International Journal of Security and Its Applications
Vol.8, No.5 (2014), pp.391-402
http://dx.doi.org/10.14257/ijsia.2014.8.5.34

A Novel Method for Detection of Internet Worm Malcodes using Principal Component Analysis and Multiclass Support Vector Machine

S.Divya [1] and Dr.G.Padmavathi [2]

[1]*Research Scholar, Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, University, Coimbatore, India.*
Email id: divya.phd.research@gmail.com
[2]*Professor and Head, Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, University, Coimbatore, India.*
Email id: ganapathi.padmavathi@gmail.com

Abstract

Internet worms are malware programs that imitate themselves and spread around the network. Internet worm, a wide spreading malcode exploits vulnerability in the operating system, hard disk, software and web browsers. This paper analyzes and classifies the Internet worm, depending on the training signatures. This work presents the Internet worm detection mechanism, using Principal Component Analysis (PCA) and Support Vector Machine (SVM). A Selective sampling technique is applied to maximize the performance of the classifier and to reduce misleading data instances. The results obtained show improved memory utilization, detection time and detection accuracy for Internet worms.

Keyword *- Malcode, Selective sampling, Multiclass SVM and PCA*

1. Introduction

Malicious software is due to malware or malicious code (Malcode). Malware is known as a code or software that is particularly designed to damage, disrupt, steal, or some other imposing of data or illegal action on data, hosts, or networks. These types of Internet worms often attack the computer through the development of vulnerabilities arising from low-level memory faults such as stack overflow, format string vulnerability, integer overflow, double free, heap overflow and return-to-libc [4]. Therefore, to protect the computer system from the Internet attacks, the scanning process is done for all available network resources using local OS services and the Internet for vulnerable hosts in the network [5]. Many real-world worms have caused notable damage to the network and computers. These worms include "Code-Red" worm in 2001,"Slammer" worm in 2003 [14],"Witty"/ "Sasser" worms in 2004, Storm worm in 2007 [1] and StuxNet worms in 2010-2012. The SQL Slammer worm infected more than 90% of hosts on the Internet within 10 minutes and Storm worm

ISSN: 1738-9976 IJSIA

infected millions of computers. Within the period of five years, 4,00,000 computers got infected by the Blaster worm.

Presently many software packages are used to detect and remove the malware. Usually antivirus software checks each file in the system looking for known signs (signatures) which uniquely identify an instance of known malcode [10]. Various approaches exist to detect or eliminate the malware code using Intrusion detection systems. These types of intrusion detection used in host level are called host-based intrusion detection systems (HIDS) [7] [13]. Though these types of HIDS have some limitations in their detection approaches, there is a need for effective detection systems. The objective of the work is to analyze the malcode and classify those identified under the labeled classes based on vulnerabilities to achieve accuracy. Given a training set of signatures, a classifier is trained to identify and classify the unknown executable as being malicious. In this paper, the principles of Principal Component Analysis (PCA) and Multiclass Support Vector Machines (MSVM) are applied to analyze and detect unknown worms based on their character code and examine using the selective sampling approach to improve the detection performance in terms of precision, recall, accuracy, time consumption and memory utilization.

The paper is organized as follows: Section 2 discusses the related works of some authors in Internet worm detection. Section 3 describes the proposed method for Internet worm detection and classification. Section 4 discusses the experimental results and Section 5 concludes the paper.

2. Related Works

Internet worms propagating in the network cause high damage and they are primarily detected using signatures. Various approaches have been proposed by different authors for monitoring and detecting Internet worms based on their characteristics. Some of them are discussed below.

Gil Tahan et al. [5] introduced form-based analysis methodology which analyzes common segments to detect the malware files. The techniques involved are the feature extraction, selection and finally the classifier is used to classify unknown malware. This detection method enabled low false positive rate.

Nir Nissim et al. [7] applied Support Vector Machines (SVM) for classifying the unknown Internet worms, which combines active learning selective sampling for its effective performance. Different kernel functions are applied with the SVM classifier to detect unknown worms based on their behavior in the host environment. Mean detection accuracy is achieved by the technique, sustaining the false positive rate at low level.

Qian Wang et al. [9] proposed a method based on maximum likelihood and linear regression estimation methods. Statistical estimation techniques are applied to monitor the Internet worm tomography. Darknet scans the infected hosts with both worm propagation and statistical model. Passive unwanted traffic is analyzed by Darknet. Estimation technique is enabled for destination detection.

Robert Moskovitch et al. [10] detect the unknown worms using machine learning methods based on their behaviors on the host. Worm malcodes are detected and feature set is reduced using feature selection methods. Classification accuracy is achieved by minimizing the false positive rate. The learning algorithms are applied for the feature subset reduction.

Wei Yu et al. [14] analyzed the C-Worm characteristics and the traffic comparison is performed using spectrum-based detection scheme. Detection scheme is developed using frequency domain analysis technique to trace Internet traffic.

Wei Yu et al. [12] used game-theoretic formulation to handle worm propagation and use defender mechanism to show their interaction. Static and dynamic self-disciplinary worms are classified and their propagation damage is detected by the combination of trace-back, threshold-based and spectrum-based schemes.

Detection of Internet worms using different techniques and metrics is listed below. Table 1 summarizes the significant literatures reviewed for Internet Worm detection.

Table 1. Literature Review on Internet Worm Detection

Year	Author	Technique(s) used	Metrics Used	Observations
2008	Robert Mosvitch et al.	Bayesian Networks	True Positive Rate, False Positive Rate, Total Accuracy	Mean detection accuracy is achieved with low false positive rate
2010	Wei Yu et al.	Game Theory	Infection Rate, False Positive Rate	Worms are classified based on growth rate propagation
2011	Wei Yu et al.	Power Spectral Density Distribution	Infection Ratio, Detection, Time, Detection Ratio	Both time and frequency domain are used for analyzing and reducing effective countermeasures
2011	Qian Wang et al.	Statistical Estimation	Error Rate Detection	Estimation methods perform better for identifying worm infection sequence
2012	Gil Tahan et al.	Boosted Decision Tree	False Positive Rate, Accuracy, AUC (Area Under ROC) Curve	Malwares are detected at the segment level using n-grams
2012	Nir Nissim et al.	Support Vector Machine	True Positive Rate, False Positive Rate, Total Accuracy	Reduces misleading instances using selective sampling and different kernels with SVM used for classifying unknown worms on hosts

Various techniques have been used to detect the Internet worms and different parameters have been applied to achieve the accuracy of detection. Though acceptable level of accuracy is achieved, there may be certain limitations on memory and system affect rates. The Internet worms are to be detected before affecting the network and the proposed approach overcomes the existing limitations.

3. Proposed Methodology

In this proposed methodology, Principal Component Analysis (PCA) and Multiclass Support Vector Machine (SVM) [2, 3, 6, 7, 11] are used to analyze and classify the Malcodes by comparing with the signatures in the labeled dataset. Using Multiclass SVM classifier, the identified malcodes are categorized based on their vulnerabilities. The detailed description of the proposed techniques is explained below.

3.1. Principal Component Analysis

Let X indicate an N x P data where N is the number of data samples, which can be regarded as N realizations of a P-dimensional random vector, which has been normalized to zero-mean and unit variance. PCA is a linear transformation R^P from, to an M-dimensional vector space, where$M \leq P$. The optimal linear mapping in the least mean square sense is the one formed by the eigenvectors of the correlation matrix of S_x.X, where $S_x = (1/(N-1))X^T X$ .Let Z denote the N x M transformed data matrix. The PCA transforms X to Z by the following equation:

$$Z = XV_M \qquad (1)$$

Where V_M, the P x M weight matrix, consists of eigenvectors corresponding to the first largest eigenvalues of the correlation matrix S_x, or V_M corresponds to the first M column vectors of matrix, which is obtained through singular value decomposition (SVD) of a scaled matrix $T = (1\sqrt{N-1})X$, i.e., $T = UDV^T$.Here, both U and V are unitary matrices, and D is a P x P diagonal matrix with nonnegative diagonal elements d_i in decreasing order. Note that the correlation matrix of Z is diagonal, which is given by $S_z = (1/(N-1))Z^T Z = \text{diag}\{d_1^2, d_2^2, \dots\}$; i.e., columns of Z are mutually uncorrelated.

Let x denote a row of X (one of the samples, or one NAV), $z_j = xv_j$ is referred to as the jth principal component (or the th PC score). The column vectors v_j of V is called the weight vector of z_j or the jth feature vector. The minimum mean-square error due to dimension reduction is $\sum_i^P = M + 1\ {d_i}^2$

Evaluating the statistical significance of a principal component (PC) is a critical part of choosing a proper dimension for PCA to capture the most salient features. As they are mutually uncorrelated, each coefficient is tested individually using only one random variable statistics to determine whether it is relevant or random.

3.2. Multi-class Support Vector Machine

The earliest used implementation of SVM multiclass classification is probably the method named one-against-all. It constructs ‘k’ SVM models where ‘k’ represents the number of classes. The ith value in SVM trained through all of the examples in the i^{th} class with positive labels, and the remaining with negative labels. Thus given one training data $(x_1, y_1), \dots, (x_l, y_l)$, where $x_i \epsilon R^n, i = 1, \dots, l$ and $y_i \in \{1, \dots, k\}$ is the class of x_i , the ith SVM solves the following problem:

$$\min_{w^i,b^i,\xi^i} \frac{1}{2}(w^i)^T w^i + C\sum_{j=1}^{l}\xi_j^i (w^i)^T \quad (2)$$

$$(w^i)^T \emptyset(x_j) + b^i \geq 1 - \xi_j^i, \text{if } y_j = i \quad (3)$$

$$(w^i)^T \emptyset(x_j) + b^i < -1 + \xi_j^i, \text{if } y_j \neq i \quad (4)$$

$$\xi_j^i \geq 0, j = 1, .., l \quad (5)$$

Here, training data x_i will be mapped with the high dimensional space by the function $\emptyset$ and C is the penalty parameter.

Minimizing $(1/2)(w^i)^T w^i$ means maximizing $2/\|w^i\|$, and the margin between two data groups. When grouped data are not separable linearly, there is a penalty term $C\sum_{j=1}^{l}\xi_j^i$ which can reduce the number of training errors. The fundamental conception behind SVM is to search for a balance between the regularization term $(1/2)(w^i)^T w^i$ and the training errors.

After solving (5), there are k decision functions

$$(w^1)^T \emptyset(x) + b^1$$
$$\vdots$$
$$(w^k)^T \emptyset(x) + b^k$$

Here x is in the class which has the largest value of the decision function

$$\text{class of } x = \arg\max_{i=1,\ldots,k}((w^i)^T \emptyset(x) + b^i) \quad (6)$$

Practically, solve the dual problem of (5) whose number of variables is the same as the number of data in (5). Hence l − variable quadratic programming problems solved. Table 2 presents the algorithm for Multiclass SVM.

Table 2. Algorithm of Multiclass SVM

Input $\{(\bar{x}_1, y_1), \ldots, (\bar{x}_m, y_m)\}$

Intialize: $\bar{T}_1 = \bar{0}, \ldots, \bar{T}_m = \bar{0}$

Loop:

Step 1: Choose an example p.

Step 2: Calculate the constants for the reduced problem

- $A_p = K(\bar{x}_p, \bar{x}_p)$
- $\bar{B}_p = \sum_{i \neq p} K(\bar{x}_i, \bar{x}_p)\bar{T}_i - \beta \bar{1}_{yp}$

Step 3: Set $\bar{T}_p$ to be the solution of the reduced problem

$$\min_{r_p} Q(\bar{T}_p) = \frac{1}{2} A_p(\bar{T}_p.\bar{T}_p) + \bar{B}_p.\bar{T}_p$$

Subject to : $\bar{T}_p \leq \bar{1}_{yp}$ and $\bar{T}_p.\bar{1} = 0$

Output: $H(\bar{x}) = \arg\max_{r=1}^{k}\{\sum_i T_{i,r} K(\bar{x}, \bar{x}_i)\}$

The above proposed algorithm in table 2 classifies the Internet worms that are monitored and analyzed by PCA with the trained signatures. They are further classified accurately by the selective sampling technique labeled based on their input.

3.3. Selective Sampling

Selective Sampling is a branch of active learning mainly used to minimize the learning process and improve classifier accuracy by adding up supplementary labels to the dataset. To perform the accurate predictions with the input, the dataset is divided and labeled. Here, these labeled datasets under the sampling technique intends to decrease the labeling effort. The signatures are sampled under the different labels named C, C++ and Java.

Table 3 below provides the pseudocode of the proposed method, which describes the overall process of monitoring and classification of Internet worms.

Table 3. Pseudo Code of Proposed Algorithm

```
Input: P->Packet
Output: CL-> Class of vulnerabilities
Begin
For each packet P transfer
    Check each line of code C with signatures S in labeled dataset
     If (C ε S)
     Marked as Malicious code
     Add malcode lines to temporary buffer
    Else
    Benign
    End
Until end of packet
   Read malcode lines from temporary buffer
   T = total number of malcode lines from temporary buffer
  Calculate Euclidean distance of first infected lines and last infected lines
  K=Sum of infected lines in packet / T
 Kernel point = K + constant
 Classify the malcode under different classes CL
End
```

In the above table, the proposed approach incremented by the threshold constant value for its better detection accuracy. The detected worms are further classified into multiple classes based on vulnerability in the proposed method.

Figure 1 below provides the flowchart of the proposed Principal Component Analysis with the Multiclass Support Vector Machine for detecting and classifying the Internet worms.

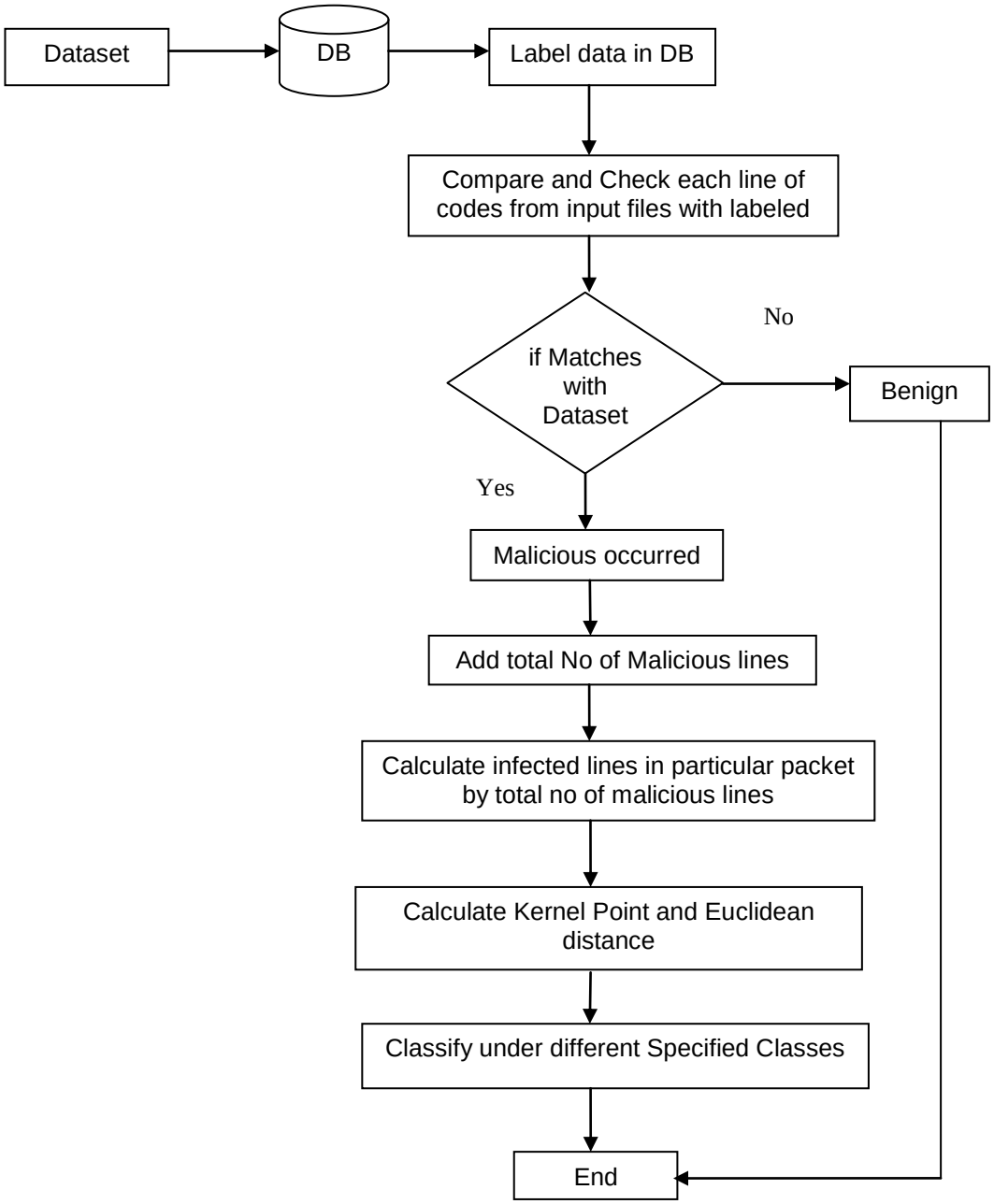

Figure 1. Flowchart of Proposed Methodology

The proposed approach at the host level is used for preventing the network from Internet worms attack entry. Figure 1 shows the proposed approach that uses the PCA and Multiclass SVM for analyzing and classifying the malcodes with the trained labeled dataset. Multiclass SVM in the proposed work classifies the identified malcodes under different classes based on their vulnerabilities and stores the identified anomalous.

4. Experimentaion and Results

The proposed method is evaluated using various parameters such as memory utilization, time utilization, precision value, recall value and accuracy. To provide high stability and performance in detection accuracy and Classification of Internet worms, memory usage and time consumption is monitored.

Memory Utilization

The usage of CPU of the system is calculated for finding memory utilization.

Memory Utilization = CPU memory consumption after process completion – CPU memory consumption before process beginning

Time Utilization

The time consumption is the period between the end of the detection system and the start of its execution process.

Time Utilization = Finishing time – Starting time of processing

Precision value

Precision refers to the retrieved document. This is calculated by the total number of relevant documents divided by the total number of resultant documents.

$$Precision\,Value = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Recall value

Recall value is referred to as the relevant documents that are related to the request search.

$$Recall\,Value = \frac{True\ Positive}{False\ Positive + False\ Negative}$$

Accuracy

Accuracy provides the required related documents/measures used for classification.

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + \ False\ Positive + True\ Negative + False\ Negative}$$

The proposed method is implemented using Java. The real traces of the dataset are taken from the Internet through web. The dataset contains 1008 signatures for worm. It is labeled with harmful code, type and programming language.

Table 4. Parameters for proposed methodology

Parameters	Existing SVM	Proposed PCA with Multiclass SVM	% of Improvement
Memory Utilization (MB)	30MB	26MB	13.33%
Time Utilization (ms)	11ms	9ms	18.18%
Precision Value (%)	36.72%	47.45 %	22.61%
Recall Value (%)	52.57%	60.87 %	13.63%
Accuracy (%)	48.72 %	56.37 %	13.57%

Table 4 provides the comparison of parameters between existing SVM and proposed PCA with SVM. The given parameters are Memory utilization in (MB), Time Utilization in (ms), Detection Range, Precision Value in (%), and Recall Value in (%) and Accuracy in (%).

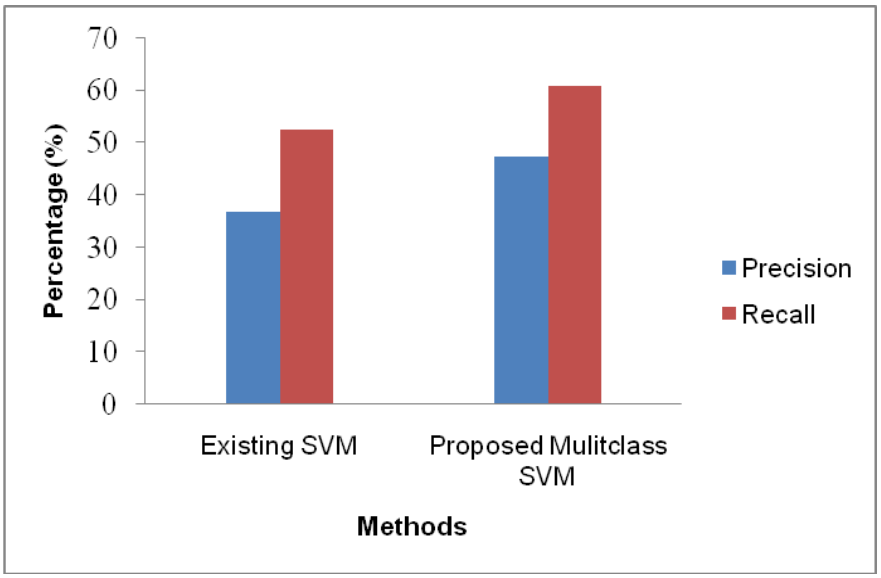

Figure 2. Comparison of Precision and Recall

The above Figure 2 illustrates the comparison between precision and recall for existing SVM and proposed PCA with SVM. From figure 2, it can be clearly observed that the proposed method of PCA with SVM give better results compared to the existing SVM.

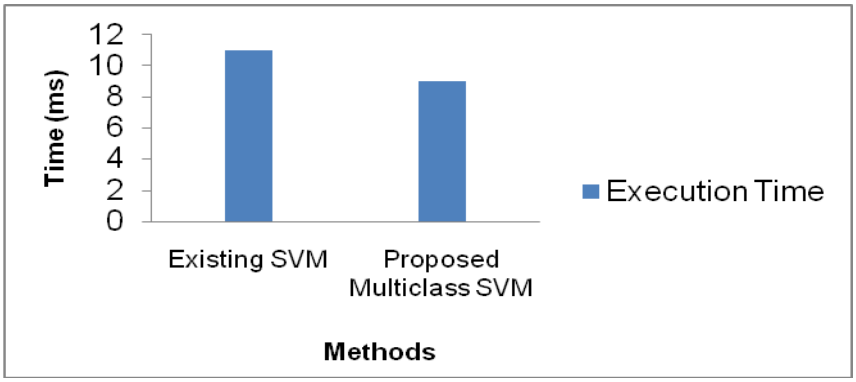

Figure 3. Execution Time Comparison

Figure 3 gives the execution time comparison between existing SVM and proposed PCA with SVM. From Figure 3, it is clearly observed that the proposed PCA with SVM provides improved 6ms result, that is less than existing SVM.

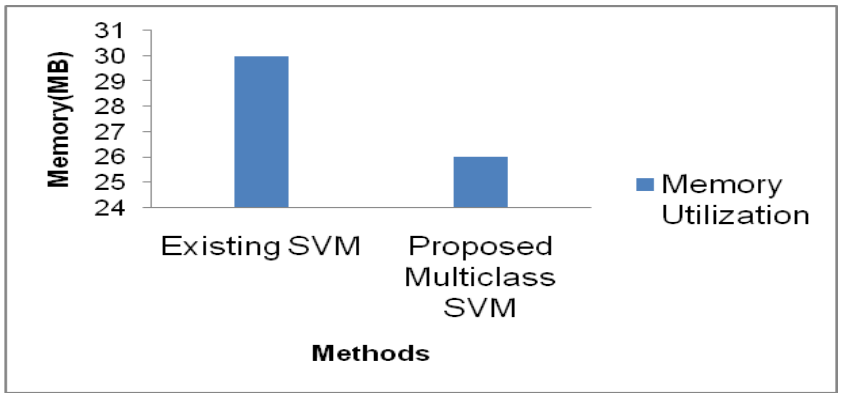

Figure 4. Comparison of Memory utilization

Figure 4 gives the comparison, that existing SVM used memory allocation of 30MB and proposed PCA with SVM, 26MB memory. From Figure 4, it is clearly observed that the proposed method has less memory space than the existing SVM.

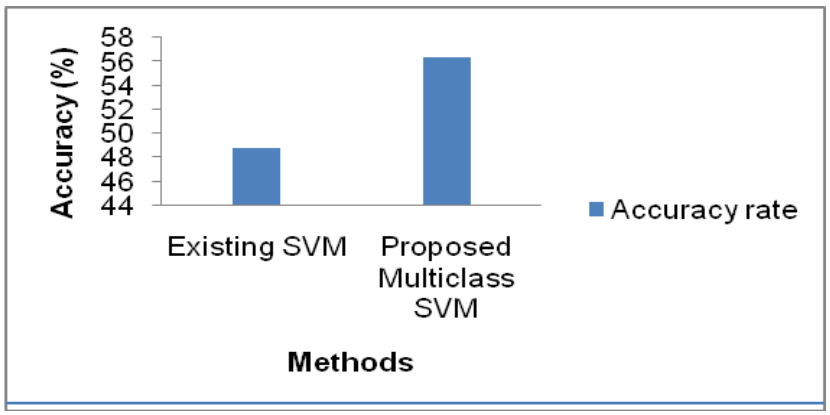

Figure 5. Overall Accuracy

Overall accuracy is shown in figure 5 for existing system of SVM and proposed method of PCA with SVM. In this, accuracy attained by existing SVM is 48.72% and proposed PCA with SVM is 56.37%.

5. Conclusion

This paper proposed the method for worm detection in Malcode. Internet Worms are most vulnerable in affecting the network operating system, software, web browser and hard disk in the computers. Some of the challenges address traffic payload analysis. The proposed method prevents the infection by analyzing the files in the client vulnerable host, before entering the network transfer. In this paper, the novel algorithm PCA with Multiclass SVM is implemented for analyzing the Internet worms affecting the network and is classified based on the operating system, web browser and hard disk vulnerabilities. Compared to existing method of SVM, proposed method shows better resultsin terms of memory utilization, time utilization, detection range, precision value, recall value and overall accuracy. The results shown in section 4 provide better results in classifying the identified anomalies under the specified classes.

References

[1] C. Chen, Z. Chen and Y. Li, "Characterizing and defending against divide-conquer-scanning worms", Elsevier, Computer Networks, vol. 54, **(2010),** pp. 3210-3222.
[2] C.-W. Hsu and C.-J. Lin, "A Comparison of Methods for Multiclass Support Vector Machines", IEEE Transactions on Neural Networks, vol. 13, no. 2, **(2002),** pp. 415-425.
[3] F. Kuang, W. Xu and S. Zhang, "A novel hybrid KPCA and sVM with GA model for intrusion detection", ELSEVIER, Applied Soft Computing, vol. 18, **(2014),** pp. 178-184.
[4] G. Qijun, F. Chistopher and N. Rizwan "A study of self-propagating mal-packets in sensor networks: Attacks and defenses", ELSEVIER, Computers & Security, vol. 30, nop. 1, **(2011)**, pp. 13-27.
[5] G. Tahan, L. Rokach and Y. Shahar, "Mal-ID: Automatic Malware Detection Using Common Segment Analysis and Meta-Features", Journal of Machine Learning Research, vol. 1, **(2012),** pp. 949-979.
[6] J. Yu, H. Lee, Y. Im, M.-S. Kim and D. Park, "Real-time Classification of Internet Application Traffic using a Hierarchical Multi-class SVM", KSII Transaction on Internet and Information Systems, vol. 4, no. 5, **(2010)** October, pp. 859-876.
[7] N. Nissim, R. Moskovitch, L. Rokach and Y. Elovici, "Detecting unknown computer worm activity via support vector machines and active learning", Springer, Pattern Analysis and Applications, vol. 15, no. 4, **(2012),** pp. 459-475.
[8] P. Li, M. Salour and X. Su, "A Survey of Internet Worm Detection and Containment", IEEE Communications Surveys, vol. 10, no. 1, **(2008),** pp. 20-35.
[9] Q. Wang, Z. Chen and C. Chen, "Darknet-Based Inference of Internet Worm Temporal Characteristics", IEEE Transactions on Information Forensics and Security, vol. 6, Issue. 4, **(2011),** pp. 1382-1393.
[10] R. Moskovitch, Y. Elovici and L. Rokach, "Detection of unknown computer worms based on behavioral classification of the host", ELSEVIER, Computational Statistics & Data Analysis, vol. 52, **(2008),** pp. 4544–4566.
[11] S. H. Hashem, "Efficiency of SVM and PCA to enhance Intrusion Detection System", Journal of Asian Scientific Research, vol. 3, no. 4, **(2013),** pp. 381-395.
[12] W. Yu, N. Zhang, X. Fu and W. Zhao, "Self-Disciplinary Worms and Countermeasures: Modeling and Analysis", IEEE Transactions on Parallel and Distributed Systems, vol. 21, no. 10, **(2010),** pp. 1501-1514.
[13] W, Yu and X, Wang, A. Champion, D. Xuan and D. Lee, "On detecting active worms with varying scan rate", Elsevier, Computer Communications, vol. 34, **(2011)**, pp. 1269-1282.

[14] W. Yu, X. Wang, P. Clayam, D. Xuan and W. Zhao, "Modeling and Detection of Camouflaging Worm", IEEE Transactions on Dependable and Secure Computing, vol. 8, no. 3, **(2011),** pp. 377-390.

Authors

S.Divya is pursuing her Ph.D. at Avinashilingam University for Women, Coimbatore. She has 2 years of teaching experience. Her areas of interest include Network and Communication Security. She has 3 publications in her research area.

Dr.G.Padmavathi is the Professor and Head, Department of Computer Science,Avinashilingam University, Coimbatore. Shae has 25 years of teaching experience and one year of industrial experience. Her areas of interest include Real Time Communication, Network Security and Cryptography. She has 100 Publications in her research area. Presently, she is guiding M.Phil researcher and Ph.D scholars. She has been profiled in various organizations for her academic contributions. She has been the Principal Investigator for four projects funded by UGC and DRDO and the Scientific Mentor for one project funded by DST. She is life member of many preferred organizations of CSI, ISTE,WSEAS, AACE and ACRS.

