XM-RF. In general, the excellence and efficiency of anomaly based detection is assessed by the false alarm value.

Journal of Theoretical and Applied Information Technology
20th October 2014. Vol. 68 No.2
© 2005 - 2014 JATIT & LLS. All rights reserved·
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The smaller amount of false alarm values the further proficient the anomaly based detection model.

Table 7: XM-RF versus RF Using Training dataset

Method	Training Data		
	Accuracy	Detection Rate	False Alarm
RF	99.99	99.81	0.01
XM-RF	99.99	99.95	0.00

Table 8: XM-RF versus RF Using Testing dataset

Method	Testing Data		
	Accuracy	Detection Rate	False Alarm
RF	99.91	99.94	0.11
XM-RF	99.96	99.99	0.02

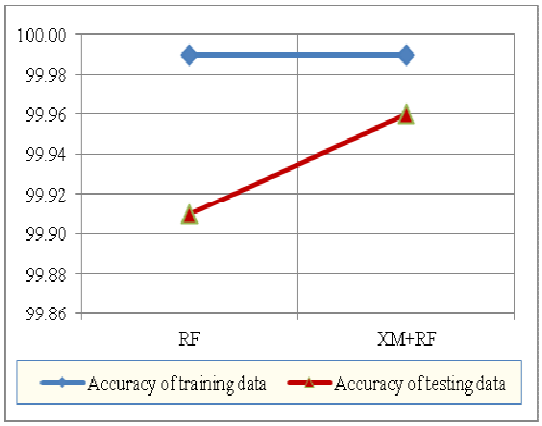

Figure 3: Accuracy for RF and XM-RF

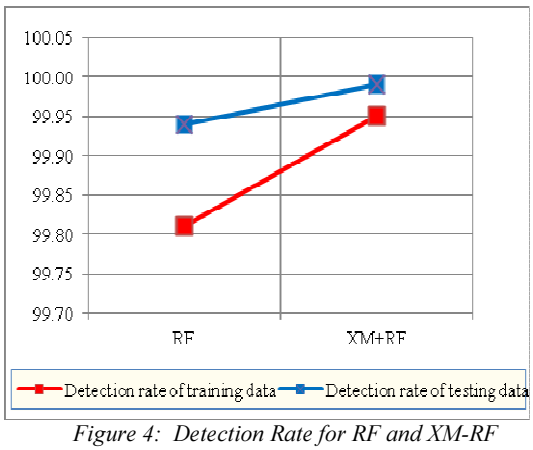

Figure 4: Detection Rate for RF and XM-RF

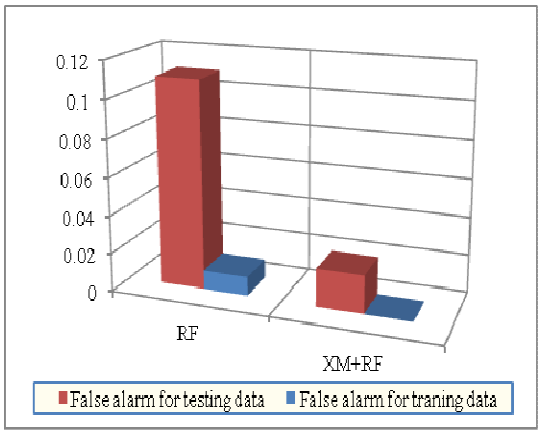

Figure 5: False Alarm for RF and XM-RF

Further comparison between the proposed method and other's has been done using the ISCX 2012 dataset in term of accuracy, detection rate as well as false alarm. Thus, the selected methods are X-Means Clustering and Naïve Bayes Classifier (XM-NB), and X-Means Clustering and OneR Classifier (XM-1R). As in Table 9, it is clearly shown that combination of X-Means Clustering and Random Forest Classifier (XM-RF) give much better result in term of accuracy, detection rate and false alarm. It is also conclude that Random Forest is the best classifier compared to Naïve Bayes and OneR.

Table 9: Other Hybrid Approach versus. XM-RF

Measurement	Other Hybrid Approaches		Proposed Approach
	XM-NB	XM-1R	XM-RF
Accuracy	88.27	93.68	99.96
Detection Rate	85.07	95.20	99.99
False Alarm	33.76	9.26	0.02

5. CONCLUSION AND FUTURE WORK

In this paper, a hybrid machine learning approach was proposed by gathering X-Means clustering and Random Forest classification (XM-RF) for anomaly-based detection in IDS. XM-RF evaluated using the ISCX 2012 intrusion detection evaluation benchmark dataset. The fundamental decree is to divide the attack and normal instances into clusters at the first stage. Then, the labeled clustered data are re-classified into attack classes and normal classes. XM-RF significantly enhances the accuracy and detection rate whereas lowering down the false alarm. This work has been applied

© 2005 - 2014 JATIT & LLS. All rights reserved·
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

on small dataset. Therefore, for future enhancement, using a larger dataset is recommended and also using online detection programs.

REFRENCES:

[1] C.F. Tsai and C.Y. Lin, "A triangle area based nearest neighbors approach to intrusion detection," *Pattern Recognition*, Vol. 43, 2010, pp. 222-229.

[2] W. Lee, J. S. Stolfo and W.K. Mok, "A data mining framework for adaptive intrusion detection," *Proceedings of the lEEE Symposium on Security and Privacy*, New York, USA, 1998, pp. 120-132.

[3] Z. Muda, W. Yassin, M.N. Sulaiman and N.I. Udzir, "Intrusion Detection based on K-Means Clustering and OneR Classification,". *7th International Conference on Information Assurance and Security (IAS)*, Melaka, Malaysia, 2011, pp. 192-197.

[4] A. Patcha and J.M. Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends," *Computer Networks*, vol. 51, no.12, 2007, pp. 3448-3470.

[5] A.C. Carlos and G.G. Carlos, "Automatic network intrusion detection: Current techniques and open issues," *Computers and Electrical Engineering*, vol. 38, no.5, 2012, pp.1062-1072.

[6] C.H. Tsang, S. Kwong and H. Wang, "Genetic-fuzzy rule mining approach and evaluation of feature selection techniques for anomaly intrusion detection," *Pattern Recognition* , Vol. 40, 2007, pp.2373–2391.

[7] S. Peddabachigaria, A. Abrahamb, C. Grosanc and J. Thomas, "Modelling intrusion detection system using hybrid intelligent systems," *Computer Applications*, vol.30, 2007, pp. 114–132.

[8] J. Zhang, M. Zulkernine and A. Haque, "Random-Forests-Based Network Intrusion," *IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews*, vol. 38, 2008, pp. 649-659.

[9] A. Tesfahun and D.L. Bhaskari, "Intrusion Detection using Random Forests Classifier with SMOTE and Feature Reduction," *International Conference on Cloud & Ubiquitous Computing & Emerging Technologies*, Visakhapatnam, AP, India, 2013, pp. 127-132.

[10] W. Yassin, N.I. Udzir, Z. Muda and M.N. Sulaiman, "Anomaly-Based Intrusion Detection through K-Means Clustering and Naives Bayes Classification," *Proceedings of the 4th International Conference on Computing and Informatics (ICOCI)*, Sarawak, Malaysia, 2013, pp. 298-303.

[11] 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*, Sarawak, Malaysia, 2013, pp. 312-320.

[12] D. Pelleg and A. Moore, "X-means: Extending K-means with Efficient Estimation of the Number of Clusters," *ICML '00 Proceedings of the Seventeenth International Conference on Machine Learning,* San Francisco, USA: Morgan Kaufmann, 2000, pp. 727-734.

[13] R.M. Elbasiony, E.A. Sallam, T.E. Eltobely and M.M. Fahmy,"A hybrid network intrusion detection framework based on random forests and weighted K-means," *Ain Shams Engineering Journal*, vol. 4, 2013, pp. 753–762.

[14] A. Shiravi, H. Shiravi, M. Tavallaee and A.A. Ghorbani, "Toward Developing a Systematic Approach to Generate Benchmark Datasets for Intrusion Detection," *Computers & Security*, vol. 31, no. 3, 2012, pp. 357-374.