2011 7th International Conference on IT in Asia (CITA)

Intrusion Detection based on K-Means Clustering and Naïve Bayes Classification

Z. Muda, W. Yassin, M.N. Sulaiman, N.I. Udzir
Faculty of Computer Science and Information Technology
University Putra Malaysia
43400 UPM Serdang, Selangor Darul Ehsan, Malaysia

***Abstract*--Intrusion Detection System (IDS) plays an effective way to achieve higher security in detecting malicious activities for a couple of years. Anomaly detection is one of intrusion detection system. Current anomaly detection is often associated with high false alarm with moderate accuracy and detection rates when it's unable to detect all types of attacks correctly. To overcome this problem, we propose an hybrid learning approach through combination of K-Means clustering and Naïve Bayes classification. The proposed approach will be cluster all data into the corresponding group before applying a classifier for classification purpose. An experiment is carried out to evaluate the performance of the proposed approach using KDD Cup '99 dataset. Result show that the proposed approach performed better in term of accuracy, detection rate with reasonable false alarm rate.**

Keywords-Intrusion Detection system; Anomaly Detection; Hybrid Learning; Clustering; Classification

I. INTRODUCTION

With the rapid growth of network technology, a cyber crime incident has also grown accordingly. A wide range of risks and threats against uncontrolled and undefended assets such as database and web server as well as entire network system become the general concern for intruders nowadays. Gaining unauthorized access to files, network and any other serious security threat can be detected by employing Intrusion Detection System. IDS identify any activity that violates the security policy from various areas within computer and network environment. An IDS is capable of sending early alarm upon risk exposure caused by any attack. This is to alert the system administrators to execute corresponding response measurements, thus to reduce the possibility of bigger losses. There are two traditional IDSs used to detect intruders: signature-based detection and anomaly-based detection [1]. A signature-based IDS match define signature with each anayzed packets on the network to detect known malicious attack as a same way like a virus scanner. These type of IDS required a frequent updating for the new signatures to keep the signature database up-to-date. Thus, it fails in discovering and detect an unknown attacks once the signature did not exist in its library. Unlike signature-based detection, anomaly-based detection is designed to capture any activities which are diviates the normal usage pattern called normal profile. If any activity deviates the normal profile it will be considered as intrusion. Anomaly-based detection have an ability to detect unforseen attacks, but at the same time it has potential to generate high volume of false alarm.

In recent years, data mining approach have been proposed and used as detection techniques for discover unknown attacks [2]. This approach has resulted in high accuracy and good detection rates but with moderate false alarm on novel attacks. In addition, some attacks and normal connections are even failed to be detected correctly. Therefore, there is a need to detect and identify such attacks accurately in an interconnected network.

In this work, we propose a hybrid learning approach based on combination of K-Means clustering and Naïve Bayes classification to improve current anomaly-based detection capabilities in the term of accuracy, detection rate as well as false alarm rate. The proposed approach is evaluated using KDD Cup '99 benchmark dataset and compared with single classifier and previous findings. The rest of the paper is organized as follows: in section 2, related works of this field are discussed. We describe the proposed model in section 3. Experimental results and comparison are presented in section 4. Finally, the conclusion and future work is presented in section 5 respectively.

II. RELATED WORK

Data mining is the latest technology introduced in network security enviroment to find regularities and irregulaties in large datasets [3, 4]. KDD CUP '99 dataset is the dominating evaluation dataset used by most of researcher to test their proposed techniques. The best possible accuracy and detection rate can be achieved by using Hybrid learning approaches [5]. However, the work to improve false alarm rate is an ongoing affair. Different classifiers can be use to formed a hybrid learning approaches such as combination of clustering and classification technique [6].

Clustering is an anomaly-based detection method that is able to detect novel attack without any prior notice and is capable to find natural grouping of data based on similarities among the patterns [7]. Reference [6] employ K-Means clustering to cluster data instances into k-clusters. Next, the research trains the new dataset, which consist of only the centers of cluster with Support Vector Machine (SVM). Reference [8] use K-Means and DBScan to efficiently identify a group of traffic behaviors that are similar to each other using cluster analysis.

978-1-61284-130-4/11/$26.00 ©2011 IEEE

Reference [9] has state that Naïve Bayes classifiers provide a very competitive result even this classifier having a simple structure on his experimental study. According to the author, Naïve Bayes are more efficient in classification task. Naïve Bayes classifier for anomaly-based network intrusion detection has proposed in [10]. He demonstrates that Naïve bayes classifier more efficient in detecting network intrusion compare to neural network.

Various data mining algorithm are compared in [11] to detect network intrusion. The author concluded that data mining approaches can increase the detection rate as well as reducing the false alarm with reasonable rate.

A comprehensive set of classifiers evaluated for detecting four type of attack category which are available on the KDD dataset [12]. The best classifier for each attack category has been chose and two appropriate classifier proposed for their selection models. Reference [13] proposed the best performed classifier for each category of attack by evaluates a comprehensive set of different classifier using the data collected from Knowledge Discovery Database (KDD).

III. HYBRID LEARNING APPROACH

Anomaly learning approaches are able to detect attacks with high accuracy and to achieve high detection rates. However, the rate of false alarm using anomaly approach is equally high. In order to maintain the high accuracy and detection rate while at the same time to lower down the false alarm rate, we proposed a combination of two learning techniques.

For the first stage in the proposed hybrid learning approach, we grouped similar data instances based on their behaviors by utilizing a K-Means clustering as a pre-classification component. Next, using Naïve Bayes classifier we classified the resulting clusters into attack classes as a final classification task. We found that data that has been misclassified during the earlier stage may be correctly classified in the subsequent classification stage.

A. K-Means Clustering

Network intrusion class labels are divided into four main classes, which are DoS, Probe, U2R, and R2L [14]. Fig. 1(a) to Fig. 1(d) shows the steps involved in K-Means clustering process. Fig.2 will later show the final overall result with application of the classification approach.

The main goal to utilize K-Means clustering approach is to split and to group data into normal and attack instances. K-Means clustering methods partition the input dataset into k- clusters according to an initial value known as the seed-points into each cluster's centroids or cluster centers. The mean value of numerical data contained within each cluster is called centroids. In our case, we choose k = 3 in order to cluster the data into three clusters (C1, C2, C3). Since *U2R* and R2L attack patterns are naturally quite similar with normal instances, one extra cluster is used to group U2R and R2L attacks.

Back to Fig. 1(b), each input will be assigned to the closest centroid by squared distances between the input data points and the centroids. New centroids will then be generated for each cluster by calculating the mean values of the input set assigned to each cluster as shown in Fig. 1(c).

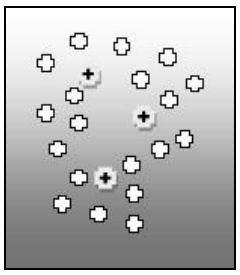

Figure 1(a). Seeds

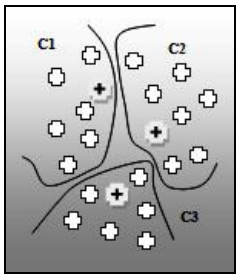

Figure 1(b). Assigns instances to cluster

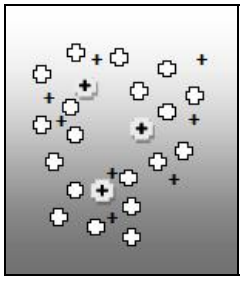

Figure 1(c). Finds

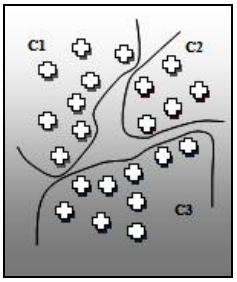

Figure 1(d). New centroid

Step in Fig. 1(b) and Fig. 1(c) are repeated until the result has reached convergence as shown in Fig. 1(d).

The K-Means algorithm works as follows:

- Select initial centers of the K clusters. Repeat step 2 through 3 until the cluster membership stabilizes.
- Generate a new partition by assigning each data to its closest cluster centers.
- Compute new clusters as the centroids of the clusters.

B. Naïve Bayes Classifier

Some behaviors in intrusion instances are similar to normal and other intrusion instances as well. In addition, a lot of algorithms including K-Means are unable to correctly distinguish intrusion instances and normal instances. In order to improve this shortcoming in classification, we combined K-Means technique with Naïve Bayes classifier. Naïve Bayes has became one of the most efficient learning algorithm [15]. Naïve Bayes are based on a very strong independence assumption with fairly simple construction. It analyzes the relationship between independent variable and the dependent variable to derive a conditional probability for each relationship. Using Bayes Theorem we write:

$$P(H|X) = P(X|H)\,P(H) / P(X) \quad (1)$$

Let X be the data record. Let H be some hypothesis represent data record X, which belongs to a specified class C. For classification, we would like to determine P(H|X), which is the probability that the hypothesis H holds, given an observed data record X. P(H|X) is the posterior probability of H conditioned on X. In contrast, P(H) is the prior probability. The posterior probability P(H|X), is based on more information such as background knowledge than the prior probability P(H), which is independent of X. Similarly, P(X|H) is posterior probability of X conditioned on H. Bayes theorem is useful because it provides ways to calculate the posterior probability P(H|X) from P(H), P(X), and P(X|H).

We consider 5 category classes (C1 = Normal, C2 = DoS, C3 = Probe, C4 = R2L, and C5 = U2R). Given X, predict C1, C2, C3, C4, and C5. The Bayes rule is shown in Equation (2).

$$P(C_i|X) = \frac{P(X|C_i).P(C_i)}{P(X)} \quad (2)$$

where C_i represents the category of classes and X is data record. X may be divided into pieces of instances, say x_1, x_2, ..., x_n which are related to the attributes X_1, X_2, ..., X_N, respectively. The probability obtained is shown in the following Equation (3).

$$P(C_i|X) = \frac{P(x_1|C_i).P(x_2|C_i)\ldots. P(x_n|C_i).\ P(C_i)}{P(X)} \quad (3)$$

The denominator P(X) always constant for all classes. Thus, it can be ignored as in Equation (4).

$$P(C_i|X) = P(x_1|C_i).P(x_2|C_i)\ldots. P(x_n|C_i).\ P(C_i) \quad (4)$$

Fig.2 shows Naïve Bayes classifier that are used to classify classifies all 3 clusters as illustrated in Fig. 1(d) into more specific categories, which are Probe, Normal, Dos, U2R, and R2L. Combination of these classifier with K-Means clustering technique showed an encouraging improvement as compared to previous approaches. The results are surprisingly better in term of accuracy, detection rate, and false alarm rate.

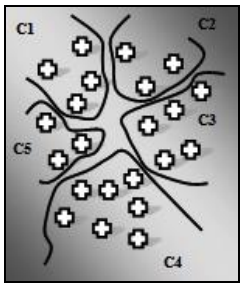

Figure 2. Classifier

IV. EXPERIMENTS AND RESULTS

A. Dataset Description

In our experiments, the KDD Cup'99 benchmark dataset [16] is chosen for evaluation and comparison between the proposed approaches and the previous approaches. The entire KDD data set contains an approximately 500,000 instances with 41 features. The training dataset contains 24 types of attack, while the testing data contains more than 14 types of additional attack. Further description for the available features and intrusion instances can be found in [17].

KDD dataset covered four major categories of attacks which is Probe, DoS, R2L and U2R. In order to demonstrate the abilities to detect different kinds of intrusions, the training and testing data covered all classes of intrusion categories as listed in the following as adopted from the [16].

Table I and Table II summarizes the distribution records for training dataset according to class type. In order to validate the overall hybrid learning approach overall, a testing dataset is also used.

TABLE I. SAMPLE DISTRIBUTION OF THE TRAINING DATASET

Class	No. of Samples	Sample Percentage (%)
Normal	97277	19.69
Probe	4107	0.83
DoS	391458	79.24
U2R	52	0.01
R2L	1126	0.23
Total	494020	100

TABLE II. SAMPLE DISTRIBUTION OF THE TESTING DATASET

Class	No. of Samples	Sample Percentage (%)
Normal	60593	19.4
Probe	4166	1.33
DoS	231455	74.4
U2R	88	0.028
R2L	14727	4.73
Total	311029	100

B. Evaluation Measurement

An Intrusion Detection System (IDS) requires high accuracy and detection rate as well as low false alarm rate. In general, the performance of IDS is evaluated in term of accuracy, detection rate, and false alarm rate as in the following formula:

$$\text{Accuracy} = (TP+TN) / (TP+TN+FP+FN) \quad (5)$$

$$\text{Detetion Rate} = (TP) / (TP+FP) \quad (6)$$

$$\text{False Alarm} = (FP) / (FP+TN) \quad (7)$$

Table III shows the categories of data behavior in intrusion detection for binary category classes (Normal and Attacks) in term of true negative, true positive, false positive and false negative.

TABLE III. GENERAL BEHAVIOR OF INTRUSION DETECTION DATA

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

- True positive (TP) when attack data detected as attack
- True negative (TN) when normal data detected as normal
- False positive (FP) when normal data detected as attack
- False negative (FN) when attack data detected as normal

C. Result and Discussion

Table IV represent the results across all category classes obtained from Naïve Bayes (NB) and proposed hybrid learning approach K-Means with Naïve Bayes (KM+NB) using the training and testing sets. KM+NB performed better than the single classifier NB in detecting Normal, Probe, and DoS instances. Since Normal, U2R, and R2L instances are similar to each other, KM+NB recorded a comparable result for R2L instances except for U2R instances.

TABLE IV. CLASSIFICATION RESULT FOR EACH CATEGORY CLASS USING TRAINING AND TESTING DATASET

Dataset	**Training**		**Testing**	
Methods	NB	KM+NB	NB	KM+NB
Normal	91.6	99.6	81	99.5
Probe	99.8	100	95.6	98.3
DoS	94.3	99.5	82.5	99.6
U2R	80	40	80	80
R2L	65.5	61.6	90.3	83.2

Table V and Table VI represent results across binary category classes obtained from NB and KM+NB using training dataset. NB is less efficient when the algorithm falsely predicted 818 data as attacks and 471 data as normal as compared to KM+NB with only 40 data and 39 data respectively.

TABLE V. DETECTION RESULT FOR THE NORMAL AND ATTACK CLASSES USING TRAINING DATASET (NB)

Actual	**Predicted Normal**	**Predicted Attack**
Normal	8909	818
Intrusions (attacks)	471	39204

TABLE VI. DETECTION RESULT FOR THE NORMAL AND ATTACK CLASSES USING TRAINING DATASET (KM+NB)

Actual	**Predicted Normal**	**Predicted Attack**
Normal	9687	40
Intrusions (attacks)	39	39636

In the case of binary class detection for testing dataset, KM+NB performed better than NB as observed from Table VII, where 49 normal data was detected as attack and only 139 attacks data was detected as normal. On the contrary, NB resulted in 1852 false positive and 6448 false negative as shown in Table VIII. In short, NB contribute in increasing false alarm rate as compared to KM+NB.

TABLE VII. DETECTION RESULT FOR THE NORMAL AND ATTACK CLASSES USING TESTING DATASET (NB)

Actual	**Predicted Normal**	**Predicted Attack**
Normal	7875	1852
Intrusions (attacks)	6448	33227

TABLE VIII. DETECTION RESULT FOR THE NORMAL AND ATTACK CLASSES USING TESTING DATASET (KM+NB)

Actual	**Predicted Normal**	**Predicted Attack**
Normal	9678	49
Intrusions (attacks)	139	39536

Table VIIII shows the measurement in terms of accuracy, detection rate, and false alarm using the training and testing sets of both single classifiers and hybrid learning approach. We can see that single classifier produced a slightly higher accuracy and detection rate but with high false alarm rates as well. Meanwhile, the hybrid approach recorded high accuracy and detection rate with low false alarm percentage. The clustering techniques used as a pre-classification component for grouping similar data into respective classes helped the proposed hybrid learning approach to produce better results as compared to single classifier. The hybrid approach also allows misclassified data during the first stage to be classified again, hence improving the accuracy and detection rate with acceptable false alarm. For instance, the hybrid learning approach enhances the accuracy for single classifier especially for KM+NB combination, which shows an increase of +16.41% while reducing the false alarm rate up to -18.5%. On the contrary, NB classifier only achieved 83.19% and 19%

respectively. In short, NB suffers in high false alarm rate as compared to KM+NB.

TABLE VIIII. SINGLE CLASSIFIERS VS. HYBRID APPROACH USING TRAINING AND TESTING DATASET

Dataset	Training		Testing	
Methods	NB	KM+NB	NB	KM+NB
Accuracy	97.39	99.84	83.19	99.6
Detection Rate	97.95	99.89	94.7	99.8
False Alarm	8.4	0.41	19	0.5

Table X show further comparisons made for the proposed hybrid learning approach using the same KDD Cup '99 dataset as in previous researches in term of accuracy (AC), detection rate (DR), false positive (FP) and false alarm (FA).

TABLE X. FURTHER COMPARISON WITH PREVIOUS FINDINGS

Approaches	AC	DR	FP	FA
KM+NB (K-Means+Naïve Bayes)	99.6	99.8	0.09	0.5
Hierarchical Clustering and SVM [18]	95.7	N/A	0.7	N/A
TANN [6]	96.91	98.95	0.8	3.83
KM-KNN [6]	93.55	98.68	0.98	4.79
Hybrid Classifier [19]	96.78	99.21	3.2	3.2
ESC-IDS [20]	65.48	95.3	N/A	1.9

Intrusion detection system based on Hierarchical Clustering and Support Vector Machine (SVM) has proposed by author [18] recently. It can be noticed that our approach achieve the best false positive and accuracy rate compared to this approach.

Author [6] proposed the Triangle Area Nearest Neighbor (TANN) and K-Means with K-Nearest Neighbor (KM-KNN) approach for better intrusion detection. This approaches showed a reasonable detection rate compare to our approach. Unfortunately, a potential drawback of this technique is the rate of false alarms.

Moreover, unlike our approach, the system proposed by author [19] have tendencies to misclassify an normal data as an attack. Thus, this approach generates high false positive rate compare to our approach.

The Evolutionary Soft Computing based Intrusion Detection System (ESC-IDS) which focuses to detect and classify intrusion has proposed by Author [20]. This approach has serious shortcomings in its low accuracy rate as well as the tendency to produce high false alarm compare to our approach.

Overall, the proposed approach detected better percentage of attacks than the rest as proven in Table X with above 99.0% of accuracy and detection rate and below 0.5% of false alarm. This is attributed to K-Means clustering technique that has been used as a pre-classification. K-Means clustering helped to group similar data respectively so the misclassified data instances during the first clustering stage were able to be correctly classified in the second stage. The hybrid learning approach is proven to be more efficient as compared to previous approach that is synonym with high false alarm rates.

V. CONCLUSION AND FUTURE WORK

In this paper, an hybrid learning approach through combination of K-Means clustering and Naïve Bayes classifier is proposed. The proposed aproach was compared and evaluated using KDD Cup '99 benchmark dataset. The fundamental solution is to separate instances between the potential attacks and the normal instances during a preliminary stage into different clusters. Subsequently, the clusters are further classified into more specific categories, for example Probe, R2L, U2R, DoS and Normal. Hybrid learning approach achieved very low false alarm rate with an average below than 0.5%, while keeping the accuracy and the detection rate on average higher than 99%. The approach are capable to classify all data correctly except for attack type U2R and R2L. In the future, we recommend considering the Hybrid Intrusion Detection System which is better at detecting R2L and U2R attacks. The misuse detection approach better at detecting R2L and U2R attacks more efficiently as well as anomaly detection approach, which is better at detecting attacks at the absence of match signatures as provided in the misuse rule files.

REFERENCES

[1] W. Lee, J. S. Stolfo, and W. K. Mok, "A Data mining framework for adaptive intrusion detection," Proceedings of the 1999 lEEE Symposium on Security and Privacy, pp.120-132, 1999.

[2] A. Patcha and J-M Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends," Computer Network, 2007.

[3] Solahuddin, "Applying knowledge discovery in database techniques in modeling packet header anomaly intrusion detection systems," Journal of Software, 2008, 3(9): 68-76.

[4] M. Xue and C. Zhu, "Applied research on data mining algorithm in network intrusion detection," International Joint Conference on Artificial Intelligence, 2009.

[5] 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, 2007, 40:2373–2391.

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

[7] Y. Li and L. Guo, "An active learning based on TCM-KNN algorithm for supervised network intrusion," Computer and Securtiy, 2007, 26: 459-467.

[8] R. Luigi, T.E. Anderson, and N. McKeown, "Traffic classification using clustering algorithms. In Proceedings of the ACM SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, Sept. 11-15, Pisa, Italy, ACM Press, pp. 281-286.

[9] B.A. Nahla, B. Salem, and E. Zied, Naive bayes vs decision trees in intrusion detection systems. In Proceeding of the ACM Symposium on Applied Computing, Nicosia, Cyprus, 2004.

[10] P. Mrutyunjaya, and R. P. Manas, "Network intrusion detection Using naïve bayes," International Journal of Computer Science and Network Security, 2007, 7(12):258-263.

[11] M. Panda and M.R. Patra, "A comparative study of data mining algorithms for network intrusion detection. In Proceedings of

ICETET, India, 2008, pp.504-507.

[12] N. H. Anh, and C. Deokjai, "Application of data mining to network intrusion detection: classifier selection model," Lecture Notes in Computer Science. 2008, 5297:399-408.

[13] G. Meera and S. K. Srivatsa, "Classification algorithms in comparing classifier categories to predict the accuracy of the network intrusion detection – a machine learning approach", Advances in Computational Sciences and Technology. 2010, (3):321–334.

[14] R.P. Lippmann, D.J. Fried, I. Graf, J.W. Haines, K.R. Kendall, D. McClung, D. Weber, S.E. Webster, D. Wyschogrod, R.K. Cunningham, and M.A. Zissman, (2000). Evaluating intrusion detection systems: The 1998 DARPA off-line intrusion detection evaluation. In Proceedings of the 2000 DARPA Information Survivability Conference and Exposition (DISCEX), Los Alamitos, CA, 2000, 2:12–26.

[15] H. Zhang and J. Su., "Naive bayes for optimal ranking", Journal of Experimental and Theoretical Artificial Intelligence. 2008, 20: 79-93.

[16] KDD. (1999). Available at http://kdd.ics.uci.edu/databases/ - kddcup99/kddcup99.html

[17] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone, Classification and regression trees. Monterey, CA: Wadsworth & Books/Cole Advanced Boks & Software, 1984.

[18] S-J Horng, M-Y Su and Y-H Chen, "A novel intrusion detection system based on hierarchical clustering and support vector machines," Expert Systems with Applications, 2011, 38:306–313.

[19] C. Xiang, P.C. Yong, and L.S. Meng, "Design of multiple level hybrid classifier for intrusion detection system using Bayesian clustering and decision tree," Pattern Recognition Letters, 2008, 29: 918-924.

[20] M. Toosi, "A new approach to intrusion detection based on an evolutionary soft computing model using neuro-fuzzy classifiers," Computer Communications, 2007, 30: 2201-221