Evaluating Performance of Intrusion Detection System using Support Vector Machines: Review

Leila Mohammadpour¹, Mehdi Hussain¹², Alihossein Aryanfar³, Vahid Maleki Raee¹ and Fahad Sattar⁴

¹Faculty of Computer Science and Information Technology, University of Malaya 50603 Kuala Lumpur, Malaysia
²School of Electrical Engineering and Computer Science, National University of Sciences and Technology, Islamabad Pakistan
³Faculty of Computer Science and Information Technology, University Putra Malaysia, Malaysia
⁴University of Management and Technology, Lahore Pakistan
le.vesa@siswa.um.edu.my, mehdi141@siswa.um.edu.my, a.aryanfar85@gmail.com, vahid_m@siswa.um.edu.my, fahad.sattar1@gmail.com

Abstract

The basic task in intrusion detection system is to classify network activities as normal or abnormal while minimizing misclassification. In literature, various machine learning and data mining techniques have been applied to Intrusion Detection Systems (IDSs) to protect the special computer systems, vulnerable traffics cyber-attacks for computer networks. In addition, Support Vector Machine (SVM) is applied as the classification techniques in literature. However, there is a lack of review for the IDS method using SVM as the classifier. The objective of this paper is to review the contemporary literature and to provide a critical evaluation of various techniques of intrusion detection using SVM as classifier. We analyze and identify the strengths and limitations of various SVM usages as classifier in IDS systems. This paper also highlights the usefulness of SVM in IDS system for network security environment with future direction.

Keywords: Intrusion Detection System; Support vector machine; intelligent IDS; Machine learning IDS

1. Introduction

Intrusion detection is one of the most essential facilities for security infrastructures in network environments, and it is widely used in detecting, identifying and tracking the intruders. The network vulnerable traffic has been excessively increasing in every year. Different computer networks vulnerabilities as malicious intrusion or attacks in computers or information systems becoming serious issues, because it violates the security parameters i.e., Confidentiality, Integrity and Availability. To best of our knowledge, the threats on computer networks or information security are still significant research issues. Though there are number of existing literatures to survey IDS [1,2,3,4,5,6]. In this paper, it gives a more contemporary image of the current IDS based literature which utilized SVM as the main part of classifier for improving the intrusion detection systems.

Many problems in the traditional IDS should be addressed, such as the low detection capability against the unknown network attack, high false alarm rate, and insufficient analysis capability and so on. Generally, intrusion detection is targeted as classification problem, to distinguish between the normal activities and the malicious activities [7]. The concerned problems of machine learning are how the systems automatically improve the performance with the increase of experience, which is consistent with that of the IDS [7].
Therefore, various machine learning methods are developed for intrusion detection, such as decision tree [8], genetic algorithm (GA) [9], neural network [10], principal component analysis (PCA) [11], fuzzy logic[12], K-nearest neighbor[13], rough set theory [14] and support vector machine [15].

SVM is an effective and popular one, against different available machine learning methods in IDA, because SVM strength as the distribution of different types of imbalanced attacks, where the learning sample size of the low-frequent attacks is too small as compared to the high-frequent attacks [7]. SVM as classifier has capability of good generalization against small sample size training or learning, which is frequently used in real world applications of classification [16]. Due to robustness and efficiency in the network action classification, SVM is widely used in IDS as a popular method [17]. Although many surveys have been carried out on IDS, to best of our knowledge no single study exists which focus the usage of SVM as classification in this field. In this paper we have critically reviewed the IDS, which applied SVM as individual or hybrid classification in intrusion detection.

The purpose of this paper is to review the literature and provide critical evaluation of IDS using SVM classifier. Rest of the paper is as follows; Section 2, describe the measurement tools which are used in IDS methodologies. Section 3, present literature review of the form of contemporary literature, Section 4 presents the critical analysis of different IDS with SVM techniques. Subsequently, Section 5 draws conclusion, and future work.

2. Measurement Tools

In this section, we introduce the most applied measurement tools which are used as the performance measurement in the IDS. There are some performance indicators for the intrusion detection system as follows: TP, FP, TN and FN, where TP represents that the normal behavior is correctly forecasted, FP indicates that the abnormal behavior is judged as normal, FN denotes that the normal behavior is wrongly thought as abnormal, and TN represents the abnormal behavior is correctly detected [18]. Where DR denotes the detection rate and FAR denotes the false alarm rate. They are important to evaluate the performance of the intrusion detection system.

\[
\text{(1) Detection rate: } DR = \frac{TN}{TN+FP} \\
\text{(2) False alarm rate: } FAR = \frac{FN}{FN+FP}
\]

3. Literature Review

The purpose of this literature review is to highlight the significance of IDS using SVM. We also outline ideas how these different IDS systems can be further improved.

Kuang Xu et al. [7], Proposed a method with the combination of KPCA, SVM and GA algorithms. Where KPCA reduce the eigenspace input space features and after that apply SVM to classify the features. Where SVM parameters are optimized by GA. However, the selection of parameters effects on the detection performance, when the difference between features are high, using the RBF kernel in the training phase, it produce a huge number of support vectors. Both of them are vital for achieving high performance of detection, intrusion and feature extraction. In this paper author use the dataset KDD CUP99 for training and testing. According to experimental results that the proposed model illustrates the faster convergence speed, Performs superior forecasting accuracy and better generalization. But this proposed is not useful for state of distributed situation, just useful for centralized. The suggested idea for improvement of this method is using some kernel methods combining more classification method such Neural networks (ANN) and optimization algorithm such as GA, QPSO as for pattern analysis and real-time intrusion detection.
Authors in [19], proposed an approach using data mining techniques. This approach used a combination of Change Detection Algorithm and a data mining. Although many IDS have been proposed, but high dimensionality data to proceed in network still is challenge. Changing data frequently is a challenge which affects the classification performance. Although, the data suffer from outliers are considered as the noise. However it could have a potentiality to carry the significant information. To capture the transmitted data from network in real time situation, JPCAP and WINPCAP tool is applied. Moreover, the result of experiments represent that this model can be used for the new feature mining in IDS. The limitation of this work can be mentioned that the attack with hardly seen cannot be recognized successfully. For suggestion it can be considered that the detection of exact source of attack also important to consider in the IDS.

Y. Li et al. in [20], introduced a approach for detecting normal or abnormal behavior established a high efficient and accurate classifiers with combining of SVM, as ant colony algorithm and clustering method. It also reduced the number of features. But, its computation is complex and time consuming for distinguishing normal and abnormal behavior and deal with the massive amount of data which contains redundant features. In this paper author use the dataset KDD 99. Furthermore, in this article, 10–fold cross validation is used for the accuracy validation. Unfortunately, hybrid method has low performance. When there is unbalanced situation, small training data is not suitable. Multiple classifications can be useful for complex program and adaptive.

According to the idea that the current and initial samples are kept during the entire training process, the paper has suggested, an enhanced learning of SVM. However, the memory of computer’s is not enough for a large training dataset. When data packets are captured from a stream of a network, it is not possible to get all information about the network in the first time, therefore, for having high learning accuracy a regular online learning is needed incrementing number of samples. Furthermore, the challenge of incremental learning is to determine how to deal with the new data sets to be added in that phase and what and how much information from the earlier training should be elected for training in the next learning phase. Thus, deal with rising data is the key of incremental learning. This paper uses KDD Cup 1999 data set. The article indicates that in comparing to general SVM the proposed method has better recognize rate and false alarm rate than usual classification for network intrusion detection in real-time. For acquire best result can change the strategy to work with multi-class [21].

A. H. Sung et al. in [22], used a method for feature deleting at each time employ to SVM and ANN, after that the features are ranked and for five different classes in DARPA intrusion data, efficient features are demonstrated. But, since the unimportant and/or pointless inputs cause a complex problem, slower and less accurate detection results. Author used data in their own experiments that initiatate from MIT's Lincoln Lab. This technique considered a benchmark for evaluating and creating an intrusion detection system by DARPA. Using the importance features gives the most significant performance as far as training time. Using 2 classifiers are time consuming and hard task to trigger them.

I. Ahmad et al. in [23], performance of intrusion detection is enhanced according to optimal feature subset election, which are gathered from GA and PCA. However, the primary problem is their performance, which can be improved by raising the recognition rates and diminishing false positives. The authors of this paper, for experimentation used KDD cup dataset. This method indicates that the suggested method improved SVM performance in recognition intrusion that becomes batter than current approaches which is capable to increase the detection rates and reduce the number of features. Using and testing other optimization techniques such as (Particle swarm optimize) PSO and (gravity search algorithm) GSA because these optimizations techniques have higher speed in convergence.
In [24], K. Atefi et al. proposed an improved anomaly detection according to GA and SVM that raised or enhanced the accuracy to recognize intrusion while focus on the hybrid model by combining (GA and SVM) rather primary algorithms (SVM). When a researcher uses hybrid model this method achieve an acceptable percentage of alarm as far as: True Negative, False Negative, True positive and False positive. But, there are many suspicious behaviors in network intrusion detection, and unfortunately firewall techniques cannot guarantee against intrusion, due to the fact that the defense is extremely vital. Although, various studies already have been carried out in this area, there are few of them that use the potential benefit to combine SVM via GA. For evaluation by this algorithm author used KDDCUP’99 data set. At last, the results illustrated the hybrid model toward primary algorithm have high accuracy to recognize intrusion, additionally, have an acceptable percentage of alarms as far as (TN, FN, TP and FP).

This paper focused on incremental SVM training algorithms to aimed network intrusion detection, and suggested an improved algorithm. It applied on hybrid with modifying kernel function U-RBF, to deal with network intrusion detection. However, given the fluctuation problem that usually happens in traditional incremental SVM's pursue learning process. Authors used benchmark KDD Cup 1999 as the dataset for experiments. Contrasted with different algorithms by experiments, the test of results shows, that the oscillation problem is more comforted by improving the incremental SVM algorithm in the training process, acquire satisfactory performance, nevertheless, its reliability is high. This work is not suitable based on the detection rate on foreseeing attacks, specifically for attacks of U2R and R2L [25].

Catania et al. in [26], proposed an approach for autonomous labeling algorithm of normal traffic in the network. This algorithm is applied to the SVM algorithm, when the class distribution is not imbalanced. The current proposed methods using SVM in IDS are accurate when the normal traffic has hugely been more than the number of attacks on the dataset; while this situation cannot be always accrued. The main advantage of using the proposed approaches is, when the training dataset which contains normal traffic and number of attacks, the approach can depict the normal traffic. The SbSVM is not evaluated for the real time situation. It is important to evaluate and improve the ability of applying this approach for real time applications.

Authors in [27], proposed a combining IG feather selection and SVM classifier in IDS model, but selection of suitable parameters effect on performance of SVM. This paper use NSL-KDD as a dataset. Results illustrate this model can give a lower false alarm and higher detection rate toward regular SVM. The limitation of this paper is limited experiments; just two swarm intelligence algorithms. Proposed algorithms can be used other kinds of swarm intelligence algorithms.

S. Araki et al. in [28], using multistage OC-SVM and feature extraction represented a method to detect unknown attacks. The unknown attacks cannot be detected using signature-based IDS. Since the detecting unknown attacks, IDS is applied; yet, this method cannot completely distinguish between sophisticated attacks and known attack. Furthermore, for evaluating the proposed method used Kyoto2006+ as data set. This represents a method achieve unknown attacks that in the archive have not been stored. This method is poor in second stage classifier to a detection rate of unknown attacks. For solving that poorness can extract more viable features at reflect exclusive behavior of unknown attacks, and also can perform clustering and filtering in order to decrease an effect of noisy data.

G. Zhao et al. in [29], represents a new network intrusion detection method using SVM. In addition the attributes are optimized using k-fold cross validation. Yet, the anomaly based and signatures base both rates as False Negative and False Positive are high. They have used their collected dataset. Comparisons on some proposed machine learning method, the result of online data experimental indicate that this technique can be used to decrease the rate of False-Negatives in the IDS.
In [30], authors reduce the dimension of the feature vectors and that decrease training time, and achieved better performance. The authors proposed to use combined KPCA SVM classifier with an ICPSO model for intrusion detection. Radial Bayes methods produced many support vectors and increased training time. In this paper for experiments, select samples randomly from KDDCUP99 as data sets. Experimental results showed that the in proposed method, performance is high for intrusion detection, also for training and testing required less time. The detection rate of forecasting attacks, specifically the attacks of R2L and U2R are poor. To improve the limitation; can investigation required in SVM parameters optimization by various optimization algorithm.

J. Song et al. in [31], have proposed a novel anomaly detection method which can be tuned and optimize automatically without pre-defining them. Moreover, evaluate this method via real traffic data achieved from Kyoto University honeypots as a dataset. It is still difficult to deploy many IDS methods into real network environments due to the fact that they need several factors during their process. Moreover, IDS managers and operators suffer from optimizing and tuning the need factors according to the alternative of their network characteristics. The experimental results illustrate that the proposed method is greater than the Song’s method from the aspect of the accuracy performance. This method is not beneficial and difficult to apply on various network environments. This method also poor in experiments due to auteurs didn’t effort in more real traffic data and superior range. The proposed method required training and testing in maximum databases and apply in various real network environments.

W. Feng et al. in [32], introduced a novel algorithm (CSVAC) for the generate classifiers using clustering techniques based on a machine-learning that is applied to intrusion detection problem. High dimensionality and exponential data cannot be handled using traditional IDS techniques, especially in real time IDS. This article used a standard benchmark KDD99 data set for implemented and evaluated. The paper uses the advantages of the hybrid method Combining Support Vectors with Ant Colony SVM (CSOACN), for the real time application of the proposed method. To have a better performance, combination between the privacy preserving OLAP with the proposed method is suggested to improve effectiveness and the flexibility of the IDS.

In [33], M. Sailaja et al. proposed an architecture called IHDAIDS for the NIDS. Beside the real time potential application of IHDAIDS, it is intelligent, hybrid and adaptive. In addition, it produced a low rate false alarm and required a lower rate of human intervention. However, encryption software and firewalls, which are used for intrusion detection, do not provide complete security of the networks. Aggregation of these techniques with IDS can provide improved security. The proposed architecture has applied and tested on the KDDCUP’99 dataset. This architecture has used the combination of host based and network based IDS to provide a high level accuracy.

4. Critical Analyses

In this section we provide a critical review of different approaches in tabulated form. In this table the name of author and the used methods and also the dataset which are used are presented. The strength and weakness and false alarm and detection rate is presented. In the case of the detection rate, the detection rate or accuracy of recognition is reported, however in the case of cross validation the maximum reported values have showed.

<table>
<thead>
<tr>
<th>Ref</th>
<th>Approach</th>
<th>Solution</th>
<th>Dataset</th>
<th>Strength</th>
<th>Weakness</th>
<th>False alarm %</th>
<th>Detection Rate % (Max)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[7]</td>
<td>N-KPCA-GA-SVM kernel principal component</td>
<td>Combination of KPCA, SVM and GA algorithms.</td>
<td>KDD CUP99</td>
<td>Faster convergence speed. Performs higher predictive</td>
<td>The Hybrid KPCA and SVM with GA is a complex and</td>
<td>0.95</td>
<td>96.37</td>
</tr>
<tr>
<td>Reference</td>
<td>Method</td>
<td>Analysis</td>
<td>Results</td>
<td>Latency</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>--------</td>
<td>----------</td>
<td>---------</td>
<td>---------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[20]</td>
<td>GFR-gradually feature removal</td>
<td>High efficient and accurate classifiers established with combining of SVM, ant colony algorithm and clustering method.</td>
<td>KDD 99</td>
<td>Hybrid method has low performance from the real time aspect</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[21]</td>
<td>CSV-B SVM Candidate Support Vector - Incremental SVM</td>
<td>Enhanced learning technique for SVM.</td>
<td>KDD Cup 1999</td>
<td>Better recognize rate and false alarm rate than usual classification</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[22]</td>
<td>SVM and NN</td>
<td>SVM and ANN, the features are ranked and for five different classes attacks.</td>
<td>DARPA</td>
<td>Time consuming and hard task to trigger</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[23]</td>
<td>SVM -GPC-10 Genetic principal component</td>
<td>According to optimal feature subset election which is gotten from GA and PCA.</td>
<td>KDD cup</td>
<td>Improved SVM performance increases the detection rates and reduces the number of features</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[24]</td>
<td>SVM-GA</td>
<td>Hybrid model by combining (GA and SVM)</td>
<td>KDD Cup'99</td>
<td>High Accuracy to recognize intrusion</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[25]</td>
<td>RS-ISVM-reserved set - Incremental SVM</td>
<td>An incremental SVM training algorithms is used, hybrid with modifying kernel function U-RBF;</td>
<td>KDD Cup 1999</td>
<td>This work is not suitable based on detection rate on foreseeing attacks, specifically for attacks of U2R and R2L</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[26]</td>
<td>ShSVM</td>
<td>Approach for autonomous labeling algorithm of normal traffic (when the class distribution is not imbalanced)</td>
<td>DARPA</td>
<td>This method is poor in second stage classifier to detection rate of unknown attacks</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[27]</td>
<td>IG-ABC-SVM Information Gain-Artificial Bee Colony</td>
<td>A combining IG feature selection and SVM classifier in IDS model is proposed</td>
<td>NSL-KDD</td>
<td>The experiments using just two swarm intelligence algorithms</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[28]</td>
<td>OC-SVM One-Class SVM</td>
<td>Multistage OC-SVM and feature extraction represented a method to detect unknown attacks.</td>
<td>Kyoto 2006</td>
<td>This method is poor in second stage classifier to detection rate of unknown attacks</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[29]</td>
<td>Multiclass SVM</td>
<td>In addition the attributes are optimized using k-fold cross validation.</td>
<td>Own collected dataset</td>
<td>This technique can be used to decrease the rate of False-Negatives in the IDS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[30]</td>
<td>KPCA-ICPSO-SVM kernel principal</td>
<td>Use combine KPCA SVM classifier with</td>
<td>KDD CUP'99</td>
<td>Detection rate on forecasting attacks</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: KDD Cup 1999 is a dataset used for benchmarking intrusion detection systems.
**5. Conclusion**

We have critically analyzed various approaches of intrusion detection system based on support vector machine for intelligent detection of normal or abnormal activity. We only evaluated those approaches that utilized the SVM in IDS. Although SVM is applied to detect known attacks and unknown attacks, hybrid of SVM and using different parameters in SVM, has impact on IDS performances. Some of these techniques using SVM are suitable for real time application, but others are not applicable in context of implementation for real time usage. Therefore, having a comprehensive view of the application of SVM in IDSs and is indispensable before practical usages. In addition, we propose a more elaborate review on IDSs. Critical evaluation in tabulated form summarizes latest contribution research and easily to grasp the overall picture in this domain. In future, systematic review on IDS based SVM and comprehensive benchmarking comparison with other detection techniques would be our main focused.

**References**


**Authors**

**Leila Mohammadpour**, She received her BS degree in Computer Engineering from Islamic Azad University of Shiraz 2010 IRAN. She obtained her MSc. degree in Computer Science from the University Putra Malaysia (UPM) in 2014. She has experience in IDS system, and her current research area is on the SDN. Presently she is PhD student at University of Malaya, Malaysia. Research interests are network security, IDS and SDN. He can be reached at le.vesal@gmail.com.

**Mehdi Hussain**, he received his B.S. degree in Computer Science from The Islamia University Bahawalpur in 2005 and the MS degree in Computer Science from SZABIST Islamabad in 2011, Pakistan. He had been employed as a senior software engineer at Streaming Networks (Software House), Pakistan from 2006 to 2014. He had selected as funded scholar at National University of Science and Technology (NUST) under faculty development program 2014. Presently he is Ph.D. scholar at University of Malaya, Malaysia. Research interests are information hiding, multimedia security, and steganography. He can be reached at mehdi141@hotmail.com

**Alihossein Aryanfar**, he is a PhD candidate at University Putra Malaysia (UPM). He has 10 years experiences in teaching, research and development. Research interests are Artificial Intelligence, Computer Vision, and Pattern Recognition. He can be reached at a_aryanfar85@yahoo.com.

**Vahid Maleki Raee** received his BS degree in Computer System and Network University of Malaya, Malaysia 2012. He obtained his MS degree in Computer System and Network University of Malaya, Malaysia 2015. He has experience in SDN, cloud computing and cisco networking. He can be reached at vahid.m19r@yahoo.com.
Fahad Sattar, he received his BS degree in Computer Engineering from University of Management and Technology Lahore, Pakistan in 2005. He obtained his MS degree in Computer Science from SZABIST Islamabad 2011 Pakistan. He has 7 years working experience in Pakistan Media (PTV). Research interests are Information security, VOIP. He can be reached at fahad.sattar1@gmail.com.