Intrusion Detection Method Based on Support Vector Machine and Information Gain for Mobile Cloud Computing

Emmanuel Mugabo and Qiu-Yu Zhang
(*Corresponding author: Qiu-Yu Zhang*)

School of Computer and Communication, Lanzhou University of Technology
No.287, Lan-Gong-Ping Road, Lanzhou 730050, China
(Email: zhangqylz@163.com)

(Received Aug. 26, 2018; Revised and Accepted Mar. 23, 2019; First Online June 5, 2019)

Abstract

Intrusion detection system (IDS) has become an important security method that monitors and investigates the network security in mobile cloud computing (MCC). However, in some existing methods, there are still some limitations such as high false positive rates, low classification accuracies, and low true positive rates. To counter these limitations, an intrusion detection method based on support vector machine (SVM) and information gain (IG) for MCC was proposed in this paper. In the proposed method, the SVM classifier is adopted to classify network data into normal and attack behaviors, and due to the irrelevant and redundant features found in KDD datasets, IG is used to select the relevant features and remove unnecessary features. The KDD'99 and NSL-KDD datasets are used to evaluate the effectiveness of the proposed method. Compared with other methods, the experimental results show that the proposed method can detect malicious attacks with high accuracy, true positive rate, low false positive rate and high training speed.

Keywords: Intrusion Detection; Information Gain; Malicious Attacks; Mobile Cloud Computing; Support Vector Machine

1 Introduction

Mobile Cloud Computing (MCC) is an exciting new technology, which integrates cloud computing into the mobile environment [4, 22]. According to Cisco IBSG (Online, 2016), close to 85% of the world's population has access to mobile devices as they bring some convenience, but at the same time, endless security issues also follow. Mobile cloud applications move the computing power and data storage away from mobile devices and into the cloud, which enable users to access network services anywhere and anytime [4, 10]. Furthermore, MCC provides simple and easy infrastructure for mobile applications and services, and it enables users to utilize resources on demand, and take full advantage of cloud computing services. However, due to its distributed nature and easy to use, MCC faces many technical challenges such as privacy, security, and so on.

To counter security issues in MCC like intruders or cyber-attacks, it is necessary to detect those attacks earlier by implementing immediate countermeasures to prevent the harmful risks [8]. Based on the existing security issues solutions, there have been two most useful techniques to defend mobile cloud services against intruders, such as firewall technology and IDS. In the research of cloud environment intrusion detection problems, many researchers have been using mainly four approaches based IDS such as clustering, classification, information theory, and statistical theories to deal with intrusion detection problems [8, 21]. Most recently, researchers have adopted different approaches like deep learning [19], Naïve Bayes [7], neural network [5, 25, 29], SVM [5, 9, 15, 16, 19, 20], genetic algorithm (GA) [13, 15], *etc.*

Meanwhile, The KDD'99 and NSL-KDD datasets have been used by many researchers to survey and evaluate research in intrusion detection. The KDD'99 dataset has not only been the most useful dataset in IDS, but also a benchmark for evaluating the best performance of intrusion detection methods [2, 19]. On the other hand, the NSL-KDD dataset is a new version of the KDD dataset, which has some advantages over the original KDD'99 dataset such as no redundant records in the training set and duplicate records in the test set of the NSL-KDD dataset [3]. After data collection, most of the datasets require feature analysis and dimensionality reduction to extract and select the data that is most likely to produce accurate results and reduce the computing cost and timing cost of the IDS [8,19]. The most recent feature analysis and dimension reduction methods used in cloud computing include principal component analysis (PCA) [19], information gain (IG) [6, 12, 25], genetic algorithm (GA)

based feature selection [15], *etc.*

However, many researchers have used different intrusion detection techniques to provide security for both mobile computing and networks, but still, have the common limitations on low detection accuracy, low true positive rate and high false positive rate. Motivated by this above, this paper proposed an intrusion detection method based on SVM and IG approach, which detects different malicious attacks with high detection accuracy and low false positive rate.

The remainder of this paper is organized as follows: Section 2 presents the recent related works. The problem statement and preliminaries of MCC and other related theories are explained in Section 3. The proposed intrusion detection method of MCC using SVM-IG is presented in Section 4. Section 5 gives the experimental results and performance analysis as compared with other related methods. Finally, we conclude our paper in Section 6.

2 Related Works

Currently, with the rapid development of cloud computing environment, the cloud security has become a serious challenge, and many researchers have adopted different techniques and methods such as machine learning techniques and data mining to improve the capability of IDS [5,29]. Among those techniques, artificial neural network (ANN) and SVM are the most useful methods in the cloud computing area [12, 19, 25, 29].

In [4, 18], the detailed surveys of data security in the MCC are discussed. Li *et al.* [16] proposed an IDS model based on rough set theory (RST) and fuzzy SVM (FSVM), the proposed method uses RST to reduce the dimensions of features, and the experimental results show that the proposed RST-FSVM can do better for IDSs. Hoque *et al.* [13] proposed an IDS model based on GA that filter and reduce the complexity of data; by using KDD'99 dataset, the reasonable detection rate has been achieved but got a slightly high false positive rate. Kannan *et al.* [15] proposed an intrusion detection model that combines genetic based feature selection and FSVM to secure the cloud networks. The proposed genetic based feature selection improved the detection accuracy of the FSVM classifier by selecting the relevant attributes in the KDD'99 dataset.

Zhang *et al.* [27] proposed an intrusion detection method based on cloud model and semi-supervised clustering, and the simulation results show that the performance of intrusion detection method has improved. Hoz *et al.* [14] proposed a network anomaly classification using support vector classifier and non-linear projection techniques; the experimental results show that the reasonable true positive rate has been achieved using NSL-KDD dataset, but has a high false positive rate. Deshmukh *et al.* [7] proposed an IDS model based preprocessing methods and Naïve Bayes classifier; The experimental results show that after applying preprocessing methods including discretization, normalization and feature selection using NSL-KDD dataset, the proposed method effectively improved the performance of IDS.

Pervez *et al.* [20] proposed a feature selection and SVM classifier using NSL-KDD dataset. Eesa *et al.* [9] proposed a new feature selection based on cuttlefish optimization algorithm (CFA) and the decision tree classifier, and by using the KDD'99 dataset, the results show that the proposed approach gives a high accuracy and detection rate with lower false positive rate compared with the results using all 41 features.

Yuan *et al.* [26] proposed a semi-supervised AdaBoost algorithm for network anomaly detection, and the experimental results show that the proposed method can achieve a good result even with a small labeled dataset. Nguyen *et al.* [19] proposed a deep learning approach that detects cyber-attacks in MCC, in this framework, PCA was used for the feature extraction and dimension reduction, and by using KDD'99, NSL-KDD and UNSW-NB-15 datasets, a good accuracy, and detection rate have been achieved, but they do not evaluate the false positive rate.

Ashfaq *et al.* [3] proposed a fuzziness based semi-supervised learning approach using neural network with random weights (NNRw) as a classifier, and through the experiments, NSL-KDDTest+ and NSL-KDDTest-21 are used to test the proposed method. Hammami *et al.* [12] proposed a cloud computing based IDS and human-immune system, the NSL-KDD dataset is used to evaluate the performance of the proposed method.

Zhao *et al.* [29] proposed an intrusion detection approach based SOM neural network in cloud computing, where the particle swarm optimization (PSO) based on simulated annealing was used to optimize the SOM neural network, and the experimental results showed that the PSO algorithm has effectively improved the performance of the system.

In [17, 27], the approaches of network intrusion detection based PCA using SVM were proposed, and PCA was used to reduce the higher dimensional KDD dataset to lower dimensional dataset, and the experimental results showed that PCA has effectively improved the performance of the IDS compared to without using PCA.

Wang [25] proposed an IDS for cloud computing using MLP and K-means algorithm. Information gain, which is a feature selection method, was used to select the most relevant features and remove unneeded features in the KDD'99 dataset. The simulation results show that the proposed approach has good performance compared to each of MLP and K-means respectively.

Aghdam *et al.* [1] proposed a feature selection for IDS using Ant Colony Optimization that identifies important features and improves the performance of IDS. The experimental results on the KDD Cup 99 and NSL-KDD datasets show that the proposed method provides high accuracy and low false positive rate in detecting intrusions.

3 Problem Statement and Preliminaries

3.1 Intrusion Detection System (IDS)

IDS is defined as a software application or a security management tool that controls all events occurring in a computing system or network and analyzing them to find the intrusions or malicious activities either within the system or outside the system [10, 20]. Intrusion is defined as the attempts that are used to compromise the data integrity, confidentiality, and data availability in a computing system [13].

Considering the methods of data collection, there are two distinct types of IDSs: "Host-based IDS" and "Network-based IDS" and considering the detection techniques of intrusions, there are two approaches to detect intrusions: "Misuse or signature detection based IDS" and "Anomaly detection based IDS" [10, 20]. In misuse detection based, the intrusions are detected by searching for activities that correspond to known attacks; While in anomaly detection based, the intrusions are detected by searching for deviations from a normal behavior model.

3.2 Mobile Cloud Computing (MCC)

MCC is a fast-growing architecture, which integrates mobile computing and wireless technology with cloud computing, where the mobile users utilize different cloud services anytime and anywhere based on the pay-as-you-use principle [10, 18]. The cloud computing provides various services such as Infrastructure as a Service (IaaS), Software as a Service (SaaS), and Platform as a Service (PaaS) to mobile computing in order to tackle the lack of enough storage space and processing power in mobile devices [11]. Although the MCC may seem to be a very exciting technology nowadays, there still remains some technical challenges, more specifically the security of data or information stored in the cloud [4, 11].

Figure 1 shows the architecture of mobile cloud computing.

3.3 Support Vector Machine (SVM)

The SVM is a supervised machine learning method that is used in data mining to analyze data and recognize patterns in the dataset for regression and classification purpose [20]. Nowadays, SVMs are used for linear and nonlinear classifications and support both binary and multiclass classifications. The datasets used in IDS are often high dimensional and heterogeneous. Because the traditional SVMs cannot directly deal with heterogeneous datasets, there is a need to use some kernel functions like Radial Bias Function (RBF), Linear Function, *etc.* in order to extend them on heterogeneous datasets [20]. In Binary classification, the SVM finds for a maximum margin hyperplane to separate the two classes, a class of positive samples and class of negative samples of the training set based on structural risk minimization analysis of statistical learning theory [23].

Let us assume that our binary classification has an n-dimensional feature space of a training set as follows:

$$T_s = \{(x_1, y_1), \ldots, (x_m, y_m)\} \in (R^n \times \{-1, +1\})^m, \quad (1)$$

where $x_i \in ([x_i]_1, [x_i]_2, \ldots, [x_i]_n)^{T_s}$ is the input feature vectors, $y_i \in \{-1, +1\}$ is the binary output of x_i, m is the number of samples in the feature space and R^n an n-dimensional real space.

The classification function of SVM is defined as follows:

$$f(x) = w^{T_s} \cdot x + b, \quad (2)$$

where b is the bias and w is a weight vector.

Thus, the training set should satisfy the following condition:

$$f(x) = \begin{cases} w^{T_s} \cdot x_i + b \geq -1, & \text{for all attack data } x_i \\ w^{T_s} \cdot x_i + b \leq +1, & \text{for all normal data } x_i \end{cases}$$

Figure 2 demonstrates the maximum-margin hyperplane and margins for an SVM classifier trained with samples from two classes either normal or malicious attack.

The main goal of SVM is to find the optimal hyperplane by maximizing the margin between two classes as can be seen in Figure 2. The distance between two hyperplanes is $\frac{2}{\|w\|}$, the optimization objective is just to minimize $||w||$, and this can be obtained by solving the following quadratic optimization problem:

$$\begin{cases} \textit{minimize } (\textit{in } \text{w,b}) \ \frac{1}{2}||w|| \\ \textit{subject to} : y_i(\langle w \cdot x_i \rangle + b) \quad \text{for } i\text{=1,2,...,m} \end{cases} \quad (3)$$

The quadratic optimization problem in Equation (3) can be solved by the sequential minimal optimization (SMO) algorithm using the Lagrange multipliers a_i as follows:

$$\begin{cases} \textit{maximize } L(\alpha) = \\ \qquad \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{j=1}^{m} \sum_{i=1}^{m} y_i y_j \alpha_i \alpha_j K(x_i \cdot y_j) \\ \textit{subject to} : \sum_{i=1}^{m} y_i \alpha_i = 0, 0 \leq \alpha_i \leq C, 1 \leq i \leq m \end{cases} \quad (4)$$

where L is the Lagrange function, C is the regularization parameter and K is the kernel function.

After solving Equation (4), we obtain $w = \sum_{i=1}^{m} y_i \alpha_i x_i$, and the decision attack function is defined as follows:

$$f(x, \alpha, b) = \{\pm 1\} = sgn(\sum_{i=1}^{m} y_i \alpha_i K(x, x_i) + b), \quad (5)$$

where b is obtained from Karush-Kuhn-Tucker condition.

To construct SVM classifier, a kernel function and some parameters have to be selected. There are three main types of SVM kernel function: Linear kernel function, Radial Bias kernel function (RBF) and polynomial kernel

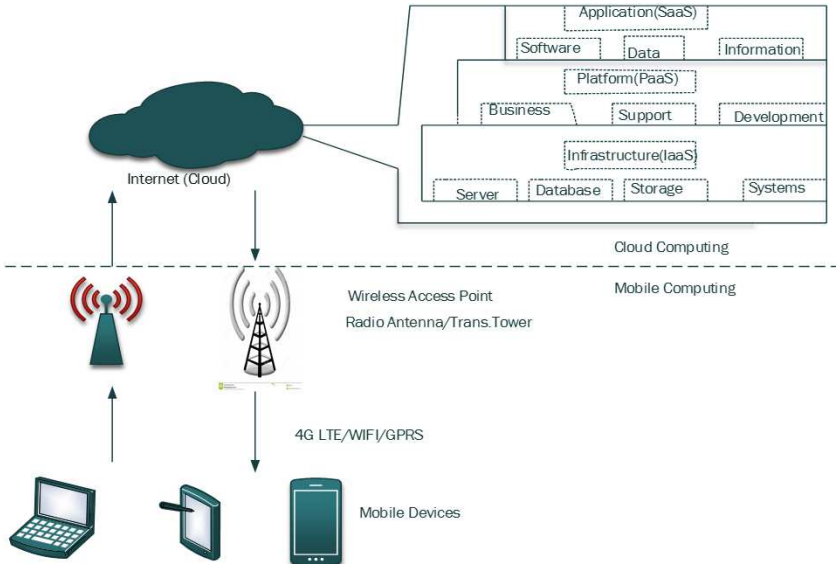

Figure 1: Architecture of mobile cloud computing

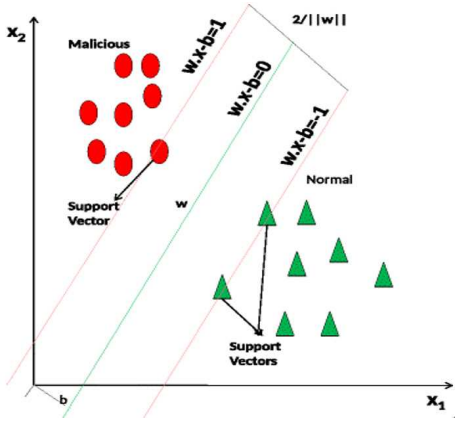

Figure 2: Maximum separating the hyperplane and the margin

function. It is very advantageous to use SVM classifier as it has the ability to give very accurate results, specifically for binary classification, and it is very effective in high dimensional datasets. In addition to that, the SVM classifier is very robust against overfitting and outliers.

3.4 Information Gain Based Feature Selection (IG)

The datasets, which are mostly used for analyzing IDSs like KDD'99 and NSL-KDD datasets, have been facing a serious problem of large amount of records including redundant and irrelevant records, as well as relevant records. To counter this above problem, many researchers have adopted various methods like dimensionality reduction, feature selection and so on. According to [1], feature selection is a task of identifying the relevant features from all features and removing the irrelevant or inappropriate ones in the dataset. In this paper, we have adopted IG based feature selection to reduce the computation complexity of the dataset.

According to [6], IG is a method used to decide which attribute in a given dataset is most important to be used in the machine learning process for classifying data. The IG uses Shannon's entropy to measure the feature set quality.

Let's consider S to be a set of training set samples of any dataset. Suppose that there are n classes and the training set contains s_i samples of class I and s is the total number of samples in the training set. Then the expected information needed to classify a given sample is solved as follows:

$$I(s_1, s_2, \ldots, s_n) = \sum_{i=1}^{n} \frac{s_i}{s} \log 2 \left(\frac{s_i}{s}\right) \tag{6}$$

The training set S can be divided into v subsets S_1, S_2, ..., S_v by an attribute A with values a_1, a_2, ..., a_v, where S_j is the training subset, which has the value a_j for attribute A. Additionally, let S_j contains s_{ij} samples of class I. The Entropy of the attribute A is as follows:

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + s_{2j} + \cdots + s_{nj}}{s} \times I(s_{1j}, \ldots, s_{nj}). \tag{7}$$

Therefore, the IG for attribute or feature A can be calculated as follows:

$$Gain(A) = I(s_1, \ldots, s_n) - E(A). \tag{8}$$

3.5 Dataset Description

The datasets are used to survey and evaluate a research in intrusion detection, some are self-created datasets, and others are publicly available. Over the last years, most researchers have adopted KDD'99 Cup and NSL-KDD datasets, which are publicly available to train and test the performance of an intrusion detection research [7,11].

3.5.1 KDD'99 Cup Dataset

The KDD'99 Cup is a version of DARPA 1998 dataset, which has been provided by MIT Lincoln laboratory in 1999. The complete KDD'99 dataset has up to 5 million input records, and every record represents a TCP/IP connection that consists of 41 features (3 of them are symbolic, and 38 are numeric) and one marked as either normal or attack, and all attacks fall into four major categories: Probe, DoS, U2R and R2L [1, 7, 11]. Because of this large amount of records in the KDD'99 dataset, it has been grouped into three independent subsets: "10% KDD", "Corrected KDD" and "Whole KDD" [11]. Most of the researchers prefer to use 10% KDD for training set and corrected KDD for test set.

3.5.2 NSL-KDD Dataset

The NSL-KDD dataset is considered as a reduced version of KDD'99 Cup, where it overcomes the problem of redundant records existing in KDD'99 Cup training set, the size of the dataset is reduced compared to that of KDD'99, and has no duplicate data in the improved test set [2]. As for the KDD'99 dataset, NSL-KDD dataset also has 41 features and one marked as either normal or attack [11,28]. Apart from having several advantages over KDD'99, the NSL-KDD is still not yet a perfect representative for existing real networks compared to the original KDD'99 dataset, due to the lack of public datasets for network-based IDSs, but it still can be used as an effective benchmark dataset to compare different intrusion detection methods [1, 11].

Table 1 describes the number of records in KDD'99 (10% KDD'99 for training and corrected KDD for testing) and NSL-KDD datasets.

4 The Proposed Method

The flowchart of the intrusion detection method based on SVM and IG for MCC is depicted in Figure 3. The proposed method is divided into two phases namely; data preparing phase and intrusion detection phase.

The data-preparing phase has two main functions, *i.e.*, data collection from the KDD'99 and NSL-KDD datasets and data preprocessing which is divided into three parts: "Data discretization", "Feature selection" and "Data normalization".

In data preparing phase, the packet features collected from KDD'99 and NSL-KDD datasets are first discretized where not all the 41 features of KDD datasets are continuous or discrete values. The features like protocol type (TCP, UDP and ICMP), network service need to be converted into numbers. Among the 41 features, some are irrelevant or redundant leading to a long detection process and degrading the system's performance. Therefore, selecting the most relevant features is an essential way to increase the performance and reduce the computing and timing cost; here the IG based feature selection is used. Thus, the data normalization part follows, in order to scale the data in a specific range; we adopted the min-max normalization method.

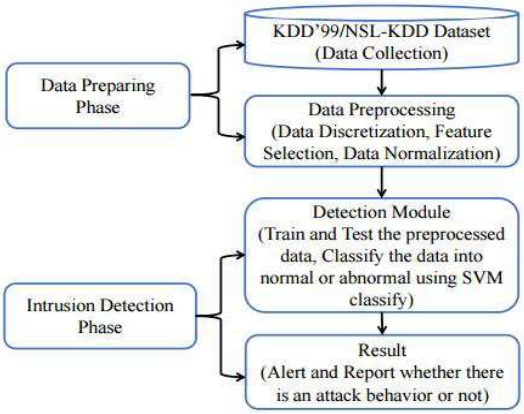

Figure 3: Intrusion detection method of MCC based on SVM and IG

In intrusion detection phase, the selected features from the data preprocessing part are trained, tested and then classified into normal or attack by using the SVM classifier. Therefore, the final result will be reported to the system administrator, and if there is an attack, the system administrator will deal with it accordingly; otherwise, the packet feature will be served as normal.

The proposed SVM-IG algorithm in this paper is defined as follows:

Algorithm 1 SVM-IG

1: Input: The KDD'99 and NSL-KDD training and test data
2: Output: The evaluation metrics of the proposed method (ACC, TPR, PPV, and FPR)
3: Obtain the input data
4: **while** training data **do**
5: Preprocessing of data
6: Consider the RBF kernel function
7: Use the cross-validation to find the best C and gamma (γ) parameters
8: Use the best parameters C and γ to train the whole training set
9: **end while**
10: **while** test data **do**
11: Preprocessing of data
12: Evaluate and predict the output
13: **end while**

5 Experimental Results and Analysis

All the experiments are performed on a Compaq-HP computer with 2.4 GHz Intel (R) Core (TM) i3-3110M

Table 1: KDD'99 and NSL-KDD training and test dataset records

Name	Dataset	Records	Normal	Probe	DoS	U2R	R2L
KDD'99 (10%)	Train	494021	97278	4107	391458	52	1126
Corrected KDD	Test	311029	60593	4166	229853	228	16189
NSL-KDD	$Train+\_20$	25192	13449	2289	9234	11	209
NSL-KDD	Test-21	22544	9711	2421	7458	200	2754

and 10 GB of RAM, and running on windows 10 Enterprise (64bits). The proposed method was implemented in MATLAB R2018b and Weka 3.8.3 data mining tool. The SVM classifier is applied with LibSVM package (MATLAB version 3.23).

5.1 Dataset and Data Preprocessing

In the evaluation process of the proposed method, we have used 10% of the full KDD'99, corrected KDD, NSL-KDD Train+_20 and NSL-KDD Test-21 datasets for training and testing our model as shown in Table 1. All these KDD datasets contain 41 features and one marked as either normal (1) or attack (0). The attacks fall into four major types: Denial of Service (DoS), Probe, Remote-to-Local (R2L), and User-to-Root (U2R).

In data preprocessing, we adopted the Weka tool to perform the data discretization and feature selection. In data discretization, the continuous features are converted to discrete or nominal features by using the Weka discretization filter and InfoGainAttributeEval with Ranker available in Weka tool is used to select the most important features. Table 2 shows the most relevant features selected by using IG based feature selection.

Through the feature selection analysis, the features {20, 21} for both KDD'99 and NSL-KDD datasets show zero information gain which means they do not contribute to the intrusion detection. The features {5, 6, 7, 9, 10, 11, 13, 14, 15, 16, 17, 18} for NSL-KDD dataset and {5, 6, 7, 9, 13, 14, 15, 16, 17, 18} for KDD'99 dataset have a very small information gain, which has a little effect to the intrusion detection. The stated above features are removed from the datasets due to the small contribution on the intrusion detection. Therefore, by using the IG based feature selection, the most relevant features used in our method for both KDD'99 and NSL-KDD datasets are shown in Table 2 and the dimension of both datasets was reduced as well. Furthermore, the features like protocol type and network service cannot be sent directly to the system, and hence they need to be preprocessed and converted into numerical digits. For example, protocol types like TCP, UDP, and ICMP are converted to number 1, 2 and 3 the same as for other network services. Therefore, the numerical values are scaled within a specified range. In the proposed method, we scaled the numerical values within a range of [0, 1] by using the min-max normalization method.

$$f(x) = \frac{(x - x_{min})}{(x_{max} - x_{min})} \in [0,1]; x \in [x_{min}, x_{max}] \quad (9)$$

where x is an attribute value, x_{min} is the minimum attribute value, x_{max} is the maximum attribute value, and $f(x)$ is the normalized value.

5.2 Performance Metrics

In order to measure the performance of the MCC based IDS, the true positive rate (TPR), false positive rate (FPR), accuracy (ACC), and precision (PPV) indicators are used for measurement. A confusion matrix is used to represent the information related to the actual and predicted classifications performed by the classification system.

The confusion matrix is shown in Table 3. In Table 3, TP indicates that the actual is a normal sample and is predicted as the number of normal samples, FN indicates that the actual is a normal sample and is predicted as the number of abnormal samples, FP indicates that the actual is an abnormal sample and is predicted as the number of normal samples, and TN indicates that the actual is an abnormal sample and is predicted as the number of normal samples.

The ACC, PPV, TPR, and FPR are the four main performance metrics used for the proposed method as described below:

$$\begin{aligned} ACC &= \frac{TP+TN}{TP+FN+FP+TN} \\ PPV &= \frac{TP}{TP+FP} \\ TPR &= \frac{TP}{TP+FN} \\ FPR &= \frac{FP}{FP+TN} \end{aligned}$$

where ACC shows a total number of corrected predictions, PPV indicates that the intrusion predicted by IDS is an actual intrusion, TPR determines the correctly identified positive instances, FPR indicates the normal cases that incorrectly identified as an anomaly.

5.3 Performance Evaluation

After the data preprocessing, the reduced data is fed into the model and processed via LibSVM, an open source for SVM Classifier and Radial Bias Kernel Function (RBF Kernel) which has two hyperparameters C and Gamma (γ), was used to study the effectiveness of the SVM classifier. The parameters C and Gamma (γ) were tuned to find the better cross-validation (Cross Val) accuracy by

Table 2: List of most relevant features in KDD'99 and NSL-KDD dataset using IG

KDD'99 dataset		NSL-KDD dataset	
No.	Selected Features	No.	Selected Features
2	protocol type	3	service
3	service	4	flag
4	flag	12	*logged_in*
23	count	23	count
24	*srv_count*	25	*serror_rate*
25	*serror_rate*	26	*srv_serror_rate*
26	*srv_serror_rate*	29	*same_srv_rate*
29	*same_srv_rate*	32	*dst_host_count*
33	*dst_host_srv_count*	33	*dst_host_srv_count*
34	*dst_host_serror_rate*	34	*dst_host_same_srv_rate*
36	*dst_host_rerror_rate*	38	*dst_host_serror_rate*
38	*dst_host_same_srv_rate*	39	*dst_host_srv_serror_rate*
39	*dst_host_diff_srv_rate*		

Table 3: Confusion Matrix

		Predicted	
		Attack	Normal
Actual	Attack	TP	FN
	Normal	FP	TN

using grid search method and the ones with good Cross Val accuracies were picked and used to train and validate the proposed method as can be seen from the Table 4, Table 5 and Figure 4, Figure 5, Figure 6 and Figure 7 below.

In the proposed method, we have used 10-fold Cross Val to tackle the overfitting problem, which divides the dataset into 10 sub-sets of size N/10 (N is the size number of the dataset) and uses 9 sub-sets for training and 1 remaining sub-set for testing. The practical way to find better parameters of C and γ is to try the exponential growing sequences of them by using coarse and fine grid-search methods. The grid-options such as $log_2 C$ and $log_2 \gamma$ are used to run the SVM classifier for a certain range of C and γ parameters. Figure 4 and Figure 5 shows the coarse grid-search and fine grid-search for the KDD'99 dataset.

For KDD'99 dataset, we conducted the coarse grid-search on $\log_2 C \in [-5, 15]$ and $\log_2 \gamma \in [-14, 2]$, Figure 4 shows the coarse grid-search with an exponential growing sequence of C and γ ($C = 2^{-5}, 2^{-4}, \ldots, 2^{14}, 2^{15}; \gamma = 2^{-14}, 2^{-13}, \ldots, 2^{3}, 2^{2}$), which gives us the best parameters with the Cross Val accuracy of 99%.

In Figure 5, the searching range was reduced to $\log_2 C \in [-4, 8]$ and $\log_2 \gamma \in [-7, 3]$ and the fine grid-search was conducted with an exponential growing sequence of C and γ ($C = 2^{-4}, 2^{-3}, \ldots, 2^{7}, 2^{8}; \gamma = 2^{-7}, 2^{-6}, \ldots, 2^{4}, 2^{3}$), the best parameters were obtained with the Cross Val accuracy of 99%. To find the best parameters of C and γ, several grid-searches were executed, and several high Cross Val accuracies were obtained.

Table 4 shows the best parameters with their Cross Val accuracies, their classification accuracy (ACC), precision (PPV), true positive rate (TPR), and false positive rate

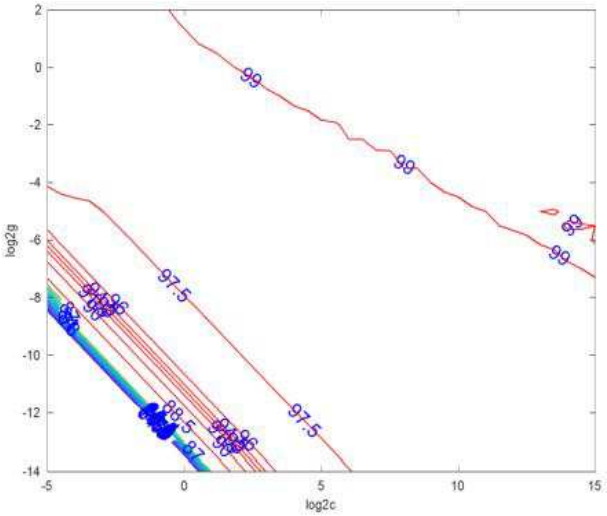

Figure 4: The coarse grid-search on $\log_2 C \in [-5, 15]$ and $\log_2 \gamma \in [-14, 2]$ for the KDD'99 dataset

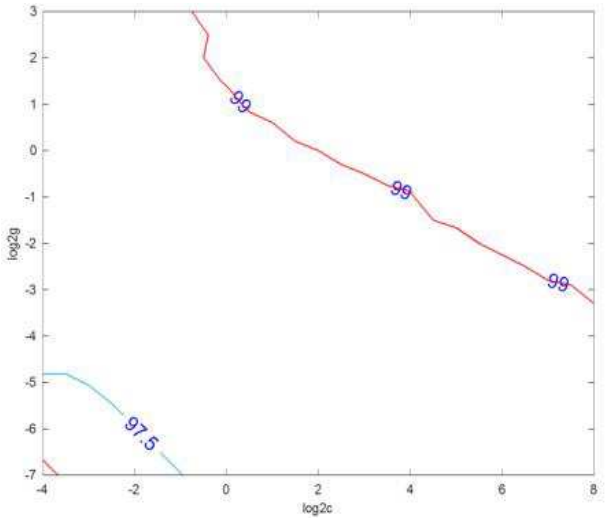

Figure 5: The fine grid-search on $\log_2 C \in [-4, 8]$ and $\log_2 \gamma \in [-7, 3]$ for the KDD'99 dataset

(FPR) for the KDD'99 dataset.

By using grid search methods mentioned above, the C and gamma (γ) parameters of RBF Kernel were tuned to select the ones with high Cross Val accuracy. As can be seen from the Table 5, the row with gamma (γ) =8, C= 1 is selected with Cross Val = 99.025%.

Figure 6 and Figure 7 displays the coarse grid-search and fine grid-search for NSL-KDD dataset.

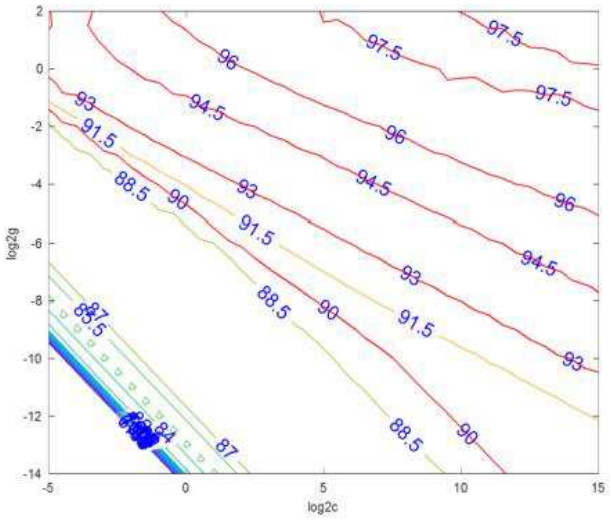

Figure 6: The coarse grid-search on $\log_2 C \in [-5, 15]$ and $\log_2 \gamma \in [-14, 2]$ for the NSL-KDD dataset

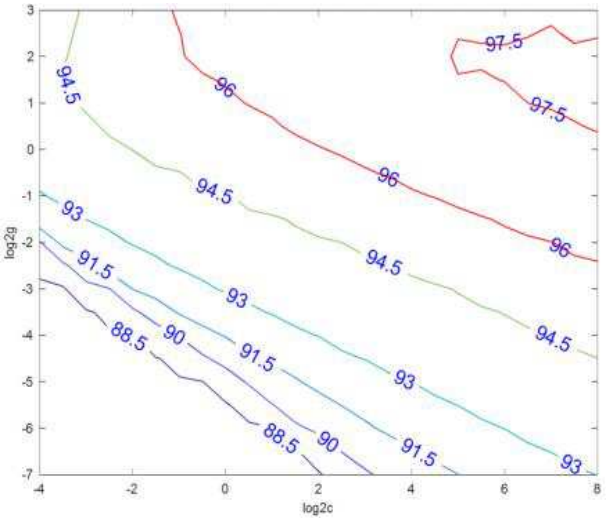

Figure 7: The fine grid-search on $\log_2 C \in [-4, 8]$ and $\log_2 \gamma \in [-7, 3]$ for the NSL-KDD dataset

For NSL-KDD dataset, the coarse grid-search was conducted in the range of $\log_2 C \in [-5, 15]$ and $\log_2 \gamma \in [-14, 2]$ as shown in Figure 6, the best parameters of C and (γ) were obtained with the Cross Val accuracy of 97.5%. Figure 7 shows the fine grid-search in the range of $\log_2 C \in [-4, 8]$ and $\log_2 \gamma \in [-7, 3]$, the best parameters were obtained with the Cross Val accuracy of 97.5%. To find the best parameters of C and (γ), we have conducted several grid-searches, and several high Cross Val accuracies were obtained. Table 5 shows the best parameters with their best Cross Val, their best classification ACC, PPV, TPR, and FPR for the NSL-KDD dataset.

As explained in Table 5, the same grid search method was performed to tune the RBF Kernel parameters, and the one with high Cross Val accuracy is picked, and as can be seen from Table 5, the row with γ=1 and C=2 is the one with the highest Cross Val = 96.6246%.

5.4 Experimental Result and Discussion

The KDD'99 Cup and NSL-KDD datasets are adopted during the experiments to evaluate the performance of the proposed method, and the performance comparison of different intrusion detection methods using KDD'99 and NSL-KDD dataset is shown in Table 6 and Table 7.

As can be seen from Table 6 and Table 7, through comparing to other research methods, the proposed approach has improved with good performance of PPV, TPR, and FPR.

1) With KDD'99 Cup dataset: The proposed method (SVM-IG) has an accuracy, which is smaller than that of Deep learning [19] and GFS-FSVM [15], but greater than the ones for RST-FSVM [16], GA-IDS [13] and SVM [9]. As can be seen from Table 6, Deep learning [19] and GFS-FSVM [15] have good ACC comparing to the proposed method, but the proposed method has good FPR compared to that of GFS-FSVM [15] and has good TPR and PPV compared to that of Deep learning [19].

   The Precision of the proposed method is good compared to other approaches except for SSA [26], but TPR of SSA [26] is slightly small compared to that of the proposed method. In addition to that, the TPR of the proposed method is good compared to other algorithms, except for GA-IDS [13], which has a slightly high FPR and small ACC and PPV compared to that of the proposed method. As can be seen from Table 6, the FPR of the proposed method is also better than the other approaches, except for SSA [26], which has a small TPR compared to the proposed method.

2) With NSL-KDD dataset: As can be seen from Table 7, the accuracy of the proposed method (SVM-IG),ACC=0.8650 is higher than that of SVM [20] and NNRw [3], but smaller than SVC [14], Naïve Bayes [7], and Deep learning [19]. On the other hand, SVC [14] and Naïve Bayes [7] have high FPR compared to that of the proposed method. As mentioned above, Deep learning [19] has good accuracy, but its PPV and TPR are smaller than that of the proposed method.

   The precision of the proposed method is better than that of SVM [19, 20] and Deep learning [19] but

Table 4: Performance of the proposed method using the Grid Search method for the KDD'99 dataset

Parameter Selection		ACC	PPV	TPR	FPR	CrossVal(%)
γ=1	C=32	0.9506	0.9647	0.9460	0.0247	98.925
γ=2	C=8	0.9507	0.9605	0.9464	0.0260	98.9
γ=2	C=16	0.9509	0.9515	0.9466	0.0253	98.95
γ=8	C=1	0.9523	0.9675	0.9489	0.0298	99.025
γ=8	C=2	0.9515	0.9624	0.9483	0.0314	98.925
γ=8	C=4	0.9510	0.9688	0.9478	0.0317	99

Table 5: Performance of the proposed method using the Grid Search method for the NSL-KDD dataset

Parameter Selection		ACC	PPV	TPR	FPR	CrossVal(%)
γ=0.25	C=32	0.8545	0.8878	0.7623	0.0657	96.0384
γ=0.5	C=4	0.8674	0.8813	0.7691	0.0651	96.5267
γ=0.5	C=8	0.8646	0.8877	0.7626	0.0655	96.6056
γ=1	C=1	0.8629	0.8781	0.7620	0.0702	96.3798
γ=1	C=2	0.8676	0.8878	0.7665	0.0646	96.6246
γ=4	C=32	0.8622	0.8757	0.7690	0.0664	96.6021

Table 6: Performance comparison of different intrusion detection methods using KDD'99 Cup

Methods	KDD'99 Cup			
	ACC	PPV	TPR	FPR
RST-FSVM [16]	0.9000	-	0.8576	0.1424
GA-IDS [13]	0.9004	0.9280	0.9500	0.3046
SVM [9]	0.9198	0.7400	0.8200	0.0391
DWIDM-CM SSC [27]	-	-	0.8989	0.0800
SSA [26]	-	0.9863	0.8902	0.0138
Deep Learning [19]	0.9711	0.9443	0.9277	-
GFS-FSVM [15]	0.9857	-	-	$\geq$0.0400
Proposed method (SVM-IG)	0.9523	0.9675	0.9489	0.0298

Table 7: Performance comparison of different intrusion detection methods using NSL-KDD

Methods	NSL-KDD			
	ACC	PPV	TPR	FPR
Naive Bayes [7]	0.9010	0.8900	0.9360	0.1340
SVC [14]	0.8970	-	0.9340	0.1400
SVM [20]	0.8350	0.7400	0.8200	0.1500
SVM [19]	0.8832	0.6470	0.7080	-
Deep Learning [19]	0.09099	0.8195	0.7748	-
NNRw [3]	0.8412	-	-	-
Human Immune System [12]	-	-	0.9860	0.0800
Proposed method (SVM-IG)	0.8676	0.8878	0.7665	0.0646

smaller than that of Naïve Bayes [7], which has a large FPR compared to that of the proposed method. In [7, 12, 14, 20], their TPR are better than that of the proposed method, but not good in terms of FPR compared to the proposed method.

Moreover, the experimental results prove that the proposed method can increase the training speed and shorten the training time cost with the elapsed time of 132.646993s for the KDD'99 dataset and 7.897525s for the NSL-KDD dataset, which is good compared to that of [24].

6 Conclusions

Over the last decades, many artificial intelligence algorithms have been applied to improve the performance of intrusion detection system (IDS). Among these algorithms, SVM is one of the most widely used and has a relatively high performance, and the performance of IDS is highly dependent on the quality of training data. In this paper, we proposed the intrusion detection method based on SVM and IG to detect cyber-attacks in MCC. The SVM classifier is used for binary classification to analyze and classify data in either normal or abnormal behavior, and the IG is used to select the most relevant features in KDD'99 and NSL-KDD dataset. Through the experimental results, we have shown that the proposed method has

good scalability and high training speed, and can detect malicious attacks with high accuracy, high detection rate, and low false positive rate.

For the future research, we will implement this method using multi-class classification and evaluate the performance on a real time basis.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 61862041), the Research Project in Universities of Education Department of Gansu Province (2017B-16, 2018A-187). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

References

[1] M. H. Aghdam and P. Kabiri, "Feature selection for intrusion detection system using ant colony optimization," *International Journal of Network Security*, vol. 18, no. 3, pp. 420–432, 2016.

[2] S. Aljawarneh, M. Aldwairi, and M. B. Yassein, "Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model," *Journal of Computational Science*, vol. 25, pp. 152–160, 2018.

[3] R. A. R. Ashfaq, X. Z. Wang, and J. Z. Huang, "Fuzziness based semi-supervised learning approach for intrusion detection system," *Information Sciences*, vol. 378, pp. 484–497, 2017.

[4] T. Bhatia and A. K. Verma, "Data security in mobile cloud computing paradigm: a survey, taxonomy and open research issues," *Journal of Supercomputing*, vol. 73, no. 6, pp. 1–74, 2017.

[5] A. N. Cahyo, R. Hidayat, and D. Adhipta, "Performance comparison of intrusion detection system based anomaly detection using artificial neural network and support vector machine," in *in Advances of Science and Technology for Society: Proceedings of the International Conference on Science and Technology*, vol. 1755, pp. 070011, 2016.

[6] L. Dali, K. Mivule, and H. El-Sayed, "A heuristic attack detection approach using the east weighted attributes for cyber security data," in *IEEE in Intelligent Systems Conference (IntelliSys'17)*, pp. 1067–1073, 2017.

[7] H. D. Deshmukh, T. Ghorpade, and P. Padiya, "Intrusion detection system by improved preprocessing methods and naïve bayes classifier using nsl-kdd 99 dataset," in *International Conference on Electronics and Communication Systems (ICECS'14)*, pp. 1–7, 2014.

[8] R. H. Dong, D. F. Wu, and Q. Y. Zhang, "Mutual information-based intrusion detection model for industrial internet," *International Journal of Network Security*, vol. 20, no. 1, pp. 131–140, 2018.

[9] A. S. Eesa, Z. Orman, and A. M. A. Brifcani, "A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems," *Expert Systems with Applications*, vol. 42, no. 5, pp. 2670–2679, 2015.

[10] K. Gai, M. Qiu, and L. Tao, "Intrusion detection techniques for mobile cloud computing in heterogeneous 5G," *Security and Communication Networks*, vol. 9, no. 16, pp. 3049–3058, 2016.

[11] Y. Hamid, V. R. Balasaraswathi, L. Journaux, and M. Sugumaran, "Benchmark datasets for network intrusion detection: A review," *International Journal of Network Security*, vol. 20, no. 4, 2018.

[12] H. Hammami, H. Brahmi, and S. B.Yahia, "Security insurance of cloud computing services through cross roads of human-immune and intrusion-detection systems," in *International Conference on Information Networking (ICOIN'18)*, pp. 174–181, 2018.

[13] M. S. Hoque, M. Mukit, and M. Bikas, "An implementation of intrusion detection system using genetic algorithm," *International Journal of Network Security & Its Applications*, vol. 4, no. 2, pp. 109–120, 2012.

[14] E. D. L. Hoz, A. Ortiz, and J. Ortega, "Network anomaly classification by support vector classi?ers ensemble and non-linear projection techniques," in *International Conference on Hybrid Artificial Intelligence Systems*, pp. 103–111, Sep. 2013.

[15] A. Kannan, G. Q. Maguire, and A. Sharma, "Genetic algorithm based feature selection algorithm for effective intrusion detection in cloud networks," in *IEEE 12th International Conference on Data Mining Workshops (ICDMW'12)*, 2012. (`https://ieeexplore.ieee.org/document/6406470`)

[16] L. Li and K. N. Zhao, "A new intrusion detection system based on rough set theory and fuzzy support vector machine," in *The 3rd International Workshop on Intelligent Systems and Applications (ISA'11)*, 2011. (`https://ieeexplore.ieee.org/document/5873410`)

[17] Q. I. M. Yu, M. Liu, and F. U. Y. Ming, "Research on network intrusion detection using support vector machines based on principal component analysis," *Netinfo Security (in Chinese)*, vol. 2, pp. 15–18, 2015.

[18] M. B. Mollah, M. A. K. Azad, and A. Vasilakos, "Security and privacy challenges in mobile cloud computing: Survey and way ahead," *Journal of Network & Computer Applications*, vol. 84, pp. 34–54, 2018.

[19] K. K. Nguyen, D. T. Hoang, and D. Niyato, "Cyberattack detection in mobile cloud computing: A deep learning approach," in *IEEE Wireless Communications and Networking Conference (WCNC'18)*, pp. 1–6, 2018.

[20] M. S. Pervez and D. M. Farid, "Feature selection and intrusion classification in NSL-KDD cup 99 dataset employing SVMs," in *The 8th International Conference on Software, Knowledge, Information Management and Applications (SKIMA'14)*, pp. 1–6, 2014.

[21] Q. S. Qassim, A. M. Zin, and M. J. A. Aziz, "Anomalies classification approach for network-based intrusion detection system," *International Journal of Network Security*, vol. 18, no. 6, pp. 1159–1172, 2016.

[22] S. Rezaei, M. Ali Doostari, and M. Bayat, "A lightweight and efficient data sharing scheme for cloud computing," *International Journal of Electronics and Information Engineering*, vol. 9, no. 2, pp. 115–131, 2018.

[23] I. S. Thaseen and C. A. Kumar, "Intrusion detection model using fusion of chi-square feature selection and multi class SVM," *Journal of King Saud University-Computer and Information Sciences*, vol. 29, no. 4, pp. 462–472, 2017.

[24] H. W. Wang, J. Gu, and S. S. Wang, "An effective intrusion detection framework based on SVM with feature augmentation," *Knowledge-Based Systems*, vol. 136, pp. 130–139, 2017.

[25] Z. Wang, "Using neural networks in intrusion detection system for cloud computing," *California State University San Marcos*, vol. 24, no. 3, pp. 579–588, 2014.

[26] Y. Yuan, G. Kaklamanos, and D. Hogrefe, "A novel semi-supervised adaboost technique for network anomaly detection," in *Proceedings of the 19th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, pp. 111–114, 2016.

[27] J. Zhang and Y. Z. Li, "Dynamic weighted intrusion detection method based on cloud model and semi-supervised clustering," *Journal of Kunming University of Science & Technology*, vol. 2013, no. 4, 2013.

[28] X. Zhang, P. Zhu, and J. Tian, "An effective semi-supervised model for intrusion detection using feature selection based LapSVM," in *International Conference on Computer, Information and Telecommunication Systems (CITS'17)*, pp. 283–286, 2017.

[29] J. Zhao and Y. Zhu, "Research on intrusion detection method based on som neural network in cloud environment," *Computer Science and Application*, vol. 6, no. 8, pp. 505–513, 2016.

Biography

Mugabo Emmanuel. He is currently pursuing his master's degree at Lanzhou University of Technology. He graduated with a bachelor degree in Electronic Science and Communication from University of Dar-es-salaam (Tanzania) in 2014. His main research focuses on the network and information security.

Zhang Qiu-yu. Researcher/Ph.D. supervisor, graduated from Gansu University of Technology in 1986, and then worked at school of computer and communication in Lanzhou University of Technology. He is the vice dean of Gansu manufacturing information engineering research center, a CCF senior member, a member of IEEE and ACM. His research interests include network and information security, information hiding and steganalysis, multimedia communication technology.