Journal of Information Assurance and Security.
ISSN 1554-1010 Volume 7 (2012) pp. 347-354

Intrusion Detection with K-Means Clustering and OneR Classification

Z. Muda[1], W. Yassin[2], M.N. Sulaiman[3] and N.I. Udzir[4]

Faculty of Computer Science and Information Technology, Universiti Putra Malaysia,
43400 UPM Serdang, Selangor Darul Ehsan, Malaysia
zaiton@fsktm.upm.edu.my

***Abstract*: Detecting malicious activities remains an elusive goal and indispensable challenge with the growing of prevalence networks attacks. In recent years, much attention has been given to anomaly detection to perform intrusion detection. Unfortunately, the major challenge of this approach is to maximize detection, accuracy and to minimize false alarm; i.e. failure in detecting certain type of attacks correctly. To overcome this problem, we propose a hybrid learning approach through a combination of K-Means clustering and One-R classification. The approach clusters all data into corresponding groups which match their natural behavior. Later, the clustered data are classified into the correct category using One-R classification. The validity of this approach is verified using the KDD Cup '99 benchmark dataset. Our experimental results demonstrate that our proposed approach performs better than existing techniques, with the accuracy, detection and false alarm rates of 99.26%, 99.33%, and 2.73%, respectively.**

***Keywords*:** Intrusion Detection System, Malicious, Anomaly Detection, Hybrid Learning, Clustering, Classification.

I. Introduction

Securing information from intrusion and malicious access has become an essential requirement today with the rapid growth of network technologies and attack. Furthermore, the number of attacks has increased dramatically with the popularity of the Internet. Thus, in the last decade IDS has become a managing scheme for network security to detect various attack and research topic [1].

Misuse detection and anomaly detection is two techniques applied in IDS today [2]. Misuse detection detects known attacks by examining attack patterns much like antivirus software. However, they cannot detect unknown attacks and need to update their attack pattern signatures with new ones [3]. Anomaly detection, on the other hand, has the capability to detect unknown attacks by identifying any unusual activity pattern which deviates from the normal usage.

In recent years, machine learning techniques have been used for intrusion detection with increasing accuracy and detection rate [5-9]. Unfortunately, it suffers from high false alarm rates [4] – incorrectly predicting an intrusion as normal, and/or normal data as attack – and it has been a challenge in building effective anomaly detection. Relying on a single machine learning algorithm is insufficient to distinguish between real attacks and normal network visit [10]. Thus, in order to achieve better performance and overcome these drawbacks, a hybrid machine learning algorithm of K-means clustering and One-R classifier is proposed. We compare the performance of our approach with single classifier and previous works. The performances of the proposed approach are better in terms of accuracy, detection rate and false alarms.

The rest of the paper is organized as follows: Section 2 discusses the related works. A brief introduction and structure of the proposed approach is described in Section 3. In Section 4, we present the evaluation of our approach. The conclusion and future work are given in Section 5.

II. Related Work

Finding regularities and irregularities in huge datasets is the latest data mining technology introduced and explored widely by the researchers in network security environment [11]. Most researchers choose the KDD CUP '99 dataset to evaluate their proposed techniques. Hybrid learning approaches promise much better possible accuracy and detection rate [12]. However, the work to detect all types of attacks and improve false alarm rate is an ongoing concern. Hybrid learning can be generated when at least two learning techniques are combined to achieve the same objective. The first technique is used to obtain an intermediate result which will be used as the input for the second technique in order to produce the final output [13]. Clustering and classification techniques can be used to form hybrid learning approaches [4, 14-16]. Anomaly-based method such as clustering is able to detect previously unseen attacks and capable of finding natural groupings of data based on similarities among the patterns [17].

Related work and research publications based on hybrid and single approaches have been widely explored such as in [16, 18-28, 33-37]. The detection rate (DR), false positive (FP), false negative (FN), true positive (TP), false alarm (FA), and accuracy for each approach are also investigated. Each approach has distinctive strengths and weaknesses. Some approaches possess strength in detection but not in reducing false alarm, and vice versa. For instance, SVM performance which normally has heavy computational challenges for huge datasets improved when using feature selection method to

eliminating the unimportant features. This approach is called ~~as~~ Least Squares Support Vector Machine (PLSSVM) [23]. Even PLSSVM able to classify Normal and Probes types of attacks correctly, but the approach also misses a large number of dynamic attacks such as DoS and U2R which are very similar to the normal behavior.

A similar approach to PLSSVM SVM-based IDS with BIRCH hierarchical clustering as a preprocessing phase and simple feature selection procedure to eliminate the unimportant features [24]. SVM correctly classifies some data upon applying a feature selection procedure while the hierarchical clustering algorithm helps to improve the performance of SVM. However the prediction percentage for R2L and Normal data decreases dramatically when the model could not differentiate between these data.

Many complex practical problems in IDS are successfully solved using Artificial Neural Network (ANN). For example, ANN-based IDS using ANN and Fuzzy Clustering called FC-ANN is proposed to enhance detection capabilities [25]. Different ANN models are trained to formulate different models and different training subset generated through fuzzy clustering approach. Later, a fuzzy aggregation module is employed to aggregate the result. ANN learns each subset more precisely in correctly detecting dynamic attacks such as U2R and R2L. However, Naive Bayes offers better detection in detecting Probe attacks compared to this approach.

Artificial Immune Network and Radial Basis Function (RBF) Neural Network are combined and proposed as a novel Intrusion Detection algorithm in [26], where the cosine RBF neural network based on gradient descent learning process is first trained, followed by the identification of a hidden neuron candidate through multiple granularities artificial immune network. The experimental results indicate that this approach has an ability to get reasonable detection but it can be further improved.

Fuzzy SVMs (FSVM) based intrusion detection is used to improve classification accuracy in [27]. First, a new training set is constructed using centers of clusters through a clustering algorithm. Later, FSVM trained this new set to obtain support vector. This method has increased the accuracy rate, but it is not of an acceptable percentage.

SVM is combined with K-Means clustering to increase accuracy and detection rate in [16]. A new dataset trained with SVM which have only the centers of clusters after K-Means clustering groups all data into k-clusters. Unfortunately, this approach generates high false alarm instead of high accuracy and detection rate.

The best performed classifier for detecting each category of attack was proposed in [28] by evaluating a comprehensive set of different classifier using data collected from the Knowledge Discovery Database (KDD). Even though a number of techniques have been evaluated and the best classifier has been identified, but the accuracy rate can still be improved.

As stated in [29], data mining approaches can reduce false alarm as well as increase accuracy and detection rate. Although an effective learning algorithm have been proposed by various researchers in intrusion detection, generally there are still rooms to improve the accuracy and detection rate with low false alarm.

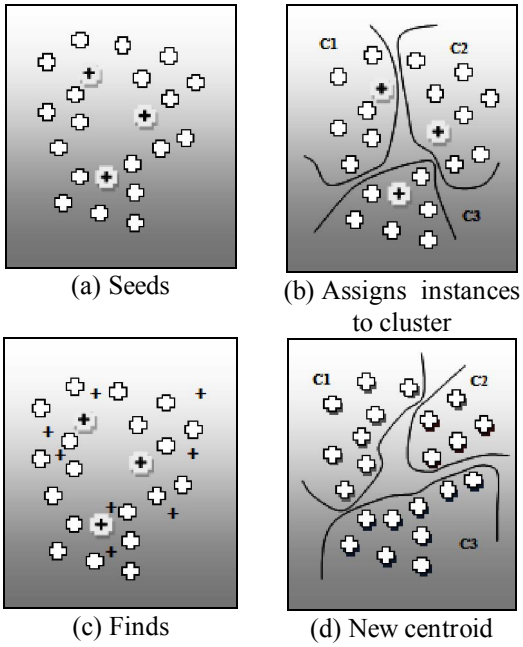
III. Hybrid Learning Approach

Learning approaches offers high accuracy and high detection rates for seen and previously unseen attacks. However, the rate of false alarm is also high. Thus, we proposed a combination of two learning techniques called KM+1R to reduce the false alarm rate while at the same time increase accuracy and detection rates. KM+1R have been deployed in a single running.

In the proposed approach, first we grouped similar data based on their natural behaviors using K-Means clustering as a pre-classification component. Next, we classified the resulting clusters into attack classes as a final classification task using One-R classifier. Misclassified data during the earlier stage are re-classified accordingly in the subsequent classification stage.

A. K-Means Clustering

DoS, Probe, U2R, and R2L are four main network intrusion attack classes [14]. Steps involved in K-Means clustering process is shown in Figure 1(a) through (d). The final classification result is presented in Figure 2.



(a) Seeds (b) Assigns instances to cluster

(c) Finds (d) New centroid

Figure 1. K-Means clustering process

We use K-Means clustering to split and group data into normal and attack instances. K-Means clustering partitions the input dataset into k-clusters according to an initial value known as the seed-points into each cluster's centroids (cluster centers), i.e. the mean value of numerical data contained within each cluster. In our approach, we cluster data into three clusters (C1, C2, C3) which is similar to k=3. One extra cluster is used to group U2R and R2L attacks, because naturally U2R and R2L attack patterns are quite similar with normal instances.

Each input will be assigned to the closest centroid by squaring distances between the input data points and the centroids as illustrated in Figure 1(b). The mean value of the

input set assigned to each clusters are calculated and new centroids will be generated for each cluster as shown in Figure 1(c). The steps in Figures 1(b) and (c) are repeated until the result reached a convergence as shown in Figure 1(d).

The K-Means algorithm works as follows:

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

B. OneR Classifier

In this technique, a set of classification rules on particular tested attributes will be generated by One-R based on the value of only a single attribute. The One-R algorithm chooses the attribute with the lowest error rate as its "one rule". A proportion of instances that do not belong to the majority class of the corresponding attribute value will contribute to the error rate.

The OneR algorithm works as follows:

1. From the clustered set, create a rule set for each value of each attribute predictor as in step i, ii, iii and iv.
   i. Count how often each value of the target class appears.
   ii. Find the most frequent class.
   iii. Make a rule set assigns that class to this value of attribute predictor.
   iv. Calculate the total error that occurs in the rule set for each attribute predictor.
2. Pick the best attribute predictors which have the smallest total error and make class attribute as a classification rules.

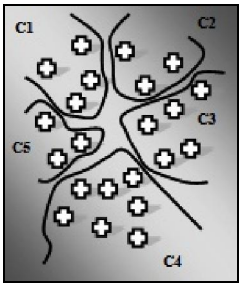


Figure 2. Classifier

Fig.2 shows OneR classifier that is used to classify all three clusters in Figure 1(d) into more specific categories, which are Probe, Normal, Dos, U2R, and R2L. The combination of these classifiers with the K-Means clustering technique showed an encouraging improvement as compared to previous approaches. The results are surprisingly better in terms of accuracy, detection rate, and false alarm rate.

IV. Implementation, Experiments and Result

A. Implementation

Learning approaches offer high detection rate and accuracy percentage for unknown attacks [30-32]. However, the limitation is the high false alarm rates. In order to achieve high accuracy and detection rate while at the same time maintaining the false alarm within an acceptable range, we propose a hybrid mining approach.

Figure 3 shows the experimental design and implementation of the proposed hybrid data mining approach. The process flow is divided into two stages. Stage 1 is designed for data preparation process while Stage 2 is designed for clustering and classification.

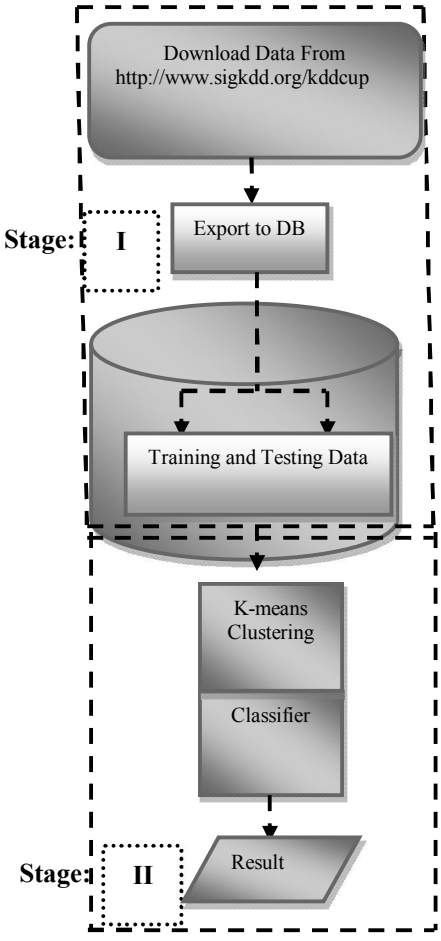


Figure 3. Experimental and implementation process flow

(i) Stage 1: Data Preparation

In this stage, training and testing data are downloaded from the ACM Special Interest Group from the Knowledge Discovery and Data Mining website. Data are available in the form of text file. We converted all data into comma separated values (*.csv*) and exported them in bulk into the SQL server database. In this work, the training set contained 24 types of known attacks as well as an addition of 14 types of unknown

attacks for the testing set, including *mailbomb*, *mscan*, and *snmpgetattack*. The entire training and testing sets are distributed between 311,029 to 494,020 records.

(ii) Stage 2: Clustering and Classification

In this stage, clustering techniques are used as a pre-classification component for grouping similar data into respective groups based on their behaviors. The main purpose of using the clustering approach is to accurately separate the attack data from the normal data. Next, the classifier model classified all clusters into five specific categories, which are *Probe*, *Normal*, *Dos*, *U2R*, and *R2L*. We implemented One-R classifier during this stage because we believed that all these classifiers will show an improvement in performance as compared to previous approaches. Furthermore, only One-R classifiers show significant improvements and much better results compared to others, namely SVM, Neural Network, Id3, NBTree, Zero-R and Linear Regression. This stage aimed for better accuracy and detection rate, as well as reducing false alarms. All the learning algorithms chosen are implemented using the Weka data mining application.

B. Dataset Description

In our experiments, the data used originate from KDD Cup'99 which is considered as a standard benchmark for evaluation of intrusion detection systems. We manage to compare and evaluate our approach with previous techniques. In order to demonstrate the abilities to detect different kinds of intrusions, the training and testing data covered all classes of intrusion categories as adopted from the KDD (1999), as follows:

- **Denial of Service (DoS):** Attacker usually occupies all system sources, disables system resources, and engages all computing or memory resources, rendering the system to be too busy to handle legitimate requests or deny legitimate users from accessing a machine. Examples of attacks are Smurf, Mailbomb, SYN Flooding, Ping Flooding, Process table, Teardrop, Apache2, Back, and Land.

- **Remote to User (R2L):** Attacker sends packets to remote machine over a network and exploits the network vulnerability to gain local access as a user of that machine. Examples of attacks are Ftp_write, Imap, Named, Phf, Sendmail, and SQL Injection.

- **User to Root (U2R):** Attacker takes the advantage of system leak by accessing a normal user's account on the system and exploits system vulnerabilities to get legal administrator access to the system. Examples of attacks are Loadmodule, Perl, Fdformat.

- **Probing:** Attacker performs some preparation step before launching attacks by scanning a network of computers to gather information or to find vulnerabilities. The attacker will use this information to determine the targets and the type of operating system. Examples of attacks are Nmap, Satan, Ipsweep, Mscan. (KDD dataset, 1999)

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, a testing dataset is also used.

Table I. Data Distribution of the Training Dataset

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

C. Evaluation Measurement

An efficient IDS requires high accuracy and detection rate as well as low false alarm rate. In general, the performance of an IDS is evaluated in terms of accuracy, detection rate, and false alarm rate as in the following formula:

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

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

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

Table III shows the categories of data behavior in intrusion detection for binary category classes (Normal and Attacks) in terms 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	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

D. Results and Discussion

A series of experiments have been conducted to compare the performance of a single classifier and previous approaches against our proposed hybrid approach using training and testing datasets. One-R (1R) has been used as the single classifier and combined with *K*-Means clustering pre-processing to form a hybrid approach known as *K*-Means+One-R (KM+1R). The experiments carried out managed to evaluate the proposed approach based on the drawbacks posed by the improved 1R classifiers in terms of accuracy, detection rate and false alarm, as presented in Tables IV through XI. In addition, we further compared the proposed hybrid approach against other related approach using the same KDD Cup '99 dataset as illustrated in Table XII.

Table IV presents the results across all types of data obtained from 1R and the proposed hybrid learning approach K-Means with 1R (KM+1R) using the training sets. Based on this table, KM+1R performed better than 1R classifier in detecting Normal, Probe, DoS, U2R and R2L data types. Behaviors of U2R, R2L and Normal in most cases are similar to each other. Thus, KM+1R recorded a low prediction for U2R based type of data compared to 1R which failed to detect any U2R data type correctly.

Table IV. Detection Result for Each Type of Data Using Training Dataset

Data	Normal	Probe	DoS	U2R	R2L
1R	95.12%	51.5%	99.68%	0%	13.27%
KM+1R	99.3%	93.4%	100%	60%	90.3%

Table V proved that 1R is less accurate when the algorithm falsely predicted 474 instances as attacks (false positive) and 414 instances as normal (false negative) as compared to KM+NB with only 68 instances (false positive) and 55 instances (false negative) respectively from Table VI. In short, 1R suffers from high false alarm rate as compared to KM+1R.

Table V. Detection Result for the Normal and Attack Classes Using Training Dataset (1R)

Actual	Predicted Normal	Predicted Attack
Normal	9253 or 95.12%	474 or 4.88%
Intrusions	414 or 1.04%	39261 or 98.96%

Table VI. Detection Result for the Normal and Attack Classes Using Training Dataset (KM+1R)

Actual	Predicted Normal	Predicted Attack
Normal	9659 or 99.3%	68 or 0.70%
Intrusions	55 or 0.14%	39620 or 99.86%

Table VII shows the measurement in terms of accuracy, detection rate, and false alarm using the training sets of both single classifiers (1R) and hybrid learning approach (KM+1R). We can see that 1R produced a slightly higher accuracy and detection rate but with high false alarm rates as well. Meanwhile, KM+1R 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 categories helped the proposed KM+1R to produce better results as compared to 1R. The KM+1R also allow misclassified data during the first stage to be classified again, hence improving the accuracy and detection rate with acceptable false alarm. For instance, the KM+1R enhances the accuracy and detection rate which shows an increase of +1.55%, +1.02% while reducing the false alarm rate up to -4.17%.

Table VII. Single Classifiers vs. Hybrid Approach Using Training Dataset

Measurement	1R	KM+1R
Accuracy	98.2	99.75
Detection Rate	98.81	99.83
False Alarm	4.87	0.70
Increment – Accuracy Rate		+1.55
Increment – Detection Rate		+1.02
False Alarm Reduced Rate		- 4.17

Table VIII represents the results across all type of data obtained from 1R and our proposed hybrid learning approach K-Means with 1R (KM+1R) using the testing sets. Based on this table, KM+1R outperformed the 1R classifier in predicting *Normal*, *Probe*, *DoS*, U2R and *R2L* data type. Due to difficulties in distinguishing *U2R* and *R2L data behavior*, the 1R classifier failed to classify any of *U2R* data type correctly while KM+1R records a low detection for this type of data.

Table VIII. Classification Result for Each Type of Data Using Testing Dataset

Data	Normal	Probe	DoS	U2R	R2L
1R	92.7%	73.2%	98.9%	0%	17.7%
KM+1R	97.26%	95.6%	99.8%	60%	91.2%

KM+1R performed better than 1R as observed from Table X where 266 normal data was detected as attack and only 99 attack data were detected as normal. On the contrary, 1R resulted in 709 false positives and 603 false negatives as illustrated in Table IX. In short, 1R contributes in increasing false alarm rate as compared to KM+1R.

Table IX. Detection Result for the Normal and Attack Classes Using Testing Dataset (1R)

Actual	Predicted Normal	Predicted Attack
Normal	9018 or 92.71%	709 or 7.29%
Intrusions	603 or 1.52%	39072 or 98.48%

Table X. Detection Result for the Normal and Attack Classes Using Testing Dataset (KM+1R)

Actual	Predicted Normal	Predicted Attack
Normal	9461 or 97.26%	266 or 2.74%
Intrusions	99 or 0.25%	39576 or 99.75%

From Table XI, it is evident that the KM+1R enhances the accuracy and detection rate for 1R, which shows an increase of +1.92%, +1.11 while reducing the false alarm rate up to -4.56%. This proves that the KM+1R are better in reducing misclassification constraints. The clustering techniques used as a pre-classification component for grouping similar data into respective classes helped the proposed hybrid approach to produce better results as compared to 1R classifier. The KM+1R also allows misclassified data during the first stage to be re-classified, hence improving the false alarm rate. These comparisons show that KM+1R are more suitable in building an efficient anomaly-based network intrusion detection model.

Table XI. Single Classifiers vs. Hybrid Approach Using Testing Dataset

Measurement	1R	KM+1R
Accuracy	97.34	99.26
Detection Rate	98.22	99.33
False Alarm	7.29	2.73
Increment Accuracy Rate		+1.92
Increment Detection Rate		+1.11
False Alarm Reduced Rate		- 4.56

Table XII shows further comparisons of the proposed hybrid learning approach using the same KDD Cup '99 dataset as used in previous researches in terms of accuracy (AC), detection rate (DR), false positive (FP) and false alarm (FA). Overall, the proposed approach performed better than the rest as proven in Table X with the accuracy, detection, and false alarm rates of 99.26%, 99.33%, and 2.73%, respectively. The KM+1R is proven to be more efficient as compared to previous approaches which are associated with high false alarm rates. This is attributed to the K-Means clustering technique used as pre-classification. K-Means clustering helped group similar data respectively so the misclassified data instances during the first clustering stage were able to be correctly classified in the second stage.

Table XII. Further Comparison with Previous Findings

Approaches	AC	DR	FA
KM+1R(K-Means+One-R)	99.26	99.33	2.73
Decision Tree+Wrapper [20]	98.38	N/A	N/A
GFR [19]	98.62	N/A	N/A
Feature Selection + SVM [23]	N/A	98.34	N/A
BIRCH Clustering + SVM [24]	95.70	N/A	N/A
ANN + Fuzzy Clustering [25]	96.71	N/A	N/A
ENLCID [18]	N/A	99.02	3.19
Fuzzy+GNP [21]	94.4	97.5	7.2
k-means-k-NN [16]	93.55	98.68	4.79
TANN [16]	96.91	98.95	3.83

V. Conclusion and Future Works

In this research, we have proposed a hybrid approach called KM+1R based on K-Means Clustering and One-R classifier to overcome problems inherent in current anomaly detection methods which are related to poor accuracy, detection and false alarm rate. We have evaluated the proposed method over the well known benchmark dataset KDD Cup '99. In our approach, all the data are grouped according to their behavior using K-Means clustering techniques, before applying the One-R classifier to re-classify all data into correct categories (Normal, R2L, U2R, Probe and DoS). The proposed approach shows better performance when compared to some recent related works. In addition, the clustering process as a preliminary stage in our works yields better generalization of accuracy, detection and false alarm upon utilizing One-R classifier for the classification purpose. More specifically, KM+1R are able to resolve incorrect detection issues for all attack types except for U2R and R2L. A possible future work may be directed towards increasing detection rate for the R2L and U2R types of attacks. Furthermore, the misuse detection approach is better at detecting R2L and U2R attacks. Hence, in future, we are considering the extension of our hybrid IDS by incorporating signature-based detection mechanism, which is better at detecting R2L and U2R attacks.

References

[1] C. Endorf, E. Schultz, J. Mellander. "Intrusion Detection & Prevention". McGraw-Hill/Osborne, 2004.

[2] J.M. Estevez-Tapiador, P. Garcia-Teodoro, J.E. Diaz-Verdejo. "Anomaly Detection Methods in Wired Networks: A Survey and Taxonomy", *Computer Communications*, Vol. 27(16), pp. 1569-1584, 2004.

[3] E. Tombini, H. Debar, L. Me, M. Ducasse. "A Serial Combination of Anomaly and Misuse IDSes Applied to Http Traffic". In *20th Annual Computer Security Applications Conference*, pp. 428-437, 2004.

[4] Y. Liu, K. Chen, X. Liao, W. Zhang. "A Genetic Clustering Method for Intrusion Detection", *Pattern Recognition*, Vol. 37(5), pp. 927-942, 2004.

[5] W. Lee, S. Stolfo, K. Mok. "A Data Mining Framework for Building Intrusion Detection Models". In: *Proceedings of the IEEE Symposium on Security and Privacy*, pp. 120–132, 1999.

[6] A. Zainal, M. A. Maarof , S. M. Shamsuddin. "Ensemble Classifiers for Network Intrusion Detection System", *Journal of Information Assurance and Security*, Vol. 4, pp. 217-225, 2009.

[7] Z. Bankovic, J. M. Moya, Á. Araujo, S. Bojanic, O. Nieto-Taladriz "A Genetic Algorithm-based Solution for Intrusion Detection", *Journal of Information Assurance and Security*, Vol. 4, pp. 192-199, 2009.

[8] G. Isazal, A. Castillo, M. López, L. Castillo. "Towards Ontology-Based Intelligent Model for Intrusion Detection and Prevention", *Journal of Information Assurance and Security*, Vol. 5, pp. 376-383, 2010.

[9] Y. Yang, D. Jiang, M. Xia. "Using Improved GHSOM for Intrusion Detection", *Journal of Information Assurance and Security*, (5), pp. 232-239, 2010.

[10] P. Tang, R. Jiang, M. Zhao. "Feature Selection and Design of Intrusion Detection System Based on k-Means and Triangle Area Support Vector Machine", In; *ICFN '10 Second International Conference on Future Network*, pp. 144-148, 2010.

[11] S.B. Shamsuddin. "Applying Knowledge Discovery in Database Techniques: Modeling Packet Header Anomaly Intrusion Detection Systems", *Journal of Software*, 3(9), pp.68-76, 2008.

[12] C.H. Tsang, S. Kwong, H. Wang. "Genetic-Fuzzy Rule Mining Approach and Evaluation of Feature Selection Techniques for Anomaly Intrusion Detection", *Pattern Recognition*, Vol. 40(9), pp. 2373–2391, 2007.

[13] J.S. Jang, C.T Sun, E. Mizutani. "Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence", *Prentice-Hall*, Englewood Cliffs, New Jersey, 1996.

[14] C. Zhang, J. Jiang, M. Kamel. "Intrusion Detection Using Hierarchical Neural Network", *Pattern Recognition Letters,* Vol. 26(6), pp.779–791, 2005.

[15] R. Luigi, T.E. Anderson, N. McKeown. "Traffic Classification Using Clustering Algorithms". In *Proceedings of the ACM SIGCOMM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, ACM Press, pp. 281-286, 2006.

[16] C.F. Tsai, C.Y. Lin. "A Triangle Area Based Nearest Neighbors Approach to Intrusion Detection", *Pattern Recognition*, 43(1), pp. 222-229, 2010.

[17] L. Yang, G. Li. "An Active Learning Based on TCM-KNN Algorithm for Supervised Network Intrusion", *Computer and Securtiy*, Vol. 26(7), pp.459-467, 2007.

[18] B. Kavitha, Dr. S. Karthikeyan, P. Sheeba Maybell. "An Ensemble Design of Intrusion Detection System for Handling Uncertainty Using Neutrosophic Logic Classifier", *Knowledge-Based Systems*, Vol. 28, pp.88-96, 2012.

[19] Y. Li, J. Xia, S. Zhang, J. Yan, X. Ai, K. Dai. "An Efficient Intrusion Detection System Based on Support Vector Machines and Gradually Feature Removal Method", *Expert System With Applications,* Vol. 39(1), 2012.

[20] S. Siva, S. Sivatha, S. Geetha, A. Kannan. "Decision Tree Based Light Weight Intrusion Detection Using A Wrapper Approach", *Expert Sysem. With Appications*. Vol. 39(1), pp.129-141. 2012.

[21] M. Shingo, C. Chen, L. Nannan, K. Shimada, K. Hirasawa. "An Intrusion-Detection Model Based on Fuzzy Class-Association-Rule Mining Using Genetic Network Programming," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* , Vol. 41(1), pp.130-139, 2011.

[22] W. Gang, H. Jinxing, M. Jian. "A New Approach to Intrusion Detection Using Artificial Neural Networks and Fuzzy Clustering", *Expert systems with applications*, Vol. 37(6), pp.6225–6232, 2011.

[23] F. Amiri, R.Y. Mohammad, S. Azadeh, Y. Nasser. "Mutual Information-Based Feature Selection for Intrusion Detection System", *Journal of Network and Computer Applications*, Vol. 34(4), pp.1184–1199, 2011.

[24] S.J. Horng. "A Novel Intrusion Detection System Based on Hierarchical Clustering and Support Vector Machines", *Expert Systems with Applications*. Vol. 38(1), pp.306-313, 2011.

[25] W. Gang, H. Jinxing, M. Jian. "A New Approach to Intrusion Detection Using Artificial Neural Networks and Fuzzy Clustering", *Expert systems with applications*, Vol. 3(76), pp.6225–6232, 2011.

[26] L. Cao, J. Zhong, Y. Feng. "Construction Cosine RBF Neural Networks Based on Artificial Immune Networks", *Lecture Notes In Computer Science*, pp.134-141, 2010.

[27] T. Shaohua, D. Hongle, W. Naiqi, Z. Wei, S. Jiangyi. "A Cooperative Network Intrusion Detection Based on Fuzzy SVMs", *Journal of Networks*, Vol. 5(4), pp.475–483, 2010.

[28] G. Meera, 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*, Vol. 3 (3), pp.321–334, 2010.

[29] M. Panda, M.R. Patra. "A Comparative Study of Data Mining Algorithms for Network Intrusion Detection". In *Proceedings of ICETET*, India, pp.504-507, 2008.

[30] Z. Muda, W. Yassin, M.N. Sulaiman, N.I. Udzir. "A K-means and naive bayes learning approach for better intrusion detection", Information Technology Journal, Vol. 10(3), pp.648-655,2011.

[31] Z. Muda,W. Yassin,M.N. Sulaiman, N.I. Udzir."Intrusion detection based on K-Means clustering and Naïve Bayes classification". In *7th Information Technology in Asia International Conference*,pp.1-6,2011.

[32] Z. Muda,W. Yassin, M.N. Sulaiman, N.I. Udzir. "Intrusion detection based on k-means clustering and OneR classification". In *7th Information Assurance and Security International Conference*,pp.192-197,2011.

[33] H. A. S. Altwaijry, "Bayesian Based Intrusion Detection System", *Journal of King Saud University - Computer and Information sciences*, Vol. 24, pp.1-6, 2012.

[34] V. Bolon-Canedo, N. Sanchez-Marono, A. Alonso-Betanzos, Feature selection and classification in multiple class datasets: An application to KDD Cup 99 dataset", *Expert Systems with Applications*, 38(5), pp. 5947-5957, 2011.
[35] M. A. Prabakar, R. Rajeswari, R. Rajaram, "Network Anomaly Detection by Cascading K-Means Clustering and C4.5 Decision Tree algorithm", *Procedia Engineering*, Vol. 30, pp. 174-182, 2012.

Author Biographies

Zaiton Muda received the B.Sc(1984) and M.Sc (1989) degrees in Computer Science from Universiti Kebangsaan Malaysia. She is a senior lecturer in Faculty of Computer Science and Information Technology, Universiti Putra Malaysia. She is the coordinator of External Education Unit, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia. Her research interests include computer security, parallel computing and intelligent computing.

Warusia Yassin is a PhD candidate from Universiti Putra Malaysia. He received his Bachelor of Computer Science (2008) and Master of Computer Science (2011) from UPM. His research interests are focused in data mining, intrusion detection and cloud computing. His profesional working experience include service as programmer, system engineer and security analyst.

Nur Izura Udzir is an associate professor at the Faculty of Computer Science and Information Technology, Universiti Putra Malaysia (UPM) since 1998. She received her Bachelor of Computer Science (1996) and Master of Science (1998) from UPM, and her PhD in Computer Science from the University of York, UK (2006). She is a member of IEEE Computer Society. Her areas of specialization are access control, secure operating systems, intrusion detection systems, coordination models and languages, and distributed systems. She is currently the Leader of the Information Security Group at the faculty.

Md. Nasir Sulaiman is an Associate Professor and the leader of the Intelligent Computing Research Group in the Dept. of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia. He obtained PhD in Neural Network Simulation from Loughborough University, U.K. in 1994. His research interests include intelligent computing, software agents, and data mining.