An Improved Malicious Behaviour Detection Via *k*-Means and Decision Tree

Warusia Yassin, Siti Rahayu, Faizal Abdollah and Hazlin Zin
Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka
Malaysia

***Abstract*—Data Mining algorithm which is applied as an anomaly detection system has been considered as one of the essential techniques in malicious behaviour detection. Unfortunately, such detection system is known for its inclination in detecting a cyber-malicious activity more accurately (i.e. maximizing malicious and non-malicious behaviours detection) and has become a persistent limitation in the deployment of intrusion detection systems. Consequently, these constraints will affect a number of important performance factors such as the accuracy, detection rate and false alarms. In this research, KMDT proposed as an anomaly detection model that utilized *k*-means clustering and decision tree classifier to maximize the detection of malicious behaviours by scrutinizing packet headers. The *k*-means clustering employed for labelling and plots the whole behaviours into identical cluster, which characterized the behaviours into suspicious or non-suspicious composition. Subsequently, these dissimilar clustered behaviours are reordered within two classes of types such as malicious and non-malicious via decision tree classifier. KMDT is a profitable finding which improved the anomaly detection performance in identifying suspicious and non-suspicious behaviours as well as characterizes it into malicious and non-malicious behaviours more accurately. These criteria have been validated by the result from the experiments throughout banking system environment dataset 2016. KMDT have detected more malicious behaviours accurately as contrast to discrete and diversely combined methods.**

***Keywords*—Intrusion Detection; Malicious Behaviours; Clustering; Decision Tree Classifier; Packet Headers**

I. INTRODUCTION

Safekeeping confidential information and computing assets from cyber threats, has turned into a foremost dispute as a consequence from sudden increases on network based malicious activities. As such, various Intrusion Detection Systems (IDSs) used to recognize, gather and analyse security infractions from diverse systems or networks [1]. In addition, the research societies have been classified these IDSs into misuse and anomaly based detection systems [2]–[6].

Misuse based detection recognizing acknowledge malicious traffic throughout the signatures which are defined and gathered earlier in the database. This facilitates the security personnel to easily create those signatures based on the seen behaviours of a malicious behaviours and determine which specific behaviours they want to detect [7]. Despite, the incapability to identifying the novel malicious behaviours remain as challenging task as these detection systems required frequent signature updates for each time novel behaviours are discovered [5].

Conversely, the research communities have claimed that the finest solution to observe unforeseen malicious behaviours without concerning signatures are with anomaly based detection systems [8]. This detection system is dependent on forming of ordinary behavioural models and conclude any attempt that is not covered within this model as malicious behaviours [9]. Nonetheless, higher false alarm or false positives are the main imperfection of this detection systems [10].

Seeing facts, for last two decades, an anomaly detection system that utilizes data mining approaches has attracted researcher interest, particularly within the concept of intrusion detection [10]–[12]. Nevertheless, maximizing the true positive (malicious behaviours which detected as malicious) and true negative (non-malicious behaviours which detected as non-malicious) as well as minimizing the false positive (non-malicious behaviours which detected as malicious) and false negative (malicious behaviours which detected as non-malicious) are not much enhanced as a whole. Consequently, this situation drags into poor performance in term of detection rate, accuracy and false alarm [13], [14] and as a result, research in data mining as anomaly detection system particularly for malicious behaviours is still in a process of improvement [15], [16].

In this research, *k*-means and Decision Tree, namely KMDT proposed in order to prevail over the aforementioned inadequacy. Even if so, much more focuses has been given to identify malicious activity than the non-malicious activities because failure in detecting malicious behaviours could lead to the losses of confidential information and computing assets. KMDT uniquely designed with clustering and classification scheme in scrutinizing malicious and non-malicious behaviours more accurately.

Figuratively, the foremost contributions of this paper include: (i) *k*-means clustering as an initial stage able to accumulate the similar and dissimilar malicious behaviours into unalike clusters. (ii) The dissimilar clustered behaviours are reorganized with decision tree classifier to increase the malicious prediction rate. By utilizing this two detection approach, the performance of the detection metric such as accuracy, detection rate and false alarm has improved.

The rest of the paper is organized as follows: In Section 2, the former similar fundamental works are briefly represented. The proposed detection model methodology has detailed out in Section 2. The experimental analysis discussed in Section 4 while conclusion and future work are covered in Section 5.

II. Related Work

Data mining approach extensively discovered and applied as detection methods in these few years in a field of an anomaly detection system. The major concentration of these whole detection methods is direct to observe, differentiate and identify i.e malicious or non-malicious behaviours [6]. Procedure of discovering fundamental patterns and concealed relationships systematically from valuable information within the data to visualize and model interrelationship of the data itself are known as data mining process [17]. Clustering and classification are common data mining algorithms and widely discovered and employed within the field of intrusion detection [6], [18], [19].

Clustering is ordinarily applied in anomaly detection to explore assemblage without prior knowledge on relationship within the data. Thus, clustering groups objects based on characterization of data points, where every single data point in a cluster is alike to those within its cluster, but different from those in different or other clusters [20], [21]. Clustering has capabilities to group similar malicious data points collectively into one or more cluster including previously unseen malicious data points [22]. In addition, *k*-means clustering is much more efficient as a contrast to other existing clustering methods as the fact that this algorithm able to process huge volumes of instances quickly and have non-linear complexity. In contrary, the challenges of such method are to define the *k* centroids for every single cluster, in which different location of an centroids possible to produce different outcome that is not much favourable [23]–[25] has tested and proven that the *k*-means algorithm is capable to detect malicious behaviours with high detection rates. They applied the algorithm to categorize normal and abnormal data points into different clusters to detect malicious behaviours more correctly, identifying unknown malicious behaviours without having prior knowledge. However, this algorithm also usually contributes in increasing false positive rates [26] even though it shows a high success rate in identifying suspicious behaviours. Such a high rate of false positives could downgrade the detection performance of IDSs as too many alerts are generated for non-suspicious behaviours wrongly detected as malicious data. Therefore, classification is introduced to reclassify the entire behaviours processed through *k*-means during the pre-processing stage [27].

Classification is a supervised learning method intended to build a detection model which could explain a data class, sort out data points or elements into classes that correspond to their features based on a sequence of pre-determined class label [13], [22]. Consequently, the structured label elements will be further used to predict the class label of new elements. For example, using training data with a pre-determined label to predict each class label in testing data [28]. Among various classification techniques, Decision Tree is faster and able to construct rules that are easy to interpret and understand [29]. However, a single classifier's impact or contributions is improved when it is integrated with other data mining algorithms even though each of these methods individually has been proven to achieve well in intrusion detection [6].

Integrated forms basically are made up of multiple clustering and classification algorithms, where clustering is executed at the beginning as a pre-processing method to reduce the noise within the dataset or to label the data for the subsequent classification stage [30], [31]. In another word, clustering is used in earlier phases to separate or label data into different clusters with a reasonable clustered data rate, so that the classifier can perform better to classify those data more correctly in the next phase. K-means clustering and Naive Bayes classifier has proposed as a an anomaly detection method. As an initial stage, the clustering method has applied to isolate a group of malicious and non-malicious activities which act analogously and un-analogously while Naive Bayes applied in the further stage to re-organize the clustered data more precisely into rightful class categories. The aim of above mentioned clustering algorithm is to minimize squared error-function, so that the optimal distance between two different data points and cluster centroids could be calculated more effectively. On the other hand, the naive bayes algorithm used to scrutinize the relationship between dependent and independent variables in deriving a conditional probability of every single relationship. This method has been evaluated and its slightly enhanced the detection performance in term of accuracy, the detection rate and false alarm [32]. A variety of combinations have been efficiently applied as intrusion detection recently. For instance, [31] integrated both k-means and k-nearest neighbour (KNN) algorithms to identify a specific form of attacks in an anomaly detection method named Triangle Area based Nearest Neighbour (TANN). K-means clustering is employed to categorize data into specific sets and type of attacks, and further reassembled through the K-NN classifier based on a feature of the triangle area. TANN is capable to identify certain types of attacks, but performs moderately in reducing the false alarm rate. Furthermore, *k*-means clustering has been combined with a decision tree algorithm as a network anomaly detection method [33]. In the initial phase, *k*-means clustering has applied to split the training data into k disjoint clusters, in which every single cluster correspond to a boundary of comparable data. Next, the above data is input into a decision tree algorithm to be classified either as normal or anomaly. The method contributes notable true positive and false positive rates with a reasonable accuracy rate which can be further improved. In contrast to these methods, an innovative anomaly detection method that distinctively combined *k*-means clustering, Naïve Bayes as feature selection and the Kruskal-Wallis statistical test is used to discover important features prior to classification. Then, this subset is classified with the Decision Tree algorithm to identify intrusion with the maximum amount of accuracy [6]. Furthermore, in [34] performance analysis on Decision Tree and Naïve Bayes for data classification has been conducted. The aim of this comparison is to find the most accurate classifier with high true positive and lower false positive rates. The author has concluded that the efficiency and accuracy of Decision Tree classifier are much better than Naïve Bayes. In addition, in [35] conducted a

similar analysis for intrusion detection using well-known data mining classifiers such as Decision Tree, Bayesnet, OneR and Naïve Bayes. In their study, they discovered that the best classification algorithm which could be applied for intrusion detection particularly novel intrusions is Decision Tree. Most researchers have claimed that the Decision Tree classifier is much more effective than others in maximizing novel malicious behaviour detection Such as from [29], [34]–[37]. Therefore Decision Tree classifier considered in this work.

In summary, even though the data mining approach, particularly clustering and classification mechanism has been applied widely in detecting malicious and non-malicious behaviours, the shortcomings such as maintaining the highest detection, accuracy and false alarm rates prohibit in assembling a proficient detection method [38], [39]. Moreover, these constraints still exist as a result of that, the whole focuses is not being given to maximizing the unseen and seen malicious as well as non-malicious behaviours identification rate. Therefore, there is also a critical requirement in designing efficient anomaly detection to identify malicious and non-malicious behaviours more accurately.

III. K-Means and Decision Tree Anomaly Detection Method

The proposed detection method *k*-means clustering and Decision Tree classifier called (KMDT) uses packet header information for anomaly detection which relies on a series of methods, i.e. employing k-means clustering in labelling data based on a clustered arrangement as a pre-processing stage and decision tree for classifying data which is affected with suspected outliers or miss-classified data during the pre-processing stage. The KMDT aim to identify the malicious behaviours more accurately and overcome the drawbacks currently faced in research in anomaly detection method. The process involves in each stage is as follows:

A. K-means Clustering

An unsupervised algorithm such as clustering method directly to discover separate data boundaries into an amount of sections called *clusters* without concerning labelled information for the learning process, in which every single data point can be apportioned or allocated a degree of relationship to each of the clusters [40], [41]. In another expression, given a sort of data that does not have any label associated with it $\{x^1, x^2, x^3 \ldots x^n\}$, whereas unsupervised algorithm (i.e. *k*-means) applied to find the structure or pattern of these data and gather it into coherent subset. Moreover, the procedure of casting clusters includes merging numerous alterable (variable) into a dissimilar or a distance dimension whose measure are thenceforth expended to construct clusters. An iterative clustering method such as k-means is an admired algorithm worthy for its easiness, immediate convergence and short period intricacy (complexity). These methods intended to the clusters sort of i input data points it $\{x^1, x^2, x^3 \ldots x^n\}$into K coherent (disjoint) subsets of $S_{(i)}$ to minimize mean-square-error, JMSE as in Equation (3):

$$JMSE = \sum_{i=1}^{K} \sum_{x_{(i)} \in S_{(i)}} dist(x_{(i)} - \mu_{r(i)})^2 \quad (1)$$

TABLE I. Banking System Dataset 2016 Training and Testing Behaviour Distribution

Variant	Non-Malicious Behaviours	Malicious Behaviours	Volume
Training	16,797	60,729	77526
Testing	33,579	22,842	56421

where $x_{(i)}$ is a vector representing the i-th input data point and $\mu_{r(i)}$ is the geometric centroid of the data points $x_{(i)}$ in $S_{(i)}$, and $(x_{(i)} - \mu_{r(i)})^2$ is selected as matrix of distances to be minimized between input data points $x_{(i)}$ and $\mu_{r(i)}$ cluster centroid [40]. In addition, Euclidean Distance (Dist) function applied to calculate the distance in a basis of similarity or dissimilarity between $x_{(i)}$ and $\mu_{r(i)}$.

K-means has been chosen and utilized in the proposed model among various conceivable clustering algorithms on the grounds that it have the capabilities to collectively cluster the suspicious and non-suspicious data points without prior knowledge. Secondly, the clustered collection can then be beneficial to label the data points into malicious and non-malicious for the classification stage as presented in the later sections. Based on preliminary studies, MacQueen *et al.* [42], is the researcher whom developed and introduce k-means algorithm in 1967. K-means algorithm can be described as in Algorithm 1.

K-means is one of the many clustering algorithms extensively used for the reason of its in complexity, proficiency and easy to implement. Once the clustered set has been identified and data points have been labelled, it is feasible to carry on the classification procedure as in the next section.

B. Decision Tree Classifier

The Decision Tree (DT) classifier is the well-known data mining method enforced these days and was firstly introduced by Quinlan (1986) [43]. Using similar methodology in [44], a tree classifier known as decision tree has been created. This developed classifier consisting three major elements i.e. decision node which signifies the conditions on an instances, a split that matches close to one of the possibility attribute values and a leaf that chooses the class whereabouts the instances fits in. In order to classify the instances, the initial

Algorithm 1: K-Means Clustering Steps

1. Define the total number of K clusters centroids. Evaluation phase involving a total number of K clusters in a variant of $2<K<7$.
2. Initialize the K clusters centroids, $\mu^1, \mu^2, \mu^3 \ldots \mu^K \in \mathbb{R}^n$. For instance, $\mu^1$and μ^2 for two different clusters (i.e. suspicious and non-suspicious).
3. Compute Euclidean Distance $dist(x_{(i)} - \mu_{r(i)})^2$. It is usually applied to calculate the distance between data points and cluster centroids.
4. Assigning data points to the nearest centroid, such that every single cluster will be occupied by possible similar data points.
5. Re-compute the mean of each cluster centroid. The location of the cluster centroids changes hereon.
6. Repeat steps (3)-(5) until the convergence conditions are fulfilled and the data point is label appropriately according to its clusters.

point defined from top of the leaf which referred as the root and subsequently the branches is established based on the outcome of every single test down along a leaf node is reached. The last part of the leaf node is considered as the classification criteria. The information gain methodology has been applied to select the optimum attribute splitting for each subset on each stage, in which the attributes that has maximum information value choose to form a decision. The formulated algorithm of above-mentioned decision tree illustrated in Algorithm 2.

In the proposed detection approach, the *k*-means clustering helps in labelling and arranging the suspicious and non-suspicious data in some form that could be usable for a classification method such as early notification on possible malicious and non-malicious data to obtain better accuracies and detection outcomes on the subsequent phase using the decision tree classifier. The next section presents the assessment to validate that the proposed method is considerably better in anomaly detection.

IV. Evaluation and Discussion

In this section, the proposed KMDT evaluated with latest datasets. In the following section, the dataset used as well as measurement of detection applied for evaluation purpose have been described and then the assessment result presented.

Algorithm 2: Decision Tree Steps

1. While every single leaf node comprises more than one instances category, remaining attributes and noteworthy information gain do the below steps,
2. Let choose the root node and single attributes,
3. Let Segregate the node population and compute an information gain values,
4. Let discover the split which have highest information gain values for an attribute,
5. Let re-compute these process for entire attributes,
6. Let discover the finest splitting attribute and split rule,
7. Utilize these attributes to split the node,
8. Repeat step 2 to 7 until no child node is remains.

A. Banking System Dataset 2016

Recently, there are huge recommendations among the research community to evaluate the detection method using an appropriate and latest dataset. As such, an analysed and validated packet header which has captured and correlated from the real-time banking system environment applied to perform intrusion detection using proposed KMDT. The detection procedure is supervised on an offline mode and the original class label i.e. malicious or non-malicious ignored for use during the prediction phase. However, those labels have been used to calculate and validate the value of false positive, false negative, true positive and true negative. Table I illustrate the distribution of training and testing behaviours. A variant of experiments were run separately to validate KMDT using aforesaid dataset.

B. Detection Measurement

The detection performance was evaluated using benchmark measurement as recommended by research communities [22], [45], in which the measurement or indicator includes detection rate, accuracy and false alarms. The formulas used to calculate those values are as follows:

$$Accuracy = (tp + tn)/(tp + tn + fp + fn) \quad (2)$$

$$Detection\ Rate = (tp)/(tp + fp) \quad (3)$$

$$False\ Alarm = (fp)/(fp + tn) \quad (4)$$

Non-malicious behaviours which is incorrectly identified as malicious is called *false positive (fp)*, while *false negative (fn)* refers to the incorrect identification of malicious behaviours as non-malicious. Conversely, malicious behaviours which is correctly identified as malicious is called *true positive (tp)*, while *true negative (tn)* is the correct identification of non-malicious behaviours as non-malicious. For better understanding, the entire above-mentioned performance metric described in percentage form in this article.

Under the circumstances of failure to achieve the best degrees in previous metrics, the current detection methods cannot contribute in high accurate detection for accuracy, detection rate and false alarm indicators. Moreover, the major significant task that necessary to be met is to come with applicable scheme that can increase the detection percentage of non-malicious behaviours (NMB) and detection percentage of malicious behaviours (MB) with the development of anomaly based detection. In addition, once the detection percentage of non-malicious and malicious behaviours is maximized, the conventional limitation in having high accuracy and detection rate as well as lowest false alarm is now achievable. The formulas used to calculate those values are as follows:

$$NMB = \left(\frac{tn}{tn+fp}\right) * 100 \quad (5)$$

$$MB = \left(\frac{tp}{tp+fn}\right) * 100 \quad (6)$$

The proposed model is promising to maximize the above-mentioned indicators rate as justified in the next section.

C. Detection Result

In order to appraise the proposed method more meticulously, various experiments have been performed using banking system dataset 2016. The banking system dataset is used to validate the effectiveness of the proposed method in terms of prediction of the malicious and non-malicious behaviours. Different stages of analysis have been conducted using clustering to choose the best optimized cluster sets while decision tree for classification purpose that can contribute to higher accuracy, detection rate and lowest false alarm as well as with the maximum non-malicious and malicious behaviours percentage. In other words, the entire behaviours have been clustered into different variants of (*k*-th) clusters and the most accurate cluster (i.e. highest true positive and true negative) arrangement is chosen for the clustering stage. Once the clustered arrangement has been identified and behaviours have labelled (i.e. suspicious or non-suspicious), the value of the behaviours was entered into the Decision Tree (DT) classification task. The DT classifier classified these

behaviours into non-malicious or malicious classes more accurately.

High detection accuracy and lowest false alarm rate represent the best model of anomaly detection. However, high NMB and MB detection percentage also need to be considered in selecting such detection model. The experimental result confirmed that the proposed KMDT (*k*-means and decision tree) is substantially effective and improve the anomaly-based detection capabilities based on the above-mentioned performance factors as compared to other combinational and individual method. The result of a variant of K-Means (*k-th*), Decision Tree (DT) as an individual method and a variant of K-Means and Decision Tree (*k-th*-DT) as a combinational method from each experiment conducted have been presented from Table 2 through Table 5.

TABLE II. DETECTION PERCENTAGE OF NON-MALICIOUS AND MALICIOUS DATA OF *K*-TH K-MEANS USING BANKING SYSTEM TRAINING DATASET

Cluster *k*-th	***k*-2**	***k*-3**	***k*-4**	***k*-5**	***k*-6**	***k*-7**	***k*-8**	***k*-9**	***k*-10**
DP-NMD	75.4	75.4	80.5	75.4	65.1	79.8	75.4	75.2	81.3
DP-MD	100	100	100	100	100	100	100	100	100

TABLE III. DETECTION PERCENTAGE OF *K*-TH K-MEANS USING BANKING SYSTEM TRAINING DATASET

Cluster *k*-th	***k*-2**	***k*-3**	***k*-4**	***k*-5**	***k*-6**	***k*-7**	***k*-8**	***k*-9**	***k*-10**
Accuracy	42.3	42.3	52.5	42.3	30.6	50.9	42.3	42	54.6
Detection Rate	4.6	4.6	5.5	4.5	3.85	5.35	4.6	4.6	5.8
False Alarm	59.4	59.4	48.9	59.3	71.4	50.5	59.3	59.6	46.7

TABLE IV. DETECTION PERCENTAGE OF *K*-8 K-MEANS & DECISION TREE USING BANKING SYSTEM TRAINING DATASET

Method	**DP-NMD**	**DP-MD**	**Accuracy**	**Detection Rate**	**False Alarm**
***k*-8**	75.4	100	42.3	4.6	59.3
DT	9.38	99.31	79.83	79.84	90.61
***k*-8-DT**	99.98	99.44	99.97	99.53	0.01

TABLE V. DETECTION PERCENTAGE OF *K*-8 K-MEANS + DECISION TREE USING BANKING SYSTEM TESTING DATASET

Method	**DP-NMD**	**DP-MD**	**Accuracy**	**Detection Rate**	**False Alarm**
***k*-8**	50.98	99.98	83.25	79.73	49.01
DT	88.39	67.27	79.84	79.77	11.60
***k*-8-DT**	99.81	99.88	99.86	99.90	0.18

TABLE VI. DETECTION PERCENTAGE OF KNOWN & UNKNOWN MALICIOUS BEHAVIOURS OF *K*-8 K-MEANS + DECISION TREE USING BANKING SYSTEM TESTING DATASET

Method	**Unknown Malicious Behaviours**		**Known Malicious Behaviours**	
***k*-8+DT**	M-UMB	D-UMB	M-KMB	D-KMB
	281	27894	58	3805
	0.01%	99.9%	0.15%	99.85%

In Table II, in a series of *k*-th cluster trials, from $2<k<10$, the entire clusters able to clusters the malicious behaviours accurately with 100% detection (MB). In contrast, for the NMB, clustering outcome are not much satisfactory. For example, *k*-10 and *k*-4 recorded approximately less than 82% as NMB while the remaining *k*-th is below than 80%. Failure in achieving high NMB and NB has caused the computed false alarm, accuracy and detection rate for overall *k*-th clusters recorded the poorest result as in Table III. Based on the analysis, the entire clusters are less effective in grouping the behaviours which are similar to each other.

In Table IV, the result of the KMDT that only has performed better during the assessment period is presented. It is noticeable that the KM-DT which employs *k-8* clusters and Decision Tree (i.e. *k*-8+DT) perform better with higher accuracy, higher detection rate and much lower false alarm rate compared to *k*-8 and Decision Tree (DT). Taking *k*-8, DT and *k*-8-DT as an example, the accuracy and detection rate have increased from 42.3%, 79.83% to 99.97%, and 4.6%, 79.84 to 99.53, respectively, and false alarm has decreased from 59.3%, 90.61% to 0.01%. This table also demonstrates the percentages of detection for the NMB and MB which vary between 75.4%, 9.38% to 99.98% and 100%, 99.31% to 99.44%, respectively. Unlike DT and *k*-8-DT which has only 2 final classes for grouping, *k*-8 has a set of clusters which facilitates to group malicious data effectively. Thus, the result of *k*-8 in MB is much better as compare to others. However, single detection method is not capable to improve the entire performance metrics using banking dataset.

In contrast to the above experiments that only use training data in assessing the proposed method, a different experiment have been further conducted using testing data. The reason is to identify the optimized clusters during the training phase (i.e. *k-8*) and apply these clusters during the testing phase evaluation. In addition, it could be helpful in practical solution and real deployment in selecting the number of cluster set (*k*) by only assessing the available data.

Table V represents the dimensions in terms of accuracy, the detection rate and false alarm of the KMDT method of *k*-8-DT that exercised on testing data. Combinational methods *k*-8+DT have outperformed with higher accuracy and detection rate and lowest false alarm at 99.86%, 99.90% and 0.18%, while *k-8* at 83.25%, 79.73% and 49.01% as well as the DT classifier at 79.84%, 79.77% and 11.60%, respectively. In addition, these combinations are more accurate as compared to *k*-8 clusters set and decision tree classifier for grouping and classifying malicious and non-malicious behaviours. For example, the detection percentages of both NMB and MB for the *k*-8+DT combination have achieved 99.81% and 99.88% which are much better than others, i.e. *k*-8 and DT with 50.98, 99.98% and 88.39%, 67.27%, respectively. The MB of *k*-8 is slightly higher than *k*-8+DT because the clustering (*k*-8) has the advantage to manipulate continuous data as compared to a Decision Tree. Thus, the value of true positive of *k*-8 at 37153 is much higher than *k*-8 +DT at 37117. The entire result signifies that a better performance can be attained using the *k-th* value of 8 combined with DT (*k*-8+DT).

D. Discussion and Further Analysis

This finding proves that the combination of *k*-means and Decision Tree (KMDT) classifier with appropriate clusters set i.e. *k*-8+DT observed earlier during the training phase could give a much better result during the testing phase or in real environment. In addition, *k*-8+DT also contribute in increasing both non-malicious and malicious behaviours detection percentage. The main reason is because the clustering is utilized as the initial element for grouping and labelling of similar data into analogous sorts (i.e. suspicious and non-suspicious), and the capability in handling continuous-value contributes in maximizing detection percentage particularly for malicious classes. For example, as in Table V, the MB of *k*-8 at 99.98% is higher than DT at 67.27% and *k*-8+DT at 99.88%. Based on the investigation, *k*-8 detected 37153 malicious behaviours (true positive), thus yields in higher MB while DT and *k*-8+DT merely achieves 15368 and 37117. However, referring to Table V, the most significant performance is shown through *k*-8+DT, found to greatly improve the detection rate and accuracy above 99% and false alarm below 1% as compared to others. The entire results signify that a better performance could be attained using *k*-8+DT. To support this fact, practically, various experiments conducted and each experiment has shown a remarkable performance for *k*-8+DT. In contrast, the individual *k*-8 clustering and DT classifier which are incapable to identify non-malicious and malicious data more precisely also shown.

The detection of malicious behaviours is most concern in developing a detection system. Failures in identifying more malicious behaviours can cause data and confidential assets compromised by another party. These facts include the identification of known and unknown malicious behaviours. Taking these facts into account, four different performance variants have been considered such as missing known malicious behaviours (M-KMB), missed unknown malicious behaviours (M-UMB), detected known malicious behaviours (D-KMB) and detected unknown malicious behaviours (D-UMB). The M-KMB refers to the percentage where the known malicious behaviours during the training phase are failed to be identified during the testing phase while M-UMB are the percentage where the unknown malicious behaviours which not been covered during the training phase also failed to be detectable during the testing phase. On the other hand, D-KMB refers to refers to the percentage where the known malicious behaviours during the training phase are correctly identified during the testing phase while D-UMP is the percentage where the unknown malicious behaviours which not been covered during the training phase also correctly detectable during the testing phase.

Further experiments and analysis has been conducted to evaluate the proposed *k*-8+DT using aforesaid variants against banking dataset. Total number of unique malicious behaviours used is 32,038 in which 28,175 are known malicious behaviours while remaining 3836 are unknown malicious behaviours. Table VI exhibits the distribution of known and unknown malicious behaviours, and the outcome of *k*-8+DT. Surprisingly, based on the result, the rate of D-UMB and D-KMB for *k*-8+DT is 99.9% and 99.85%, which means *k*-8+DT able to detect more unforeseen malicious behaviours as a contrast to seen behaviours accurately. Moreover, *k*-8+DT also recorded the lowest rate of M-UMB and M-KMB at 0.01% and 0.15%. Therefore, *k*-8+DT are more suitable to be utilized as anomaly detection systems particularly for detecting unknown malicious behaviours.

V. Conclusion and Future Works

Packet header for intrusion detection has attracted researchers' attention in the field of data mining based anomaly detection. Although a number of anomaly detection methods have been proposed, the common drawback is to achieve a high rate of accuracy, detection rate as well as a lower false alarm with high non-malicious and malicious behaviours detection remain as an unsolved problem, and directly affects the integrity of the said detection method to be widely adopted. In this work, a combined anomaly detection model named KMDT based on data mining methods that focuses on examining the entire features of a packet header to detect malicious behaviours is proposed. The *k*-means clustering is utilized to label the entire behaviour based on the behaviour characteristic such as suspicious or non-suspicious. In subsequent stages, the clustered behaviours input into Decision Tree classifier for classification purpose. The evaluation phase using banking dataset 2016 validates that the combinational method between *k*-means and Decision Tree shows an effective performance, such as higher accuracy and detection rate with lower false alarm as contrast to others. Thus, the KMDT approach could be a better anomaly detection method in identifying abnormal behaviour and determining it to be malicious or non-malicious behaviour more correctly. Besides, the ability of the KMDT focuses on identifying cyber-attack without emphasis on processing time or prompt detection can be considered for future research. For example, efforts to reduce the number of features that need to be examined are necessary and could be performing through feature reduction approach. This directly improves the processing time and the malicious data can be observed quickly.

Acknowledgment

The authors would like to thank the anonymous reviewers for their significant comments. Furthermore, the research project has been conducted in Universiti Teknikal Malaysia Melaka using Grant Scheme FRGS/1/2014/ICT4/FTMK?02/F00212.

References

[1] M. A. Faysel and S. S. Haque, “Towards Cyber Defense : Research in Intrusion Detection and Intrusion Prevention Systems,” vol. 10, no. 7, pp. 316–325, 2010.

[2] A. Patcha and J.-M. Park, “An overview of anomaly detection techniques: Existing solutions and latest technological trends,” Comput. Networks, vol. 51, no. 12, pp. 3448–3470, Aug. 2007.

[3] C.-M. Chen, Y.-L. Chen, and H.-C. Lin, “An efficient network intrusion detection,” Comput. Commun., vol. 33, no. 4, pp. 477–484, 2010.

[4] Y. Waizumi, Y. Sato, and Y. Nemoto, “A Network-Based Anomaly Detection System Based on Three Different Network Traffic Characteristics,” J. Commun. Comput., vol. 9, no. 7, p. 805, 2012.

[5] H.-J. Liao, C.-H. Richard Lin, Y.-C. Lin, and K.-Y. Tung, “Intrusion detection system: A comprehensive review,” J. Netw. Comput. Appl., vol. 36, no. 1, pp. 16–24, Jan. 2013.

[6] P. Louvieris, N. Clewley, and X. Liu, "Effects-based feature identification for network intrusion detection," Neurocomputing, vol. 121, no. 0, pp. 265–273, 2013.

[7] S. Juma, Z. Muda, and W. Yassin, "Machine Learning Techniques For Intrusion Detection System: A Review.," J. Theor. Appl. Inf. Technol., vol. 72, no. 3, 2015.

[8] M. Woźniak, M. Graña, and E. Corchado, "A survey of multiple classifier systems as hybrid systems," Inf. Fusion, vol. 16, no. 0, pp. 3–17, 2014.

[9] Y. Cho, K. Kang, I. Kim, and K. Jeong, "Baseline Traffic Modeling for Anomalous Traffic Detection on Network Transit Points," in Proceeding APNOMS'09 Proceedings of the 12th Asia-Pacific network operations and management conference on Management enabling the future internet for changing business and new computing services, 2009, pp. 385–394.

[10] Y. Xie, S. Tang, X. Huang, C. Tang, and X. Liu, "Detecting Latent Attack Behavior from Aggregated Web Traffic," Comput. Commun., vol. 36, no. 8, pp. 895–907, 2013.

[11] K.-C. Lee, J. Chang, and M.-S. Chen, "PAID: Packet Analysis for Anomaly Intrusion Detection," in Advances in Knowledge Discovery and Data Mining SE - 58, vol. 5012, T. Washio, E. Suzuki, K. Ting, and A. Inokuchi, Eds. Springer Berlin Heidelberg, 2008, pp. 626–633.

[12] M. Z. Mas'ud, S. Sahib, M. F. Abdollah, S. R. Selamat, and R. Yusof, "Analysis of Features Selection and Machine Learning Classifier in Android Malware Detection," in 2014 International Conference on Information Science & Applications (ICISA), 2014, pp. 1–5.

[13] L. Koc, T. A. Mazzuchi, and S. Sarkani, "A network intrusion detection system based on a Hidden Naïve Bayes multiclass classifier," Expert Syst. Appl., vol. 39, no. 18, pp. 13492–13500, 2012.

[14] S. Juma, Z. Muda, and W. Yassin, "Reducing False Alarm Using Hybrid Intrusion Detection Based On X-Means Clustering and Random Forest Classification," J. Theor. Appl. Inf. Technol., vol. 68, no. 2, pp. 249–254, 2014.

[15] W. Yassin, N. Udzir, A. Abdullah, M. Abdullah, Z. Muda, and H. Zulzalil, "Packet Header Anomaly Detection Using Statistical Analysis," in International Joint Conference SOCO'14-CISIS'14-ICEUTE'14 SE - 47, vol. 299, J. G. de la Puerta, I. G. Ferreira, P. G. Bringas, F. Klett, A. Abraham, A. C. P. L. F. de Carvalho, Á. Herrero, B. Baruque, H. Quintián, and E. Corchado, Eds. Springer International Publishing, 2014, pp. 473–482.

[16] M. Z. Mas`ud, S. Sahib, .., M. F. Abdollah, S. R. Selamat, and R. Yusof, "Android Malware Detection System Classification," Res. J. Inf. Technol., vol. 6, no. 4, pp. 325–341, Apr. 2014.

[17] A. Urtubia, J. R. Pérez-Correa, A. Soto, and P. Pszczólkowski, "Using data mining techniques to predict industrial wine problem fermentations," Food Control, vol. 18, no. 12, pp. 1512–1517, 2007.

[18] K. Abhaya, R. Jha, and S. Afroz, "Data Mining Techniques for Intrusion Detection: A Review," Int. J. Adv. Res. Comput. Commun. Eng., vol. 3, no. 6, pp. 6938–6942, 2014.

[19] S. Agrawal and J. Agrawal, "Survey on Anomaly Detection using Data Mining Techniques," Procedia Comput. Sci., vol. 60, pp. 708–713, 2015.

[20] J. F. Hair, R. Tatham, R. E. Anderson, C. T. Reviews, and B. Black, Multivariate Data Analysis, 6th ed. Academic Internet Publ., 2006.

[21] A. B. S. Serapião, G. S. Corrêa, F. B. Gonçalves, and V. O. Carvalho, "Combining K-Means and K-Harmonic with Fish School Search Algorithm for data clustering task on graphics processing units," Appl. Soft Comput., vol. 41, pp. 290–304, Apr. 2016.

[22] D. M. Farid, L. Zhang, C. M. Rahman, M. A. Hossain, and R. Strachan, "Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks," Expert Syst. Appl., vol. 41, no. 4, Part 2, pp. 1937–1946, 2014.

[23] Y. Guan, A. A. Ghorbani, and N. Belacel, "Y-means: a clustering method for intrusion detection," in Electrical and Computer Engineering, 2003. IEEE CCECE 2003. Canadian Conference on, 2003, vol. 2, pp. 1083–1086 vol.2.

[24] M. Kaushik and B. Mathur, "Comparative Study of Various Clustering Techniques," Int. J. Softw. Hardw. Res. Eng., vol. 3, no. 10, pp. 497–504, 2014.

[25] M. Ahmed, A. Naser Mahmood, and J. Hu, "A survey of network anomaly detection techniques," J. Netw. Comput. Appl., vol. 60, pp. 19–31, Jan. 2016.

[26] Y. Zhang, W. Lee, and Y.-A. Huang, "Intrusion Detection Techniques for Mobile Wireless Networks," Wirel. Netw., vol. 9, no. 5, pp. 545–556, Sep. 2003.

[27] S. Duque and M. N. bin Omar, "Using Data Mining Algorithms for Developing a Model for Intrusion Detection System (IDS)," Procedia Comput. Sci., vol. 61, pp. 46–51, 2015.

[28] P. García-Teodoro, J. Díaz-Verdejo, G. Maciá-Fernández, and E. Vázquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges," Comput. Secur., vol. 28, no. 1–2, pp. 18–28, 2009.

[29] P. Kapoor and R. Rani, "Efficient Decision Tree Algorithm Using J48 and Reduced Error Pruning," Int. J. Eng. Res. Gen. Sci., vol. 3, no. 3, pp. 1613–1621, 2015.

[30] S. Zhong, T. M. Khoshgoftaar, and N. Seliya, "Clustering-Based Network Intrusion Detection," Int. J. Reliab. Qual. Saf. Eng., vol. 14, no. 02, pp. 169–187, Apr. 2007.

[31] C.-F. Tsai and C.-Y. Lin, "A triangle area based nearest neighbors approach to intrusion detection," Pattern Recognit., vol. 43, no. 1, pp. 222–229, 2010.

[32] M. Mohammadi, Z. Muda, W. Yassin, and N. Izura Udzi, "KM-NEU: An Efficient Hybrid Approach for Intrusion Detection System," Res. J. Inf. Technol., vol. 6, no. 1, pp. 46–57, Jan. 2014.

[33] A. P. Muniyandi, R. Rajeswari, and R. Rajaram, "Network Anomaly Detection by Cascading K-Means Clustering and C4.5 Decision Tree algorithm," Procedia Eng., vol. 30, no. 0, pp. 174–182, 2012.

[34] T. R. Patil, "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification," Int. J. Comput. Sci. Appl. ISSN 0974-1011, vol. 6, no. 2, pp. 256–261, 2013.

[35] Y. K. Jain, "An Efficient Intrusion Detection Based on Decision Tree Classifier Using Feature Reduction," Int. J. Sci. Res. Publ., vol. 2, no. 1, pp. 1–6, 2012.

[36] D. P. Gaikwad and R. C. Thool, "Intrusion Detection System Using Bagging with Partial Decision TreeBase Classifier," Procedia Comput. Sci., vol. 49, pp. 92–98, 2015.

[37] A. Rai, K., Devi, M.S. and Guleria, "Decision Tree Based Algorithm for Intrusion Detection," Int. J. Adv. Netw. Appl., vol. 2834, pp. 2828–2834, 2016.

[38] C. A. Catania and C. G. Garino, "Automatic network intrusion detection: Current techniques and open issues," Comput. Electr. Eng., vol. 38, no. 5, pp. 1062–1072, 2012.

[39] M. Panda, A. Abraham, and M. R. Patra, "A Hybrid Intelligent Approach for Network Intrusion Detection," Procedia Eng., vol. 30, pp. 1–9, Jan. 2012.

[40] A. K. Jain, "Data Clustering: 50 Years Beyond K-means," Pattern Recogn. Lett., vol. 31, no. 8, pp. 651–666, 2010.

[41] N. Jain and V. Srivastava, "Data Mining Techniques: A Survey Paper," IJRET Int. J. Res. …, vol. 2, no. 11, pp. 116–119, 2013.

[42] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967, pp. 281–297.

[43] J. R. Quinlan, "Induction of decision trees," Mach. Learn., vol. 1, no. 1, pp. 81–106, 1986.

[44] D. AL-Nabi and S. Ahmed, "Survey on Classification Algorithms for Data Mining:(Comparison and Evaluation)," Comput. Eng. Intell. Syst., vol. 4, no. 8, pp. 18–25, 2013.

[45] W. Xiong, N. Xiong, L. T. Yang, J. H. Park, H. Hu, and Q. Wang, "An Anomaly-based Detection in Ubiquitous Network Using the Equilibrium State of the Catastrophe Theory," J. Supercomput., vol. 64, no. 2, pp. 274–294, 2013.