Bayesian Learning Networks Approach to Cybercrime Detection

N S ABOUZAKHAR, A GANI and G MANSON
The Centre for Mobile Communications Research (C4MCR),
University of Sheffield, Sheffield
Regent Court, 211 Portobello Street, Sheffield S1 4DP,
UNITED KINGDOM
N.Abouzakhar@dcs.shef.ac.uk,
A.Gani@dcs.shef.ac.uk
G.Manson@dcs.shef.ac.uk

M ABUITBEL and D KING
The Manchester School of Engineering,
University of Manchester
IT Building, Room IT 109,
Oxford Road,
Manchester M13 9PL,
UNITED KINGDOM
mostafa.abuitbel@stud.man.ac.uk
David.king@man.ac.uk

Abstract: The growing dependence of modern society on telecommunication and information networks has become inevitable. The increase in the number of interconnected networks to the Internet has led to an increase in security threats and cybercrimes such as Distributed Denial of Service (DDoS) attacks. Any Internet based attack typically is prefaced by a reconnaissance probe process, which might take just a few minutes, hours, days, or even months before the attack takes place. In order to detect distributed network attacks as early as possible, an under research and development probabilistic approach, which is known by Bayesian networks has been proposed. This paper shows how probabilistically Bayesian network detects communication network attacks, allowing for generalization of Network Intrusion Detection Systems (NIDSs). Learning Agents which deploy Bayesian network approach are considered to be a promising and useful tool in determining suspicious early events of Internet threats and consequently relating them to the following occurring activities.

Keywords: Networks Intrusion Detection, Bayesian Networks, and Bayesian Learning

1-Introduction

Generally, inference methods for detecting network attacks either use signature analysis or statistical anomaly detection approaches. The main advantage of the former approach is attack specificity, however, it may not be able to be generalised. The latter detects attacks probabilistically, allowing for generalization, however, it may not be able to specify. Statistics involve the fitting of models to data and making inferences from these models [1]. In pattern recognition the inferences that one wants to make are ones of assignment. For example, given a traffic trace of a network, which could be represented by a form of a dataset, we want to classify it either as an attack or not. Regardless the models used, the ultimate goal would be, how well does our model detect or classify attacks and respond to them later on? Therefore, in any statistical anomaly detection approach, the system requires the estimation of two quantities: the probability of detection (PD) and the probability of false alarm (PFA). In general, it is not possible to simultaneously achieve a PD of 1 and PFA of 0, and one generally can not simply state the PD and PFA that one wants and design the algorithm to provide them [1]. When we say or hear a statement of the kind "the network has 70% probability that it has been attacked", this probability expresses a subjective degree of belief, and this leads to the Bayesian formulation of probability.

This research paper proposes an innovative prediction approach for network intrusion detection. Section 2 discusses the concept of Bayesian learning networks model within the context of intrusion detection systems (IDSs). A model of Bayesian network for detecting network Distributed Denial of Service (DDoS) attacks is introduced graphically. The model is further developed in section 3 using powerful learning methods to extract the variable nodes of Bayesian

network directly from a dataset, which is generated by MIT Lincoln Lab. This dataset is meant to evaluate and enhance the research in the field of networks IDSs. The developed Bayesian learning model corresponds to the adaptive knowledge, which is mainly associated with the dynamic organization of the learning detector and the association discovery of parent variables and their events, while automatically gathering information that is contributed to the knowledge of all the parameters within the target variable node. Finally, the detector model performance in terms of the error rate has been measured using lift chart. The Lift chart is a standard detection evaluation method to validate machine learning algorithms. Useful and promising results were achieved and analysed to determine the intrusion detection prediction rate for the network attacks target variable value.

2-Bayesian Learning Network Extraction from a dataset

A Bayesian network is a graphical model that encodes probabilistic relationships among variables of interest. When used in conjunction with statistical techniques, the graphical model has several advantages for data analysis [2]. Bayesian network provides a set of learning entities that compute models over data stored within a network, and each model encodes dependencies among all variables. A Bayesian network can be used to learn causal relationships, and hence can be used to gain understanding about a problem domain such as networks security and to predict the consequences of intervention by combining a prior knowledge and data.

Bayesian networks have been turned into a powerful tool for data analysis, when the statistical roots of Bayesian networks led to the development of learning methods to extract them directly from databases rather than relying on human domain experts [3]. Therefore, in order for us to learn a Bayesian network, it is required to have a dataset that at least roughly represents our main high level defined variable nodes in Figure 3 model. The 2000 DARPA Intrusion Detection Evaluation Program which was prepared and managed by MIT Lincoln Labs has provided the necessary dataset [4]. The main objective of that program was to enhance the research evaluation in intrusion detection. A standard set of data to be audited, which includes a wide variety of intrusions simulated in a military network environment, was provided. The 1999 Third International Knowledge Discovery and Data Mining Tools Competition KDD intrusion detection contest used 1998 dataset version [5]. The competition task was to build a network intrusion detector, a predictive model capable of distinguishing between ``bad'' connections, called intrusions or attacks, and ``good'' normal connections. MIT Lincoln Labs set up an environment to acquire several weeks of raw TCP dump data for a local-area network (LAN) simulating a typical U.S. Air Force LAN. The generated raw dataset contains about few million connection records. A sample of this data is shown in Table 1. Table 1 lists three lines sample of the data set. The first line represents the names of the variables whereas the second and third correspond to the variables information for each particular connection.

Table 1: A sample dataset

duration	protocol_ type	service	count	srv_ count	serror_ rate	srv_serror_ rate	rerror_ rate	srv_rerror_rate	attack_type
22	tcp	ftp_data	1	1	0	0	1	0	normal.
0	icmp	ecr_i	508	508	0	0	1	0	smurf.

3-Bayesian Network Inference and Intrusion Detection Learning

Bayesian network applies adaptive knowledge, which is concerned with the dynamic organization of the detector and the association discovery of parent variables while automatically gathering information that is contributed to the knowledge of the target variable. Adaptive knowledge is only available in validation mode. Figure 1 shows the marked attack_type node as the target variable and those variables, which are heavily contributed to its knowledge. The arc's thickness is proportional to the strength of the probabilistic relations between the target node and the other nodes within the network. In order to see the amount of information being contributed by each node to the knowledge of the target node is to monitor the shade of each variable node. The lighter the shade of the square inside the node, the greater the amount of information it carries, as shown in Figure 1.

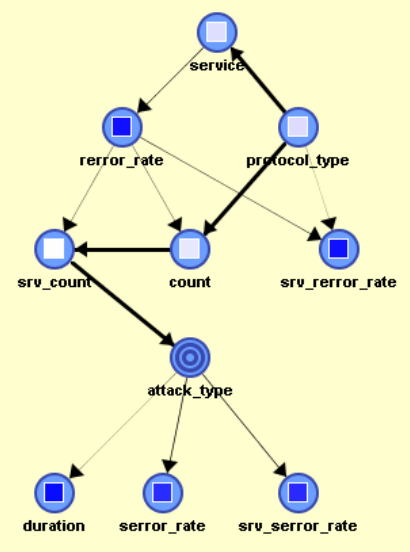


Figure 1- Bayesian network representation of the dataset

With the model learning phase completed, the next step involves the validation mode, which gives access to the communication network state, so to interpret the events that are associated with any particular network attack. Also, the validation mode enables the identification of the conditional probabilities for different states of any variable, by assigning a particular value to any variable. As soon as this kind of variable observation takes place, the probabilities of every node are updated to take the new information into account. This inference process would allow the verification of the network's consistency from the point of view of the choice of nodes and their conditional probabilities. Since the used dataset contains continuous variables, those variables must be made discrete. Therefore, decision tree has been chosen, because the intervals are selected according to the information they contribute to the target variable. Due to the decision tree's discretization, with attack_type as the target node at the detector, all continuous contributed variables are automatically cut up into a number of intervals. The attack_type can be verified by the detector by investigating the parents' nodes at the network. From the detector output shown in the following

figure, the algorithm did in fact find that there is an ecr_i (ECHO_REPLY) service and an ICMP protocol_type association that increases the probability of having smurf attack.
Figure 2 shows srv_count contribution to the new smurf attack probabilities. Because of the propagation of the new information, the probability distributions for the majority of nodes have been updated and as we can see that the probability of having a smurf attack increased from 86.69% to 99.97%. Let's look at the problem from the opposite direction. If the probability of portsweep attack, which represents a reconnaissance process is set to 100% as shown in Figure 3, then the state value of some associated variables would inevitably increase. What characterizes portsweep attack is the TCP protocol and http service association probability have been increased from 8.95% to 57.29% and from 7.15% to 31.21% respectively. Also, we can notice an increase in the srv_serror_rate (% of connections that have "SYN" errors). In the case of the buffer_overflow attack, the rerror_rate variable which represents the % of connections that have "REJ" errors has increased from 72.93% to 86.09%. The approach taken here is completely different from the previous one. In here, we are not only interested in verifying the model. However, the goal is to discover knowledge from the dataset. Bayesian networks learning leads to findings that would never have been as simple to discover by reading more than 100000 records of the dataset.

4-Learning Evaluation and Survival Analysis

It is natural to measure a classifier's prediction performance in terms of the error rate [6]. If the prediction is correct, then it is counted as a success, otherwise it is an error. Generally, in order to predict the performance of a classifier on new data, such as new incoming network traffic, it is required to assess its error rate on an independent dataset that played no part in the formulation of the classifier. The lift chart is a standard measure for evaluating detectors performance. It represents the detection rate of the target variable value (Vertical axis) proportional to the number of processed cases (Horizontal axis) based on the order defined by the learned model [7]. The Y-axis represents the number of responses obtained, and the X-axis represents the sample size as a proportion of the total possible mailout. Figure shows the Lift curve for the attack_type parameter ipsweep. Normally, we'd like to be in a lift chart is near the upper left-hand corner, at the very best, i.e.., the further to the northwest the better. The upper lift point (0,100) denotes the ideal case for accurate detection with minimum cost. The Lift curves shown in Figures 4 (a) and (b) indicate how the Baysian network model performs for detecting ipsweep and saint attacks.

The lift chart represents an effective measure for the validation of the detection process and on whether a given attack classification is valid or not. Lift curve also indicates how far the detector model is effective from the point of view of reducing the false alarms. Therefore, Lift curve could be used to compare various intrusion detection models and provide a common approach for future intelligent detectors evaluations. However, different intelligent network intrusion detection models are based or trained on different datasets, so it is quite difficult to ensure effective comparison methods unless a common dataset is used for that purpose.

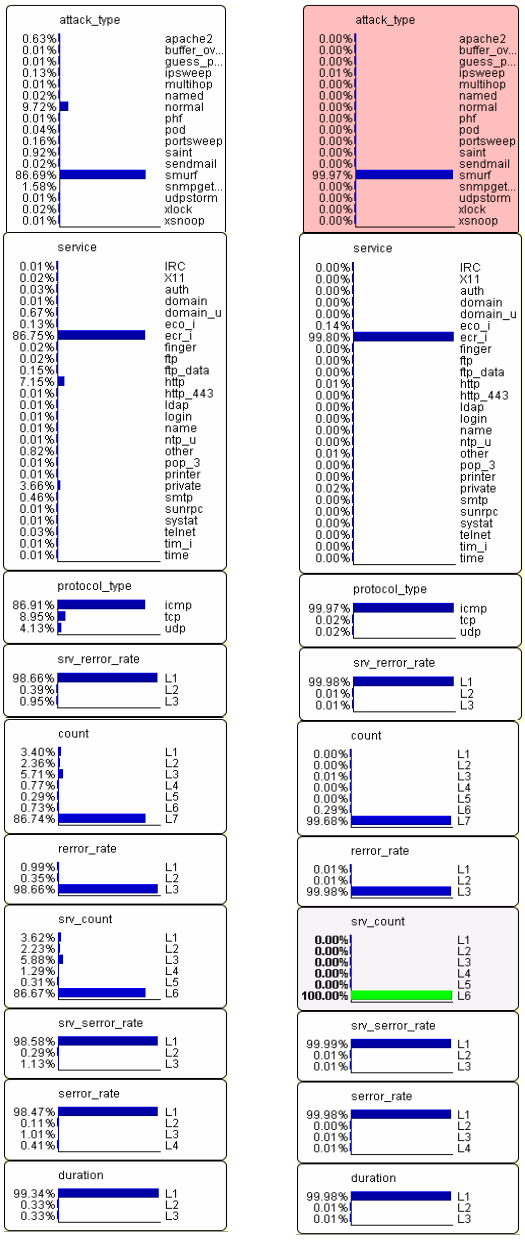


Figure 2- Bayesian detector learns smurf attack and its conditional dependence on srv_count

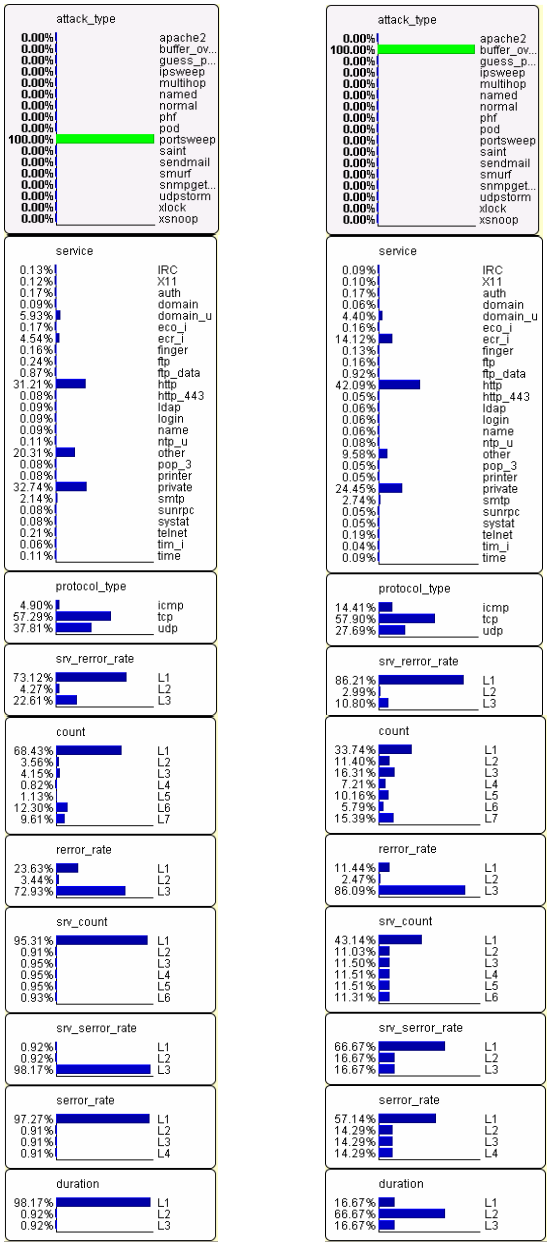
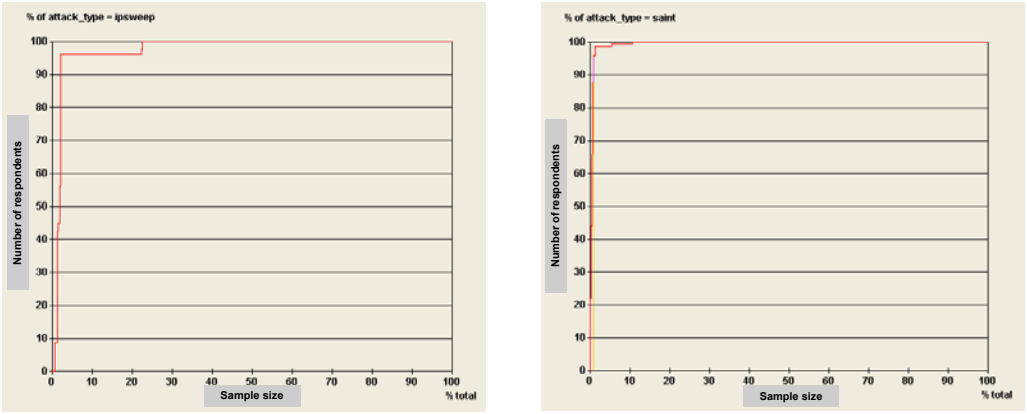


Figure 3- Bayesian detector learns the variable nodes contribution to portsweep & buffer_overflow

5-Conclusion

The Bayesian network applied an adaptive knowledge, which is concerned with the dynamic organization of the detector model and the association discovery of parents variables while automatically gathering information that is contributed to the knowledge of the detector target variable. This process represents the learning phase of the model. The validation mode enabled the identification of the conditional probabilities for different states of any network traffic state variable, by assigning a particular value to any variable. The decision tree approach as an inference process allowed the verification of the Bayesian network's consistency from the point of view of the choice of nodes and their conditional probabilities. Due to the decision tree's discretization,

with attack_type as the target node of the detector, all continuous contributed variables are automatically cut up into a number of intervals.



(a) (b)

Figure 4- Lift curves of ipsweep and saint attacks detection

In order to validate the Bayesian learning detector model, a standard detection evaluation method known as a lift chart has been applied. Promising results were obtained and analysed by generating lift curves to determine the detection rate of the target variable value, which represents the number of responses obtained proportional to the processed cases, which represents the sample size for a previously unseen test data. The lift curve performance results for a sample of DDoS attacks are given and, the results showed that Bayesian unsupervised learning models could reliably detect the majority of the existing attacks with minimum false alarm rates. The more promising is the attack detection models, the better automated response actions to be performed. Therefore, further research should emphasize the need to develop intelligent automated response actions to minimise the progress of the attack and consequently its possible effects.

References

1- David J. Marchette, Computer Intrusion Detection and Network Monitoring, A statistical Viewpoint, 2001, Springer-Verlag, New York, Inc, USA.

2- Heckerman, D. (1995), A Tutorial on Learning with Bayesian Networks, Technical Report MSR-TR-95-06, Microsoft Corporation.

3- Michael Berthold and David J. Hand, Intelligent Data Analysis, An Introduction, 1999, Springer, Italy.

4- http://www.ll.mit.edu/IST/ideval/data/data_index.html, accessed on 01/12/2002

5- http://kdd.ics.uci.edu/, accessed on 01/12/2002.

6- Ian H. Witten and Eibe Frank, Data Mining, Practical Machine Learning Tools and Techniques with Java Implementations, 2000, Morgan Kaufmann, USA.

7- http://www.bayesia.com, accessed on 20/12/2002