Feature Selection for Detection of Ad Hoc Flooding Attacks

Sevil Sen and Zeynep Dogmus

Abstract In recent years ad hoc networks have become very attractive for many applications such as tactical and disaster recovery operations. However they are vulnerable to many attacks. The vulnerabilities of wired networks such as denial of service (DoS), eavesdropping, spoofing and the like, becomes more acute in these networks. Especially it is hard to differentiate DoS attacks in these highly dynamic systems. In this research, we design an intrusion detection model using Support Vector Machines (SVM) in order to detect a popular DoS attack on these networks, namely ad hoc flooding attacks. We evaluate its performance on simulated networks with varying traffic and mobility patterns. Furthermore we investigate to choose the relevant features using Genetic Algorithms (GA) in order to increase SVM performance on detection of these attacks.

1 Introduction

Mobile ad hoc networks are one of the fastest growing areas of research. This new type of self-organizing network combines wireless communication with a high degree node mobility. They do not have any fixed infrastructure such as base stations or centralized management points as in conventional networks. The nodes cooperate with each other to provide basic functionality such as routing in a network, independent of any fixed infrastructure or centralized management. This flexibility makes them attractive for many applications such as military applications, disaster recovery operations, and virtual conferences.

Sevil Sen
Department of Computer Engineering, Hacettepe University, e-mail: ssen@cs.hacettepe.edu.tr

Zeynep Dogmus
Faculty of Engineering and Natural Sciences, Sabanci University e-mail: zeynepdogmus@sabanciuniv.edu

These networks have different properties than conventional networks such as lack of central points, dynamic topology, resource-constraints and the like. This new networking is by its very nature more vulnerable to attacks than wired networks. Furthermore their specific features present a challenge for security solutions such as intrusion detection systems. Especially the impact of mobility on intrusion detection is a complex matter. It has both positive and negative impacts on security. On one hand, mobility helps security that victims could receive (direct or promiscuously) only parts of the falsified packets due to link breakages caused by mobility. So, the attacker might partially achieve his goal. On the other hand, the attacker hides himself from the detection system under high mobility which makes it difficult to differentiate normal behaviour of the network from anomaly behaviour in this environment. We particularly focus on ad hoc flooding attacks in the second group. These highly mobile, complex systems should be modelled in order to detect attacks against them. In this research we employ SVM algorithms to achieve that. We evaluate our model created by SVM on networks with varying traffic and mobility patterns. Moreover, we investigate the selection of relevant features by using GA in order to model these complex systems better. We know of no other proposed approach in the literature on selecting features for intrusion detection in these networks.

2 Related Work

The specific features of ad hoc network make application of existing intrusion detection approach to this environment problematic. Therefore researchers have been working on new approaches or adaptation of existing approaches to ad hoc networks.

One of the most commonly proposed techniques on these networks is specification-based intrusion detection. This technique has been applied to a variety of routing protocols on ad hoc networks [1][2]. A few signature-based IDSs have also been proposed. In [3], an approach is proposed based on a stateful misuse detection technique which defines state transition machines for detecting known attacks on AODV [4]. In [5], an IDS is proposed which uses a specification-based technique for attacks that violate the specifications of AODV directly and an anomaly-based technique for other kinds of attacks such as DoS. Since wireless nodes can overhear traffic in their communication range, promiscuous monitoring can also be used to detect attacks such as dropping and modification. Mobile agents have been suggested as another way to provide communication between IDS agents.

Few artificial intelligence based intrusion detection systems have been proposed to explore the complex behavioural space of ad hoc networks. In the first proposed intrusion detection systems for these networks [6], statistical anomaly-based detection is chosen over misuse-based detection, so unknown attacks could be detected. The SVM Light and RIPPER classifiers are employed and compared in that research. In [7] a Markov-chain based local anomaly detection model is proposed for a Zone-Based IDS architecture. The network is partitioned into zones based on geographic location. Another approach which constructs an anomaly-detection model automatically by extracting the correlations among monitored features is proposed in [8]. Furthermore, they introduce simple rules to determine attack types and sometimes attackers after detecting an attack using cross-feature analysis. In [9], an approach which takes into account limited power issues of nodes by using multi-objective evolutionary computation techniques is proposed. In [10], four classification algorithms are applied for intrusion detection in ad hoc networks and compared.

SVM has been applied to ad hoc networks in order to detect known attacks [6][10]. In [6], while SVM shows a good performance on the routing protocol DSR, its performance on distance-vector routing protocols such as DSDV and AODV is poorer. In [10], SVM is shown to be the best algorithm among four classification methods on detection of some known attacks against AODV (average detection rate: %77, average false positive rate: %0.97). In this research we aim to investigate if we could achieve a good performance with SVM by using an expanded feature set.

3 Intrusion Detection by Using Support Vector Machines

Support Vector Machines (SVM) is a supervised learning algorithm used mainly for classification and regression analysis. It is one of the popular techniques for intrusion detection in conventional networks due to their good performance both in unbalanced and balanced datasets. In this research, we employ this promising technique to ad hoc networks environment.

3.1 Support Vector Machines (SVM) Model

SVM algorithm basically constructs a hyperplane or a set of hyperplanes. In order to have a good separation, the hyperplane's distance to the nearest training point of any class should be maximized. The larger the distance the better classification is achieved. In our experiments, we use libSVM library [11].

In this research the networks are simulated by ns-2 [12]. Mobility patterns of the nodes are simulated by the Random Waypoint model which is created using BonnMotion [13]. Different network scenarios are created with different mobility levels and traffic loads. 50 nodes are placed in a topology of 1000 m by 500 m. TCP traffic is used for communication. The maximum number of connections is set to 20 and 30 to simulate different traffic loads. The maximum speed of nodes is set to 20 m/s and the pause time between movements is set to 40, 20, and 5 s to simulate low, medium, and high mobility respectively. AODV periodic hello messages are used for local link connectivity. The simulations run 5000 s for training and 2000 s for testing.

Table 1 shows the features maintained at each node. This feature set is wider than other approaches which use SVM for intrusion detection in ad hoc networks [6][10]. The features can be categorised into two main groups: mobility-related and packet-related features. Mobility-related features help reflect the mobility model of a node or the network. Some of the mobility features give information about mobility directly such as changes in the number of neighbours. Others can be the results of mobility such as changes in the routing table (*e.g.* number of new routes, number of invalidated routes) in a time interval. Packet-related features include information about the frequency of the routing protocol control packets (RREQ, RREP, RERR) sent, received, or forwarded in a time interval. All features are local to a node, so no communication with other nodes is needed to gather them.

Table 1 Features

	Features (of a node)
1	no. of neighbours
2	no. of added neighbours
3	no. of removed neighbours
4	no. of active routes
5	no. of routes under repair
6	no. of invalidated routes
7	no. of added routes by route discovery mechanism
8	no. of added routes by overhearing
9	no. of updated routes (modifying hop count, sequence number)
10	no. of added routes under repair
11	no. of invalidated routes due to expiry
12	no. of invalidated routes due to other reasons
13	no. of received route request packets destined to this node
14	no. of received route request packets to be forwarded by this node
15	no. of broadcasted route request packets from this node
16	no. of forwarded route request packets from this node
17	no. of received route reply packets destined to this node
18	no. of received route reply packets to be forwarded by this node
19	no. of initiated route reply packets from this node
20	no. of forwarded route reply packets from this node
21	no. of received broadcast route error packets (to be forwarded or not)
22	no. of broadcasted route error packets from this node
23	no. of received total routing protocol packets
24	no. of received total routing protocol packets to be forwarded
25	no. of initiated total routing protocol packets from this node
26	no. of forwarded total routing protocol packets by this node

In order to evaluate our model, we focus on two metrics: detection rate and false positive rate. The detection rate (DR) shows the ratio of correctly detected intrusions to the total intrusions on the network. The false positive rate (FPR) shows the ratio of normal activities that are incorrectly marked as intrusions to the total normal activities on the network. An acceptable low rate of false alarms is as important as a high detection rate.

For training we use a dataset obtained from a network under medium mobility and traffic. It is an unbalanced dataset where normal cases are much more than abnormal cases. That is the reason we use weight parameter during the construction of our model. We try different weight parameters empirically and obtain different trade-offs between detection rate and false positive rate. Based on that, we choose the following weight parameters: 0.005 for normal cases, 0.1 for abnormal cases. We use C-SVC algorithm which is the default algorithm in libSVM. The cost parameter between 5 and 50 is evaluated, and is chose to be 5. The model trained with these parameters which happens to be the optimal model considering trade-offs between detection rate and false positive rate is chose empirically.

3.2 Experimental Results

We evaluate our model on six networks under varying mobility and traffic levels and the results are demonstrated in Table 2. SVM shows a good performance on detection of ad hoc flooding attacks. As it is shown clearly, detection rate decreases and false positive rate increases under high traffic. It is the traffic level which affects the performance of the model much more than mobility.

Table 2 SVM Performance on Detection of Ad Hoc Flooding Attacks

Simulations	Detection Rate	False Positive Rate
low mobility, medium traffic	97.03%	0.83%
low mobility, high traffic	94.17%	2.10%
medium mobility, medium traffic	97.86%	0.76%
medium mobility, high traffic	96.20%	1.70%
high mobility, medium traffic	97.40%	0.95%
high mobility, high traffic	90.02%	1.83%

As it is stated before, SVM is one of the most promising techniques used for intrusion detection in wired networks. That is the reason we aim to increase its performance for ad hoc networks successfully in this research. Since the right choice of features is much important for any machine learning method, we mainly focus on reducing the features given in Table 1 for a better performance by using genetic algorithms.

4 Selection of Features by Using Genetic Algorithms

The choice of which network characteristics can be used for machine learning is very important. They must contain sufficient information to allow the fundamentals to be developed. However irrelevant and too many features could degrade the performance of the learning algorithms. In this research, we investigate if we could increase the performance of our model with the selection of right features for training.

Since ad hoc networks have complex properties, we use all possible features which could represent its behaviour at the routing level as given in Table 1. However some of these features could not be representative for detecting ad hoc floodding attacks. Genetic algorithms have been used for selection and reduction of features in many areas successfully [14, 15, 16, 17], that's why we use this technique to increase the performance of our model by using the relevant features.

4.1 Genetic Algorithms (GA)

Genetic Algorithms is an evolutionary computation technique inspired from biological evolution. The algorithm starts with creating individuals (generally randomly) which are the candidates solutions for the problem. Traditionally, individuals are represented in binary as strings of 0s and 1s. Each individual is assigned a *fitness* value which shows how the individual solves or comes close to the solution. Some genetic operators (selection, crossover, mutation, reproduction and etc.) are applied on individuals based on their fitness values until the termination criteria is satisfied. The aim is to provide better individuals in the new population.

4.2 Feature Selection

At first, random individuals are created for the solution. These individuals represent which features are used for the SVM model, and which not. SVM algorithm is run for each individual and a fitness value is assigned to each individual based on the formula below. The GA algorithm continues until a defined generation is reached.

$$fitness = detection\,rate - false\,positive\,rate \tag{1}$$

We use ecj 20 toolkit [18] for the GA implementation in our experiments. The GA parameters are selected as follows: 100 for population size, 100 for generation size, 0.9 for crossover probability, 0.1 for reproduction probability. Other parameters used are the default parameters of the toolkit. At each generation 100 individual is evaluated by creating a SVM model for each individual. Since our training dataset is huge, fitness values might not be obtained in a reasonable time. That's the reason we use a balanced small subset of our training data and do not employ any weight parameter consequently.

4.3 Experimental Results

GA algorithm is run ten times and the feature set with the highest fitness value is selected. A SVM model is run with this feature set ($\{1, 2, 9, 10, 16, 19, 26\}$ in Table 1) and all training dataset, and evaluated on different network simulations again. The results are demonstrated in Tablo 3. Both an increase in detection rate and a decrease in false positive rate is seen in the results. The false positive rate is below 1% in most of the cases. A noticeable performance increase is achieved by the feature reduction.

Table 3 SVM Performance after Feature Selection

Simulations	Detection Rate	False Positive Rate
low mobility, medium traffic	97.69%	0.23%
low mobility, high traffic	96.49%	1.29%
medium mobility, medium traffic	99.67%	0.19%
medium mobility, high traffic	97.57%	1.08%
high mobility, medium traffic	97.44%	0.39%
high mobility, high traffic	93.25%	1.00%

5 Conclusion

In this research, we aim to detect ad hoc flooding attacks. We investigate the use of SVM in this environment and evaluate its performance on networks with varying traffic and mobility patterns. In order to increase its performance, we explore the selection of relevant features by using GA. It is shown that the performance of SVM has increased with the reduced feature set obtained by GA, both an increase in detection rate and a decrease in false positive rate is observed. As far as we know this is the first attempt on selecting relevant features by using artificial intelligence based techniques for intrusion detection in these networks. In this research, the parameters of SVM is chose empirically. In the future, the effects of these parameters could be investigated by using GA as well.

References

1. Tseng C-Y, Balasubramayan P, Ko C, Limprasittiprn R, Rowe J, Lewitt K (2003) A Specification-Based Intrusion Detection System for AODV. In: Proceedings of the ACM Workshop on Security in Ad Hoc and Sensor Networks 125–134

2. Tseng CH, Wang S-H, Lee W, Ko C, Lewitt K (2006) DEMEM: Distributed Evidence Driven Message Exchange Intrusion Detection Model for MANET. In: Proceedings of the 9th International Symposium on Recent Advances in Intrusion Detection, Springer: 249–271
3. Vigna G, Srinivasan K, Belding-Royer EM, Kemmerer RA (2004) An Intrusion Detection Tool for AODV-Based Ad hoc Wireless Networks. In: Proceedings of the 20th Annaual Computer Security Applications Conference, IEEE Computer Society: 16–27
4. Perkins CE, Royer EM (1999) Ad-hoc on demand distance vector routing. In Proceedings of IEEE Workshop on Mobile Computer Systems 90–100
5. Huang Y, Lee W (2004) Attack Analysis and Detection for Ad hoc Routing Protocols. In: Proceedings of the 7th International Symposium on Recent Advances in Intrusion Detection, Springer: 125–145
6. Zhang Y, Lee W, Huang, Y. (2003) Intrusion detection techniques for mobile wireless networks. Wirel Netw J 9(5):545–556, doi: 10.1023/A:1024600519144
7. Sun B, Wu K, Pooch U (2003) Zone-based intrusion detection for mobile ad hoc networks. Int J of Ad Hoc and Sens Wirel Netw 2
8. Huang Y, Fan W, Lee W, Yu PS (2003) Cross-feature Analysis for Detection Ad-hoc Routing Anomalies. In: Proceedings of the 23rd International Conference on Distributed Computing Systems, IEEE:478–487
9. Sen S, Clark JA (2011) Evolutionary Computation Techniques for Intrusion Detection in Mobile Ad Hoc Networks. Comput Netw 55(15):3441–3457
10. Mitrokotsa A, Tsagkaris M, Douligeris C (2008) Intrusion Detection in Mobile Ad Hoc Networks Using Classification Algorithms., In: Proceedings of the Seventh Annual Mediterranean Ad Hoc Networking Workshop - Advances in Ad Hoc Networking Springer: 133–144
11. LibSVM: A Library for Support Vector Machines. http://www.csie.ntu.edu.tw/ cjlin/libsvm/.
12. The network simulator. http://www.isi.edu/nsnam/ns/. Cited 15 Feb 2012
13. BonnMotion: A mobility scenario generatin and analysis tool. http://net.cs.uni-bonn.de/wg/cs/applications/bonnmotion/. Cited 15 Feb 2012
14. Wroblewski J (1995) Finding Minimal Reducts Using Genetic Algorithms. In: Proceedings of the Second Annual Joint Conference on Information Sciences 186–189
15. Lanzi PL (1997) Fast Feature Selection with Genetic Algorithms: A Filter Approach. In: Proceedings of IEEE Conference on Evolutionary Computation 537–540
16. Yang J, Honavar V (1998) Feature Subset Selection Using A Genetic Algorithm. IEEE Intell Syst 12(2):44–49
17. Huang C-L, Wang C-J (2006) A GA-Based Feature Selection and Parameters Optimization for Support Vector Machines. Expert Syst Appl 31:231–240
18. ecj20: A Java-based Evolutionary Computation Research System. http://cs.gmu.edu/ eclab/projects/ecj/. Cited 15 Feb 2012