ENHANCE INTRUSION DETECTION CAPABILITIES VIA WEIGHTED CHI-SQUARE, DISCRETIZATION AND SVM

*WARUSIA YASSIN, [1]MOHD FAIZAL ABDOLLAH, [1]MOHD ZAKI MAS'UD, [1]ROBIAH YUSOF, [1]RAIHANA ABDULLAH, [2]ZAITON MUDA

*[1]Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Malaysia

[2]Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Malaysia

E-mail: *s.m.warusia@utem.edu.my

ABSTRACT

Anomaly Intrusion Detection Systems (ADSs) identify patterns of network data behaviour to determine whether they are normal or represent an attack using the learning detection model. Much research has been conducted on enhancing ADSs particularly in the area of data mining that focuses on intrusive behaviour detection. Unfortunately, the current detection models such as the support vector machine (SVM) is affected by high dimensional data which limits its ability to accurately classify data. Moreover, the data points which appear similar between intrusive and regular behaviours could be problematic as some innovated attack behaviours may not be detected. To overcome this SVM drawback, we propose a combination of weighted chi-square (WCS) as a feature selection (FS) and a Discretization process (D). The WCS method is used firstly to reduce the dimensionality of data following which the assembled records are transformed into interval values via the D process before the SVM is used to identify groups of samples that behave similarly and dissimilarly such as malicious and non-malicious activities. Experiments were performed with well-known NSL-KDD data sets and the results show that the proposed method namely WCS-D-SVM (weighted chi-square, discretization and support vector machine) significantly improved and enhanced accuracy and detection rates while decreasing the false positives which the single SVM classifier produces.

Keywords: *Intrusion Detection, Data Mining, Feature Selection, Weighted Chi-square, Discretization, Support Vector Machine.*

1. INTRODUCTION

Intrusion Detection Systems (IDSs) monitor and detect malicious activity on networks or computer resources. In recent years, their management and development for network and computer security has become a major concern [1] due to their basic function of recognizing and notifying administrators of the different types of attacks in the network environment. Further, IDSs are also able to monitor any attempts at malicious activity and unauthorized entry [2]. There are two widely applied approaches to identify intrusion activity, namely the Signature Based Detection Systems (SBDSs) and the Anomaly Based Detection Systems (ABDSs) [3].

Comparing the collected and analyzed information against attack signatures in large databases is known as an SBDS [4]. As SBDSs comprise regular known attacks, efforts are always needed to create novel attack signatures. In addition, SBDSs usually have a common function with antivirus needed to investigate a specific attack or a documented malwares activity. On the other hand, in the ABDSs, the administrator usually defines a baseline or normal state of network traffic protocol and typical packet size. The incoming network segments will be compared with this baseline and seek out any deviations, known as anomalies (Juma, et al., 2014) and the identified anomalies are always considered as attacks. A false detection on normal activity or behaviour becomes a major challenge of these detection systems.

Recently, the Data Mining (DM) approach has received much attention from the research community as it could be applied for Anomaly Detection (AD) purposes [5]. In DM, anomalies are shapes of data behaviour which do not fit perfectly and are usually referred to as an anomaly or intrusion activity [6]. Although the Naive Bayes (NB) approach has been utilized as an AD, there are several constraints faced by this classifier which researchers are actively seeking to address. The major theoretical limitation of this classifier is related to a poorly created detection model that makes it difficult for intrusive behaviour to be detected as well as other unknown behaviour difficulties yet to be determined [4]. Technically, however, this method has limitations over high dimension data which consist of a number of irregular features apart from constraints in distinguishing data points more accurately. This situation is exacerbated where the detection model focuses on detecting the unknown behaviour of innovative attacks. This failure directly creates a high possibility of an inrease in false detections, i.e. false positives and false negatives.

This study discover the solution to overcome the limitation in differentiate the similar behaviour of an anomalous and non-anomalous behaviour that usually effect the detection capabilities of classifiers. In addition, this study will be beneficial to researcher in how to distinguish the similar data point of an innovated attacks against legitimate data point particularly for high dimensional data which hard to classify nowadays. For this purpose, the weighted chi-square as feature selection, discretization process and support vector machine as classifier are considered. The proposed approach is known as the WCS-D-SVM (weighted chi-square, discretization and support vector machine).

2. RELATED WORK

Various researchers have reviewed the Data Mining (DM) methods with a view to Anomaly Detection (AD) for cyber intrusion detection. They include, [7][8][9] [10] [11] [12] [13] [14]. An analysis of earlier unknown behaviour inside the data is the main focus of DM [15]. For example, in the case of AD, the observed packet behaviour is interpreted in the training phase, defined as the known behaviour (a collection of this known behaviour is also identified as learned model), while in the testing phase this learned model is applied against every single new packet to determine whether its behaviour is non-anomalous or anomalous. The DM based anomaly detection can be divided into three broader categories, i.e., unsupervised, semi-supervised and supervised [16] [17] [18]. This section focuses on semi-supervised and supervised learning, specifically feature selection and classification.

2.1 Feature Selection

The DM procedure can be segregated into several major phases; where, in most cases, in the first phase, the semi-supervised method has been applied as a preprocessing task before the supervised algorithm by groups of previous researchers [19][20] specifically to reduce the dimensionality of the data sets [16][21][22]. Preprocessing directly influences the subsequent phase produced learning models, particularly when its outcome turns as an input for the following phases such as for classification [22]. Recently, the study of feature selection (FS) as an example of preprocessing task has received much attention and plays a crucial function within the DM procedure in some of the literature. In addition, even though the preprocessing task always uses up more processes, appropriate and accurate execution of the FS function allows for good quality data to be generated. For instance, unnecessary features could affect the prediction performance of the computed models, while the FS facilitates the removal of those redundant and irrelevant features for better understandability [22] as well as directly improving predictions and reducing training time [18][23]. Feature selection methods can be divided into two major groups: wrapper and filter [18]. A subdivision of features that are selected independently in the pre-processing step of selecting a classification approach is known as the filter method while wrapper is used more to assess those subdivisions of features based on a fixed predictive function. Each of the above groups can be further divided into categories based on several criteria or procedures.

However, there is significant variation of feature selection algorithms which can be applied depending on certain criteria. This extensive accessibility causes difficulty in selecting the most adequate among a number different characteristic of feature selection algorithms [16][24]. Priority in the selection of a specific FS algorithm must take into account its ability to resolve specific related issues. In addition, the selected FS algorithm allows for not only elimination of redundant and irrelevant features, but addresses the imperfections of high-dimensional data, missing values and inconsistencies such as those presented by huge raw network traffic [25][26]. Moreover, the accuracy of the posterior learning methods is questionable if such low quality data is presented. There is a need to conduct appropriate preprocessing steps that address the quality of following analyses and assessments [17]. As such, number of researchers have proposed and applied the preprocessing method such as feature selection that has become a major concern in the area of data mining specifically in facilitating anomaly detections [27] For example, they have focused in either reducing or selecting the most appropriate features before performing the intrusion detection. [28] and [29] conducted a survey on the number of feature selection of algorithms including genetic-based ones, rough sets and particle swam intelligence which have been applied by various researchers in the field of intrusion detection for tackling performance issues. They claimed that it is essential to conduct an analysis throughout the significant features as a solution to overcome performance degradation in terms of time consume. Moreover, a novel algorithm called the rule-based attribute selection has also been proposed for effective feature selection. As such, an intelligent rule-based attribute selection using information generation as well as tuple selection is considered. The detection accuracy of a proposed method, namely the Rule-based Enhanced Multiclass SVM (IREMSVM) against a specific attack i.e., probe and DoS, improved by more than 99% compared to a single SVM's less than 92% after the experiment was conducted with a series of selected features. [30] and [31] clarified that implementing the feature selection as a component inside the IDSs framework may raise the rata of intrusion activity detection. Thus, this author has developed a novel approach for automated labelling of incoming traffic and to conduct feature selection tasks using Genetic Algorithm (GA). The GA is applied to determine the best solution for optimization limitation using a number of various parameters such as preliminary population size, chromosome length, number of repetitions, crossovers and mutation probability. Different experiments with different sets of selected metrics have generated different detection results. On the other hand, an Optimal Feature Selection (OFS) algorithm based on Information Gain Ratio (IGR) is proposed in [32] to reduce the time required for classification execution. The IGR is used to select the optimal volume of features on the dataset; so that data extracted from the traffic are not complex and facilitates in promptly detecting intrusion activity. In addition, OFS increased the detection accuracy of DoS, Probe, U2R and R2L by up to 99%, 95%, 95% and 95% respectively as well as minimized false alarm rates. As stated by [33], in the field of intrusion detection the algorithms usually applied as feature selection are Information Gain (IG) and Chi-Square (CS). They also stated that the feature selection algorithm is not only helpful in reducing time and improving accuracy, but also has the capability to reduce the false positive rate specifically on NIDS based systems. Moreover, issues in the selection of an optimal feature subset within a feature selection algorithm can also can affect detection performance. Thus, they propose resolving issues related to the process of selecting the best feature subset through a correlation and redundancy algorithm. The mutual information algorithm is employed to represent the largest minimum redundancy of a nonlinear relationship and to determine the relationship of the chosen feature. The final selection of the best feature subsets is determined based on the computed value of redundancy among them and representations of correlation among the labels. This approach has reduced false positives and detection times as well as improved accuracy rate. On the other hand, [34] and [35] introduced an Ant Colony Optimization (ACO) approach in selecting optimized features for enhanced intrusion detection performance. Since the feature sets are simplified through this approach, computational difficulty become minimal. First, the ACO algorithm is employed to establish the space of entire feature subsets of the available feature set. These selected feature subsets are measured using an evaluation function and the best is the subset considered as the finest set which could be applied in IDS. Using these algorithms, the false positive and true positive rates improved. [34] and [36] highlight that the pre-processing stage is necessary to significantly enhance the entire detection performance while mining big datasets. Thus, they propose a feature selection mechanism throughout the random forest in two major steps: choosing the best feature with a greater variable significant score and conducting or managing the preliminary of the search process. These subsequent steps produce the last set of features for the classification method. Remarkably, the accuracy for the R2L based attack and the detection time of the selected feature sets are a lot better than those for the full sets of features.

All the proposed approaches have been tested using nominal attributes since the feature selection method assumes the features or variables as discrete and only deal with numerical values [17]. Moreover, there are distinct advantages and disadvantages in both the filter and wrapper FS based methods [18]. However, the implementation method selected, especially in the field of intrusion detection usually depends on the speed and accuracy of the related algorithm's performance against a large number of features in the network traffic [35]. Moreover, effectively identifying the most significant features which contribute to or aid in intrusion detection has also to be considered.

2.2 Classification

In contrast to the feature selection approach, the classification of a supervised learning approach in anomaly detection is also an active research area nowadays [4][37]. These learning methods usually seek to create a detection model that is able to describe the arrangement of data points corresponding to the elements that are connected with a series of class labels [7]. Subsequently, the learned model which contains the structured data will be applied in predicting data class in testing sets using a pre-determined label from the training set [5].

A number of surveys of this work have been conducted and using the similar concept explained above, various works on the subject have been proposed. For example, [38], reviewed four well-known diverse classification methods used to perform intrusion detection, that is, the Support Vector Machine (SVM), Decision Tree (DT), Naïve Bayes (NB) and Neural Network (NN). According to the author each of these detection methods have their own strengths, i.e., the SVM is much faster and more accurate in performing intrusion detection than the NN while the NB classifier's attack detection accuracy slightly is slightly less than DT, and so forth. The author concludes that a single classifier by itself is unable to overcome its limitations, thus requiring an additional algorithm before or after the classification for better performance. For instance, the author addressed that [39][40] utilized the Rough Set Theory (RST) and Principle Component analysis (PCA) before classification for reducing the irrelevant features and dimensionality of the data resulting in improved detection performance. Similarly, [41] surveyed classification methods specifically used for improving the performance of Intrusion Detection Systems (IDS). They reviewed a number of earlier works and claim that these techniques still required improvement as unknown attacks are difficult to detect. Moreover, [42] explored a series of previous flow-based intrusion detection systems frameworks and assessment outcomes, particularly those that employed the machine learning approach. The SVM, DT and NN are among the reviewed methods as the learning methods have received much attention in the IDS field recently as claimed in [43][44]. The authors mention that network traffic is one of network flow data being analysed widely for intrusion detection. These flows are defined as a sorting of packets which underwent the observation process with common properties in a particular period of time on the network. In addition, they highlight the common drawbacks of such methods such as being computationally costly, have false positive rates, and being analysed with few flow features which could contribute to missing out useful information. Thus, the exploration of flow attributes is mandatory as it is related to the performance of an IDS.

Apart from conducting a survey, [45] designed and evaluated several machine learning algorithm performances, such as Logistic Regression (LR), Gaussian Naïve Bayes (GNB), SVM and Random Forest (RF) focusing on intrusion detections. According to them, the performance quality of the classifier models is more appropriate for evaluation with the involvement of two major measurement factors i.e., the false positive rate and true positive rate. Further, from the experiment conducted, the RF outperformed and showed better results compared to others. However, as the RF algorithm needs to generate more DT and select the best RF as the detection models, detection time could increase if huge volumes of data need to be analysed. On the other hand, [37] analysed several classification methods including NN, NB, DT, RF and SVM which were utilized as an IDS to facilitate the researcher in recognizing the pros and cons of the chosen methods. The authors note that although various methods can be employed to identify cyber-attacks, the process of raising the percentage of attack identification is quite challenging because it depends on the trained model and dataset. Further, the intrusion detection framework was proposed in [46] where they included two neural networks (NN) that apply the soft max PEs and feature selection approach to classify the packet's behaviors of network probe attacks such as host sweep and port scan. Each NN consists of different layers of functionality and pattern recognition entities specific to each of the above-mentioned attacks. In addition, the total number of attacks identified through the system are equalized with output layers and processed items. As opposed to the general approach of NN, the training data has been considered from the output of the recognition stage. The proposed detection system frameworks were created, tested and compared with the output of SNORT. The 24 attacks, the proposed framework consisting of the NN could detect all of them compared to only 15 by SNORT. As stated in [47], in reality the network data usually comprises a huge number of noisy data which IDS needs to evaluate for any intrusive activity. This is not an easy to accomplish particularly using the data mining (DM) approach that might end with unfavourable results. Moreover, the authors conducted a series of assessments using noisy data against various prominent DM methods aimed at observing the consequences of such noisy data against the DM's detection capability. As the noisy data increased in the evaluation stage, methods were unable to classify it more accurately including DT, NN, SVM, and RF. [48] highlight that SVM is able to project real assessed feature vectors into higher structural feature spaces via non-linear projections, which facilitate the capability to achieve real-time detection of, difficult to detect intrusive activity inside an incremental dataset, overcome high dimensional data, as well as multiple class classification issues. Thus, the researcher proposed an optimum allocation-based least square support vector machine (OA-LS-SVM) specifically for handling intrusion detection mechanisms using incremental datasets. Under this approach, the dataset is initially isolated into several pre-arranged arbitrary sub-divisions before samples that are able to reflect the whole dataset from these divisions are selected by the proposed algorithm. Then, based on the diversity of inspections on these related sub-division, an optimum allocation pattern is derived. Finally, the least square SVM is employed against the derived samples for intrusion detection purposes. Based on the analysis conducted, the proposed algorithm managed to work in both static and incremental datasets in identifying intrusive activity.

Unfortunately, based on the above literature study, the outcomes from classifiers did not necessarily perform well in the malware behaviour detection with the major drawbacks being sustaining with maximum accuracy, detection and minimizing false alarm rates. The impact of such classifiers could be enhanced if the pre-liminary phases such as applying the feature selection method is considered. However, not every combinational method could produce more accurate results and it is extremely difficult or challenging to identify the ideal pair of algorithms as the researcher must go through a scientific review process of selected methods. In this paper, a weighted chi-square was proposed as to perform the feature selection task and the SVM for classification purposes.

3. PROPOSED METHODOLOGY

The learning approaches may produce high detection rates for seen attacks behaviour but they might also be unable to predict unseen attack behaviors more accurately. In addition, not every data that has similar patterns can be considered as belonging to behaviour that is normal or anomalous, particularly data that behaves in the same manner but makes the detection algorithm classify it falsely. Furthermore, the detection might be affected by irrelevant features that contribute to falsely classified issues. Thus, we propose a weighted chi-square as the feature selection process in the first stage, the discretization process in the second, and finally a support vector machine as a classifier in the third stage for enhanced prediction of normal and attack behaviours.

Stage 1: Feature Selection Process

The chi-square is based on the fundamental statistical approach where a nonparametric method is employed to resolve whether hypothetical expected frequencies are dissimilar to the distribution of observed frequencies. Specifically, for feature selection these approaches are used to evaluate even if the frequencies of the entity and class are dependent. This method only applies frequency values as its always deal with nominal values rather than the mean and variances. The formula to compute chi-square is as follows:

$$X^2 = Sigma\left[\frac{(o-E)^2}{E}\right] \qquad (1)$$

Chi-square can be referred as X^2, while *O* and *E* represent the value of observed and expected frequencies. Usually, inconsistencies among every single outcome of expected quantity frequency (if the model is true) and observed quantity frequency are encapsulated by a chi-square statistic through a total of the squares of inconsistencies while the expected frequencies are normalized over all the classes. In this work, the weighted chi-square is able to compute the weight value of an entity, namely the feature or attributes with the class. The greater the weight value of an entity, the more relevant it is.

Stage 2: Discretization Process

In recent years, the discretization method has been widely used in the field of data mining and as an anomaly detection method. In this subsequent stage, the discretization procedure is applied to transfigure a continuous-value to intervals which will facilitate the SVM i.e., aid it in overcoming numeric attribute issues and outlier limitations which subsequently enable the classifier to classify anomalous and non-anomalous records more accurately.

Stage 3: Support Vector Machine Classifier

The Support Vector Machine (SVM) is a space based model that has been widely applied recently as a supervised machine learning algorithm for data scrutinizing in the field of IDSs. The algorithm was initially introduced by Vapnik in 1995 [49]. The SVMs has been established as a significant learning methods, particularly for classification analyses as a consequence of its ability effectively achieve high classification results as claimed in [50]. The fundamental SVM equation applied is as follows:

$$D = \{(p_i, q_i) | p_i \in R^N, q_i \in \{1, -1\}\}_{i=1}^{N} \quad (2)$$

Let D be the set of training data, (p_i, q_i) where p_i represent n scrutinizes of malware, pi E R^N , i=1,….,N; and the respective corresponding class label q_i could possess either one value of malicious (1) or non-malicious (-1), expressing the scrutinizes instance in which p_i belongs to which classes, $q_i \in \{1, -1\}$, assigned to every single scrutinizes p_i. Each malware p_i is of dimension d corresponding to the number of organized variables.

In general, the SVM is formulated via judgement boundaries based on the theory of judgement planes. An assortment of entities (also known as vectors or objects) disjoints into distinct groups of classes through the judgement plane. For example, as illustrated in Figure 1, let us assume the entities to have disjointed into two dissimilar groups (where ✦ and ✦ are considered malicious and non-malicious, respectively) through a boundary line.

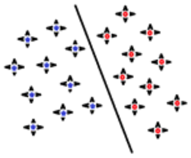

Figure 1: Disjointed Groups Using a Boundary Line

However, the above SVM approach is unable to formulate the hyperplane which may not group the entities into two non-overlapping classes and causes the learned model to miss-classify the related entities. Consequently, the hyperplane that is able to minimize the miss-classify has been considered. Thus, as an initial step a curve approach is employed (Figure 2). As Figure 2, clearly shows the entities in Figure 1 must be separated into two groups based on a curve as this is much more effective in covering up the malicious and non-malicious entities.

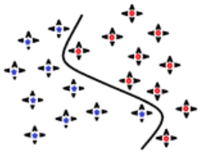

Figure 2: Disjoint Groups Using a Curve Line

Next, as illustrated in Figure 3, the procedure for mapping; containing the process of reorganizing the earliest entities throughout the kernel function is conducted. Using this function, the mapped entities are linearly isolated and the process of optimizing the line achieved.

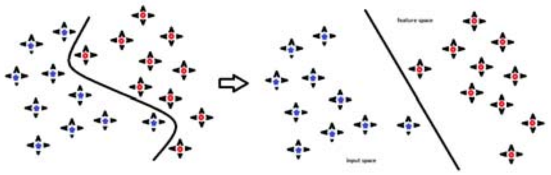

Figure 3: Disjoint Groups Linearly Isolated

In summary, the developed SVM algorithm conducts the classification task faster than identified by the hyperplane which optimizes the margin bounded by two classes. The vectors that formulate the hyperplane are considered as support vectors. The detection of malware could be enhanced once the relevant or significant vectors are selected through the weighted chi-square and feed into the SVM.

4. EXPERIMENTAL OPERATION

The effectiveness of the proposed detection model was evaluated using the NSL-KDD which is an upgraded benchmark dataset derived from the KDD Cup 1999. In general, the KDD Cup 1999 dataset comprises 41 features of traffic header records with one additional labelling feature describing whether the traffic belongs to an attack or is a normal connection. Of the 41 features, 32 are in continuous form while the remaining 9 are in nominal form. Also, the entire dataset can be grouped into three major protocols i.e., UDP, TCP and ICMP. Researchers have recently considered the KDD Cup 1999 dataset as having some limitations as contains various re-produced header records which could affect the trained classifier's and influence them against the more repeatedly record. To address such drawbacks, [51] produced the more applicable, practical and efficient NSL-KDD dataset which is derived from the KDD Cup 1999, but without the re-iterating records and having new structures.

Table 1: Distribution of Records in NSL-KDD Data Set

Behaviour	NSL-KDD			
	Training	%	Testing	%
Non-anomalous	67344	53.46	9712	43.08
Anomalous	58631	46.54	12834	56.92
Total	125975	100	22546	100

Table 2: Different Behaviour of Records

Actual	Predicted as Non-anomalous	Predicted as Anomalous
Non-anomalous	TN	FP
Anomalous	FN	TP

Various researcher's haves applied this dataset in assessing their detection models as it is better suited for intrusion detection research. The distribution and the percentage of the NSL-KDD dataset in terms of classification of records described in Table 1 while Table 2 shows the different behaviours of predicted records. Various factors are currently used by the research community to evaluate the efficiency of IDS and the general measurements are based on calculating the detection rate, accuracy and false alarms such as, *Accuracy = ((TP+TN/TP) / (TN+FP+FN))* and *Detection Rate = ((TP) / (TP+FP))*. In addition to the above measurements, the percentage of correctly classified non-anomalous and anomalous behaviours are also considered in evaluating the proposed detection model.

5. EXPERIMENTAL OUTCOME AND DISCUSSION

This study designed and evaluated a support vector machine (SVM) and the proposed method of the weighted chi-square as feature selection, discretization process and SVM as classifier, namely the WCS-D-SVM against the NSL-KDD dataset. Figure 4 shows the experimental outcome of SVM and WCS-D-SVM in terms of accuracy (AC) and detection rate (DR) while Figure 6 shows the detection percentage of predicted anomalous and non-anomalous behaviour using the training set. As seen in Figure 4, the single SVM only obtained 97.42% and 98.33% as AC and DR, while the proposed WCS-D-SVM produced 99.84% and 99.88% an increase of 2.42% and 1.55% respectively. In addition, WCS-D-SVM predicted non-anomalous and anomalous behaviour more accurately at 99.89% and 99.77% as compared to the SVM's 98.58% and 96.08% as in Figure 5. It is noticed that the SVM missed more anomalous behaviour at 3.92% compare to the WCS-D-SVM at 0.23%. Based on the analysis, there are similar records of anomalous with non-anomalous behaviour which make it difficult for the WCS-D-SVM and SVM to differentiate causing 136 and 2301 records to be identified as non-anomalous (this scenario refers to false positive records). However, the SVM is less accurate in terms of identifying anomalous and non-anomalous records compared to the WCS-D-SVM which is able to significantly minimize the redundant records and directly improves the detection results of the SVM after the feature selection and discretization phases. Furthermore, the arrangement of the values in interval form using the discretization process has facilitated the WCS-D-SVM in obtaining higher classification results.

Figure 6 and Figure 7 illustrate the proportion of AC, DR and the detection percentage of predicted anomalous and non-anomalous behaviour using the testing set. The WCS-D-SVM managed to get higher results for the AC and DR at 98.4% (an increase of 3.8%) and 98.29% (an increase of 3.75%) compared to SVM with only 94.6% and 94.54% as in Figure 6.

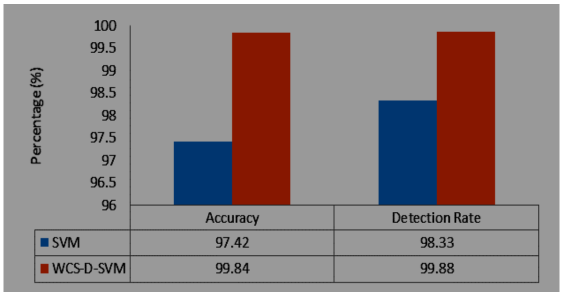

Figure 4: Accuracy and Detection Rate Using Training Records

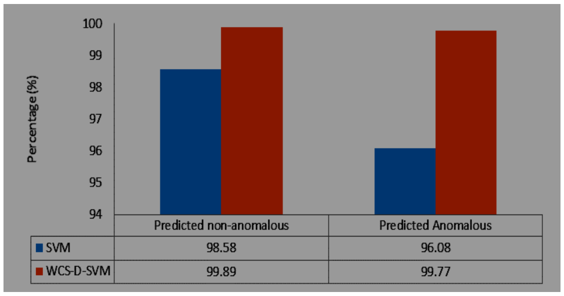

Figure 5: Percentage of Predicted Anomalous and Non-Anomalous Behaviour Using Training Records

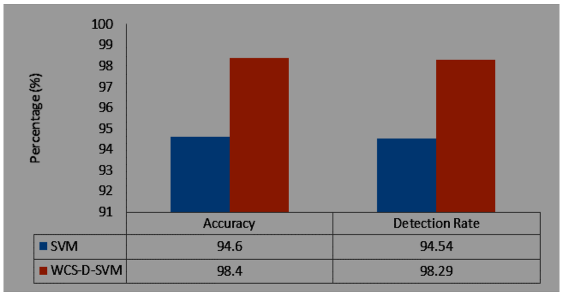

Figure 6: Accuracy and Detection Rate Using Testing Records

It is noticeable in Figure 8 that the SVM recorded poor results in predicting non-anomalous and anomalous records when it missed them by 7.33% (712 records) and 3.94% (505 records) compared to WCS-D-SVM at 2.28% (221) and 1.09% (140), respectively. Although the proportion of anomalous records (56.92%) are higher than the non-anomalous ones (43.08) as seen in Table 1, the WCS-D-SVM still performs significantly better than the SVM. In addition, the non-anomalous records which are similar to the anomalous one are much better than presented in the testing set. However, in the ability to handle high dimensional data, removing irrelevant features and interval values via the SVM, weighted chi-square and discretization process, the WCS-D-SVM is able to classify those similar behaviours more correctly than the SVM.

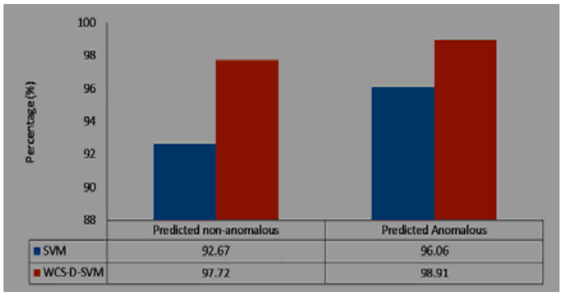

Figure 7: Percentage of Predicted Anomalous and Non-Anomalous Behaviour Using Testing Records

Table 3 illustrates a further comparison formed for the WCS-D-SVM using the similar NSL-KDD dataset as in previous related research in term of accuracy (AC), the detection rate (DR), false positive (FP) and true positive (TP). Furthermore, the selected method of the previous research is based on combination of feature selection as well as classifier that aimed for intrusion detection. For example, intrusion detection system based on self-organizing maps and statistical analysis (PSOM) hybridize as a classifier while the fisher's discriminant ratio (FDR) and principal component analysis (PCA) hybridize as a feature selection approach has been proposed in [52] with the aim to form a feature space which effective to differentiate the legitimate and illegitimate connections. In addition, the author has compared the proposed method called PSOM+PCA+FDR with a series of method such as PSOM+PCA and PSOM+FDR. It can be observed that WCS-D-SVM recorded 98.4% as an accuracy rate as compare to PSOM+PCA+FDR, PSOM+PCA and PSOM+FDR at 90%, 90% and 89% respectively. Furthermore, in [53] detection method through a combination of genetic algorithm (GA) and bagging approach (Bagged) with a series of classifier such as naïve bayes (NB), C4.5 and PART has proposed. The GA applied in a process of selecting certain relevant features to improve the accuracy. The GA+Bagged NB, GA+Bagged C4.5 and GA+Bagged PART have a drawback in its low AC rate at 73.97%, 77.86%, 78.37% and TP at

74%, 77.9%,78.4% as well as high FP at above 17% respectively. However, WCS-D-SVM AC and TP rate increased approximately by +20%, while FP reduced by -17%.

Table 3 illustrates a further comparison formed for the WCS-D-SVM using the similar NSL-KDD dataset as in previous related research in term of accuracy (AC), the detection rate (DR), false positive (FP) and true positive (TP). Furthermore, the selected method of the previous research is based on combination of feature selection as well as classifier that aimed for intrusion detection. For example, intrusion detection system based on self-organizing maps and statistical analysis (PSOM) hybridize as a classifier while the fisher's discriminant ratio (FDR) and principal component analysis (PCA) hybridize as a feature selection approach has been proposed in [52] with the aim to form a feature space which effective to differentiate the legitimate and illegitimate connections. In addition, the author has compared the proposed method called PSOM+PCA+FDR with a series of method such as PSOM+PCA and PSOM+FDR. It can be observed that WCS-D-SVM recorded 98.4% as an accuracy rate as compare to PSOM+PCA+FDR, PSOM+PCA and PSOM+FDR at 90%, 90% and 89% respectively. Furthermore, in [53] detection method through a combination of genetic algorithm (GA) and bagging approach (Bagged) with a series of classifier such as naïve bayes (NB), C4.5 and PART has proposed. The GA applied in a process of selecting certain relevant features to improve the accuracy. The GA+Bagged NB, GA+Bagged C4.5 and GA+Bagged PART have a drawback in its low AC rate at 73.97%, 77.86%, 78.37% and TP at 74%, 77.9%,78.4% as well as high FP at above 17% respectively. However, WCS-D-SVM AC and TP rate increased approximately by +20%, while FP reduced by -17%.

Moreover, in [54] a trade-off function to increase detection rate and decrease false alarm rate has proposed. As such, an efficient, robust and accurate optimization approach for feature selection and classification has been considered such as time varying chaos particle swarm optimization (TVCPSO), multiple criteria linear programming (MCLP) and support vector machine (SVM). The empirical result shows that TVCPSO-MCLP and TVCPSO-SVM obtain 96.88%, 97.84 as AC rate and 97.23%, 97.03% as detection rate. For the false positive rate, both recorded 2.41% and 0.87% respectively. In contrast, WCS-D-SVM approximately achieved better result in AC and DR with the increment of 1%, while decreasing false positive up to 0.22%. Recently, in [55], as realizing the difficulties in detecting intrusion that remain as open challenge, the researcher has proposed hypergraph based genetic algorithm (HG-GA) and support vector machine (SVM) as an adaptive and robust detection method. HG-GA applied as a weighted objective function for future selection while SVM as a classifier. The performance of related method has been compared with a number of combinational methods such s GA SVM and PSO SVM. It was concerning to mention that even though the HG-GA SVM outperform in term of FP at 0.83%, but still have room for improvement, and in fact WCS-D-SVM could achieved minimum FP at 0.022%. On the other hand, in [56], in order to identify various attacks in the networks, chi-square feature selection and multi class support vector machine namely CS-MCSVM has been proposed. Based on the result, the detection rate is nearly

Table 3: Performance Comparison of WCS-D-SVM With Recent Development in IDS

Author/Year	Methods	AC (%)	DR (%)	FP (%)	TP (%)
Eduardo DelaHoz et al. (2015) [52]	PSOM+FDR	89	-	-	-
	PSOM+PCA	90	-	-	-
	PSOM+PCA+FDR	90	-	-	-
Gaikward et al. (2015) [53]	GA+Bagged NB	73.97	-	21.2	74
	GA+Bagged C4.5	77.86	-	17.4	77.9
	GA+Bagged PART	78.37	-	17.2	78.4
Seyed et al. (2016) [54]	TVCPSO-MCLP	96.88	97.23	2.41	-
	TVCPSO-SVM	97.84	97.03	0.87	-
Gautama et al. (2017)[55]	GA-SVM	-	95.32	0.92	-
	PSO-SVM	-	93.49	1.09	-
	HG-GA SVM	-	97.14	0.83	-
Sumaiya et al. (2017) [56]	CS-MCSVM	98	-	0.13	-
The proposed method	**WCS-D-SVM**	**98.4**	**98.29**	**0.022**	**98.9**

0.4% below than WCS-D-SVM while FP lower at 0.1%. Hence, all the above results justify that there is a need for a feature selection and classifier in the development of an efficient detection method. The WCS-D-SVM could be an efficient IDS.

5. CONCLUSION AND FUTURE WORK

In this research, a weighted chi-square as a feature selection and a support vector machine as a classifier, namely WCS-SVM has been proposed for better intrusion behaviour detection. The WCS-SVM was evaluated using the well-known NSL-KDD dataset. The main aim of the WCS-SVM is to improve the detection of anomalous and non-anomalous behaviours. After the irrelevant features were terminated via the weighted chi-square and the SVM is able to handle high dimension data in various forms, the WCS-SVM significantly improved the accuracy of overall detection compared to the single support vector machine. Although researchers regularly propose and work on enhancing the detection capability, few focus on real time detection. Moreover, the primary concern currently is on the conventional types of attacks and not as much on critical insider and advanced persistent threats. As such, it is proposed that future research address this by exploring opportunities to detect and prevent such critical attacks in real time.

ACKNOWLEDGEMENT

The authors would like to express the appreciation to Inforsenet Group Research of Universiti Teknikal Malaysia Melaka (UTeM) and Cyber Security Malaysia (CSM) in encouraging the authors to publish this paper. This work was supported by UTeM and CSM under CMERP Grant – GLUAR/CSM/2016/FTMK-CACT/100013.

REFRENCES:

[1] Mohammadi, M., Muda, Z., Yassin, W., & Izura Udzi, N. (2014). KM-NEU: An Efficient Hybrid Approach for Intrusion Detection System. *Research Journal of Information Technology*, *6*(1), 46–57.

[2] Yassin, W., Udzir, N. I., Abdullah, A., Abdullah, M. T., Zulzalil, H., & Muda, Z. (2014). Signature-Based Anomaly intrusion detection using Integrated data mining classifiers. In *2014 International Symposium on Biometrics and Security Technologies (ISBAST)* (pp. 232–237). IEEE.

[3] Muda, Z., Yassin, W., Sulaiman, M., & Udzir, N. (2014). K-Means Clustering and Naive Bayes Classification for Intrusion Detection. *Journal of IT in Asia*, *4*.

[4] Yassin, W., Rahayu, S., Abdollah, F., & Zin, H. (2016). An Improved Malicious Behaviour Detection Via k-Means and Decision Tree. *International Journal of Advanced Computer Science and Applications*, *7*(12).

[5] Juma, S., Muda, Z., & Yassin, W. (2015). Machine Learning Techniques For Intrusion Detection System: A Review. *Journal of Theoretical and Applied Information Technology*, *72*(3).

[6] Dokas, P., Ertoz, L., Kumar, V., Lazarevic, A., Srivastava, J., & Tan, P. (2002). Data Mining for Network Intrusion Detection (pp. 21–30). proceedings of the nsf workshop on next generation data mining, Baltimore.

[7] Duque, S., & Omar, M. N. bin. (2015). Using Data Mining Algorithms for Developing a Model for Intrusion Detection System (IDS). *Procedia Computer Science*, *61*, 46–51.

[8] Fernandes, G., Carvalho, L. F., Rodrigues, J. J. P. C., & Proença, M. L. (2016). Network anomaly detection using IP flows with Principal Component Analysis and Ant Colony Optimization. *Journal of Network and Computer Applications*, *64*, 1–11.

[9] Fernandes, G., Rodrigues, J. J. P. C., & Proença, M. L. (2015). Autonomous profile-based anomaly detection system using principal component analysis and flow analysis. *Applied Soft Computing*, *34*, 513–525.

[10] Gaikwad, D. P., & Thool, R. C. (2015). Intrusion Detection System Using Bagging with Partial Decision TreeBase Classifier. *Procedia Computer Science*, *49*, 92–98.

[11] Grill, M., Pevný, T., & Rehak, M. (2016). Reducing false positives of network anomaly detection by local adaptive multivariate smoothing. *Journal of Computer and System Sciences*.

[12] Ji, S.-Y., Jeong, B.-K., Choi, S., & Jeong, D. H. (2016). A multi-level intrusion detection method for abnormal network behaviors. *Journal of Network and Computer Applications*, *62*, 9–17.

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

[14] Rai, K., Devi, M.S. and Guleria, A. (2016). Decision Tree Based Algorithm for Intrusion Detection. *International Journal of Advanced*

Networking and Applications, *2834*, 2828–2834.

[15] Buczak, A. L., & Guven, E. (2016). A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. *IEEE Communications Surveys & Tutorials*.

[16] Parmezan, A. R. S., Lee, H. D., & Wu, F. C. (2017). Metalearning for choosing feature selection algorithms in data mining: Proposal of a new framework. *Expert Systems with Applications*, *75*, 1–24.

[17] Ramírez-Gallego, S., Krawczyk, B., García, S., Woźniak, M., & Herrera, F. (2017). A survey on data preprocessing for data stream mining: Current status and future directions. *Neurocomputing*, *239*, 39–57.

[18] Sheikhpour, R., Sarram, M. A., Gharaghani, S., & Chahooki, M. A. Z. (2017). A Survey on semi-supervised feature selection methods. *Pattern Recognition*, *64*, 141–158.

[19] Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*(1–2), 273–324.

[20] Witten, I. H. , Frank, E., & H. (2016). *Practical Machine Learning Tools and Techniques* (4th ed.). Morgan Kaufmann.

[21] Chuvakin, A., Schmidt, K. J., Phillips, C., & Moulder, P. (2013). *Logging and Log Management: The Authoritative Guide to Understanding the Concepts Surrounding Logging and Log Management*. Elsevier.

[22] Liu, H. , & Motoda, H. (2013). *Feature selection for knowledge discovery and data mining*. (Springer, Ed.). United States of America: The Springer International Series in Engineering and Computer Science.

[23] Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16–28.

[24] Miao, J., & Niu, L. (2016). A Survey on Feature Selection. *Procedia Computer Science*, *91*, 919–926.

[25] Benabdeslem, K., & Hindawi, M. (2014). Efficient Semi-Supervised Feature Selection: Constraint, Relevance, and Redundancy. *IEEE Transactions on Knowledge and Data Engineering*, *26*(5), 1131–1143.

[26] Reif, M., & Shafait, F. (2014). Efficient feature size reduction via predictive forward selection. *Pattern Recognition*, *47*(4), 1664–1673.

[27] Visalakshi, S., & Radha, V. (2014). A literature review of feature selection techniques and applications: Review of feature selection in data mining. In *2014 IEEE International Conference on Computational Intelligence and Computing Research* (pp. 1–6). IEEE.

[28] S. Juma, Muda, Z., & Yassin, W. (2014). Reducing False Alarm Using Hybrid Intrusion Detection Based On X-Means Clustering and Random Forest Classification. *Journal of Theoretical and Applied Information Technology*, *68*(2), 249–254.

[29] Ganapathy, S., Kulothungan, K., Muthurajkumar, S., Vijayalakshmi, M., Yogesh, P., & Kannan, A. (2013). Intelligent feature selection and classification techniques for intrusion detection in networks: a survey. *EURASIP Journal on Wireless Communications and Networking*, *2013*(1), 1–16.

[30] Liu, H., & Motoda, H. (2007). *Computational Methods of Feature Selection (Chapman & Hall/Crc Data Mining and Knowledge Discovery Series)*. Chapman & Hall/CRC.

[31] Aparicio-Navarro, F. J., Kyriakopoulos, K. G., & Parish, D. J. (2014). Automatic Dataset Labelling and Feature Selection for Intrusion Detection Systems. In *2014 IEEE Military Communications Conference* (pp. 46–51). IEEE.

[32] Balakrishnan, S., K, V., & A, K. (2014). Intrusion Detection System Using Feature Selection and Classification Technique. *International Journal of Computer Science and Application*, *3*(4), 145.

[33] Yin, C., Ma, L., Feng, L., Yin, Z., & Wang, J. (2015). A Feature Selection Algorithm towards Efficient Intrusion Detection. *International Journal of Multimedia and Ubiquitous Engineering*, *10*(11), 253–264.

[34] Yassin, W., Udzir, N., Abdullah, A., Abdullah, M., Muda, Z., & Zulzalil, H. (2014). Packet Header Anomaly Detection Using Statistical Analysis. In J. G. de la Puerta, I. G. Ferreira, P. G. Bringas, F. Klett, A. Abraham, A. C. P. L. F. de Carvalho, … E. Corchado (Eds.), *International Joint Conference SOCO'14-CISIS'14-ICEUTE'14 SE - 47* (Vol. 299, pp. 473–482). Springer International Publishing.

[35] M. Aghdam, & Kabiri., P. (2016). Feature Selection for Intrusion Detection System Using Ant Colony Optimization. *International Journal of Network Security*, *18*(3), 420–432. Retrieved from

[36] Hasan, M. A. M., Nasser, M., Ahmad, S., & Molla, K. I. (2016). Feature Selection for

Intrusion Detection Using Random Forest. *Journal of Information Security*, *7*(3), 129–140.

[37] Nkikabahizi, C., Cheruiyot, W., & Kibe, A. (2017). Classification and Analysis of Techniques Applied in Intrusion Detection Systems. *International Journal of Scientific Engineering and Technology*, *6*(7), 216.

[38] Sapate, P., & A.Raut, S. (2014). Survey on Classification Techniques for Intrusion Detection. In *Computer Science & Information Technology (CS & IT)* (pp. 223–231). Academy & Industry Research Collaboration Center (AIRCC).

[39] Chen, R.-C., Cheng, K.-F., Chen, Y.-H., & Hsieh, C.-F. (2009). Using Rough Set and Support Vector Machine for Network Intrusion Detection System. In *2009 First Asian Conference on Intelligent Information and Database Systems* (pp. 465–470). IEEE.

[40] Wang, H., Zhang, G., Mingjie, E., & Sun, N. (2011). A novel intrusion detection method based on improved SVM by combining PCA and PSO. *Wuhan University Journal of Natural Sciences*, *16*(5), 409–413.

[41] Rajesh, W., Chole, V., & Shruti, K. (2015). A Review On Intrusion Detection System Using Classification Technique. *International Journal of Advanced Computational Engineering and Networking*, *3*(12), 62–65.

[42] Umer, M. F., Sher, M., & Bi, Y. (2017). Flow-based intrusion detection: Techniques and challenges. *Computers & Security*, *70*, 238–254.

[43] Gyanchandani Manasi, J.L.Rana, & R.N.Yadav. (2012). Taxonomy of anomaly based intrusion detection system: A review. *International Journal of Scientific and Research Publications (IJSRP)*, *2*(12), 1–14.

[44] Liao, H.-J., Richard Lin, C.-H., Lin, Y.-C., & Tung, K.-Y. (2013). Intrusion detection system: A comprehensive review. *Journal of Network and Computer Applications*, *36*(1), 16–24.

[45] Belavagi, M. C., & Muniyal, B. (2016). Performance Evaluation of Supervised Machine Learning Algorithms for Intrusion Detection. *Procedia Computer Science*, *89*, 117–123.

[46] Al-Jarrah, O., & Arafat, A. (2015). Network Intrusion Detection System Using Neural Network Classification of Attack Behavior. *Journal of Advances in Information Technology*, 1–8.

[47] Hussain, J., & Lalmuanawma, S. (2016). Feature Analysis, Evaluation and Comparisons of Classification Algorithms Based on Noisy Intrusion Dataset. *Procedia Computer Science*, *92*, 188–198.

[48] Kabir, E., Hu, J., Wang, H., & Zhuo, G. (2017). A novel statistical technique for intrusion detection systems. *Future Generation Computer Systems*.

[49] Cortes, C., & Vapnik, V. (1995). Support Vector Networks. *Machine Learning*, *20*(3), 273–297.

[50] Wang, P., & Wang, Y.-S. (2015). Malware behavioural detection and vaccine development by using a support vector model classifier. *Journal of Computer and System Sciences*, *81*(6), 1012–1026.

[51] Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. a. (2009). A detailed analysis of the KDD CUP 99 data set. *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, (Cisda), 1–6.

[52] Eduardo DelaHoz, De La Hoz, E., Ortiz, A., Ortega, J., & Prieto, B. (2015). PCA filtering and probabilistic SOM for network intrusion detection. *Neurocomputing*, *164*, 71–81.

[53] Gaikwad, D. P., & Thool, R. C. (2015). Intrusion Detection System Using Bagging with Partial Decision TreeBase Classifier. *Procedia Computer Science*, *49*, 92–98.

[54] Seyed Hosseini Bamakan, S. M., Wang, H., Yingjie, T., & Shi, Y. (2016). An effective intrusion detection framework based on MCLP/SVM optimized by time-varying chaos particle swarm optimization. *Neurocomputing*, *199*, 90–102.

[55] Gauthama Raman, M. R., Somu, N., Kirthivasan, K., Liscano, R., & Shankar Sriram, V. S. (2017). An efficient intrusion detection system based on hypergraph - Genetic algorithm for parameter optimization and feature selection in support vector machine. *Knowledge-Based Systems*, *134*, 1–12.

[56] Sumaiya Thaseen, & Kumar, C. A. (2013). An analysis of supervised tree based classifiers for intrusion detection system. In *2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering* (pp. 294–299).