Mining TCP/IP Traffic for Network Intrusion Detection by Using a Distributed Genetic Algorithm

Filippo Neri

DSTA - University of Piemonte Orientale
Corso Borsalino 54, 15100 Alessandria (AL), Italy
neri@di.unito.it
neri@mfn.unipmn.it

Abstract. The detection of intrusions over computer networks (i.e., network access by non-authorized users) can be cast to the task of detecting anomalous patterns of network traffic. In this case, models of normal traffic have to be determined and compared against the current network traffic. Data mining systems based on Genetic Algorithms can contribute powerful search techniques for the acquisition of patterns of the network traffic from the large amount of data made available by audit tools.
We compare models of network traffic acquired by a system based on a distributed genetic algorithm with the ones acquired by a system based on greedy heuristics. Also we discuss representation change of the network data and its impact over the performances of the traffic models.
Network data made available from the Information Exploration Shootout project and the 1998 DARPA Intrusion Detection Evaluation have been chosen as experimental testbed.

1 Introduction

The raise in the number of computer break-ins, virtually occurring at any site, determines a strong request for exploiting computer security techniques to protect the site assets. A variety of approaches to intrusion detection do exist [2]. Some of them exploit signatures of known attacks for detecting when an intrusion occurs. They are thus based on a model of virtually all the possible misuses of the resource. The completeness request is actually a major limit of this approach [8].

Another approach to intrusion detection tries to characterize the normal usage of the resources under monitoring. An intrusion is then suspected when a significant shift from the resource's normal usage is detected. This approach seems to be more promising because of its potential ability to detect unknown intrusions [7,3]. However, it also involves major challenges because of the need to acquire a model of the normal use general enough to allow authorized users to work without raising alarms, but specific enough to recognized unauthorized usages [9,4,11].

R. López de Mántaras, E. Plaza (Eds.): ECML 2000, LNAI 1810, pp. 313–322, 2000.

Our approach follows the last philosophy for detecting intrusion and we describe here how it is possible to learn a model of normal use of a network from logs of the network activity. A distributed genetic algorithm REGAL [5,14] is exploited for mining the network logs searching for interesting traffic patterns.

We are well aware that many aspects of deploying in practice learning system to acquire useful traffic patterns are still open including: selecting or building informative data representations, improving recognition performances (i.e., reducing both the rate of false alarms and of undetected intrusions), representing the traffic models for real world deployment (real-time classification of packets), and dealing with the shift in the patterns of normal use of the resources [10].

We concentrate here on the first two issues and we report our findings concerning the impact of different learning methods and of alternative data representation, with respect to the ones used in previous works, on the detection performances. As learning methods, we exploited two rule based systems: a heuristic one, RIPPER [1], and an evolutive one (based on genetic algorithms), REGAL [5,14]. The first system has been selected because of its previous use [11]; it will thus act as benchmark. The second system has been selected because we believe that its intrinsically stochastic behavior should allow the acquisition of alternative robust and simpler models [14].

In the following, a description of the used learning systems (Section 2) and of the experiments performed in the Information Exploration Shootout (IES) and DARPA contexts (Section 3 and Section 4) are reported. Finally, the conclusions are drawn.

2 The Learning Tools

In this section, a brief description of the two learning systems that have been exploited will be provide. Extended description of both systems can be found in the literature.

REGAL [5,14] is a learning system, based on a distributed genetic algorithm (GA). It takes as input a set of data (training instances) and outputs a set of symbolic classification rules characterizing the input data. As usual, learning is achieved by searching a space of candidate classification rules; in this case the searching method consists in a distributed genetic algorithm.

The language L used to represent classification rules is a Horn clause language in which terms can be variables or disjunctions of constants, and negation occurs in a restricted form [13]. An example of an atomic expression containing a disjunctive term is *color(x,[yellow, green])*, which is semantically equivalent to *color(x,yellow) or color(x,green)*. Such formulas are represented as bitstrings that are actually the population individuals processed by the GA. Classical genetic operators, operating on binary strings, with the addition of task oriented *specializing* and *generalizing* crossovers are exploited, in an adaptive way, inside the system (for details see [5].

REGAL is a distributed genetic algorithm that effectively combines the Theory of Niches and Species of Biological Evolution together with parallel pro-

cessing. The system architecture is made by a set of extended Simple Genetic Algorithms (SGA) [6], which cooperates to sieve a description space, and by a Supervisor process that coordinates the SGAs efforts by assigning to each of them a different region of the candidate rule space to be searched. In practice this is achieved by dinamically devising subsets of the dataset to be characterized by each SGA. Such a form of cooperation is obtained by exploiting a coevolutive approach [15,5].

The system RIPPER [1] also takes as input a set of data and outputs an ordered sequence of classification rules. As usual, learning is achieved by searching a space of candidate classification rules; in this case the searching method consists in the iterative application of a greedy heuristic measure, similar to the Information Gain [16], to build conjunctive classification rules. At each iteration, those training instances correctly classified by the found rules are removed and the algorithm concentrate on learning a classification rule for the remaining one. The system output is an ordered list of classification rules (possibly associated to many classes); they have to be applied in that same order to classify a new instance. An interesting features of this learning method is that it exploits online rule pruning while incrementally building a new classification rule to avoid overfitting.

3 Intrusion Detection in the Information Exploration Shootout Contest

An evaluation of REGAL over an intrusion detection task, by exploiting data from the Information Exploration Shootout Project (IES), is reported in this section. The IES made available network logs produced by 'tcpdump' for evaluating data mining tool over large set of data. These logs were collected at the gateway between an enterprise LAN and the outside-network (Internet). In the IES context, detecting intrusions means to recognize the possible occurrence of unauthorized ('bad') data packets interleaved with the authorized ('good') ones over the network under monitoring. The IES's project makes available four network logs: one is guarantee not to contain any intrusion attempts, whereas the other ones do include both normal traffic and intrusions attempts. In the IES context, no classification for each data packets is requested, instead an overall classification of a bunch of the network traffic, as containing or not attacks, is desired.

An approach to intrusion detection, based on anomaly detection, has been selected. We proceed as follows. IES data can be partitioned, on the base of their IP addresses, into packets exiting the reference installation (Outgoing), entering the installation (Incoming) and broadcasted from host to host inside the installation (Interlan). Three models of the packet traffic, one for each direction, have been built from the intrusion-free dataset. Then, these models have been applied to the three datasets containing intrusions. We expect to observe a significant variation in the classification rate between intrusion-free logs and logs containing intrusions because of the *anormal* characteristics of the traffic produced by the

Table 1. Experimental results of applying RIPPER to IES datasets using the raw data representation

Dataset	interlan	incoming	outgoing
normal	0.04	0.04	0.04
intrusion1	0.23	0.07	0.04
intrusion2	0.09	0.07	0.05
intrusion3	0.08	0.14	0.04

intrusive behavior. If this would actually occur, we could assert that the learned traffic models correctly capture the essential characteristics of the intrusion-free traffic. Experiments have been performed both with RIPPER and REGAL.

When RIPPER is applied to the IES data, the classification rate appearing in Table 1 becomes evident [11]. This results have been obtained by applying RIPPER to the data as available from the tcpdumped files (see Appendix A). No preprocessing over the data, such as feature construction, has been applied. The experimental findings shows that the acquired models do not exhibit very different classification rate when applied to logs containing intrusions with respect to intrusion-free logs. These findings may suggest that the exploited data representation is too detailed with respect to the capability of the learning system. In turn, this causes the learned models to miss the information characterizing intrusion-free traffic.

Table 2. Experimental results of applying RIPPER to IES datasets using a compressed data representation

Dataset	interlan	incoming	outgoing
normal	0.02	0.05	0.04
intrusion1	0.11	0.11	0.21
intrusion2	0.03	0.13	0.12
intrusion3	0.11	0.21	0.12

Following this observation, we develop a more compact representation for the packets that consists in mapping a subset of feature's values into a single value, thus reducing the cardinality of possible features values (see Appendix B). Exploiting this representation, RIPPER's performances become the ones reported in Table 2 and REGAL's performances exploiting the same compact data representation appear in Table 3. The observed figures show a more stable classification behavior of the models across different traffic conditions. Also a more distinct classification performance between the intrusion-free log and the logs including intrusions is evident. A compression-based representation is then a valuable way of increasing classification performances without introducing complex feature that may involves additional processing overhead. An evaluation of

```
IF      srcprt(x,[[0,20],[40,100],[150,200],[>500]]) and
        dstprt(x,[>1024]) and flag(x,[FP,pt]) and
        seq1(x,[[100,150],[200,300],[500,5000],[>10000]]) and
        seq2(x,[[50,100],[200,300],[500,20000]]) and
        ack(x,[[0,3000],[5000,10000]]) and
        win(x,[[0,2000],[>3000]]) and
        buf(x,[<=512])
THEN    IncomingPacket(x)
Coverage: (Interlan, Incoming, Outgoing) = (0, 7349, 0)
```

Fig. 1. Example of a rule characterizing part of the incoming traffic. The rule describes 7349 incoming packets without confusing them with any outgoing or interlan packet

the effect caused by the addition of complex features to the raw network data representation has been performed in [11].

For the sake of clarity, an example of rule characterizing intrusion-free Incoming packets, learned by REGAL, appears in Figure 1. The Incoming packets are characterized in term of the values of the features from their TCP/IP header. This rule successfully covers 7349 Incoming packets without being fooled by any Interlan or Outgoing ones. A description of the predicates appearing in the rule is provided in Appendix A.

4 Intrusion Detection in the 1998 DARPA Intrusion Detection Evaluation Programme

We also performed an additional evaluation of our approach over network logs from 1998 DARPA Intrusion Detection Evaluation Programme [12] whose objective was to survey and evaluate research in intrusion detection. A standard set of data to be audited, which includes a wide variety of intrusions simulated

Table 3. Experimental results of applying REGAL to IES datasets using a compressed data representation

Dataset	interlan	incoming	outgoing
normal	0.02	0.04	0.04
intrusion1	0.12	0.15	0.11
intrusion2	0.06	0.11	0.12
intrusion3	0.12	0.15	0.11

in a military network environment, was provided. We exploited data available from the KDD'99 Intrusion Detection Contest[1].

The raw training data was about four gigabytes of compressed binary TCP dump data from seven weeks of network traffic. This was processed into about five million connection records. Similarly, the two weeks of test data yielded around two million connection records. A connection is a sequence of TCP packets starting and ending at some well defined times, between which data flows to add from a source IP address to a target IP address under some well defined protocol. Each connection is labeled as either normal, or as an attack, with exactly one specific attack type. Each connection record consists of about 100 bytes. Attacks fall into four main categories:
DOS: denial-of-service, e.g. syn flood;
R2L: unauthorized access from a remote machine, e.g. guessing password;
U2R: unauthorized access to local superuser (root) privileges, e.g., various "buffer overflow" attacks;
Probe: surveillance and other probing, e.g., port scanning.
In practice two datafiles containing classified connections are available: one has to be used for acquiring a model of the traffic and the other one for testing its performances. The distinction is important because the test file contains attack types not occurring in the learning file. This is intended to make the task more realistic.

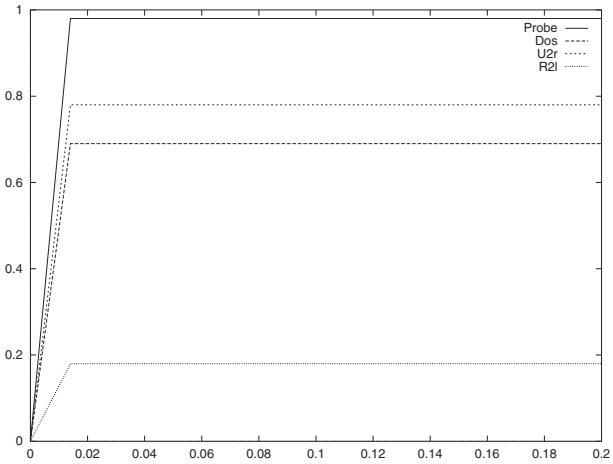

Fig. 2. Detection performances exhibited by RIPPER plus Meta-Learnig on the DARPA test data. An extended representation of the data and a complex learning approach (meta-level learning) have been exploited

[1] Information about KDD'99 Intrusion Detection Contest is available on-line at http://www.epsilon.com/kdd98/task.html.

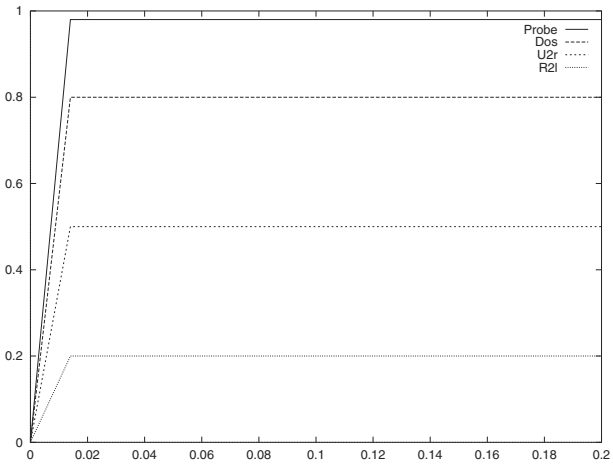

Fig. 3. Detection performances exhibit by REGAL on DARPA test data (no additional Meta-Learning has been used). A compressed data representation has been exploited

In figure 2 and figure 3, performances of RIPPER plus Meta-Learning (as used in [11]) and REGAL over DARPA's data are respectively shown. In the figures, the x axis represents the false alarm rate, i.e. the percentage of 'Normal' connections labeled as intrusions, whereas the y axis represents the detection rate, i.e. the percentage of intrusions that have been correctly recognized. The reported performances have been obtained on the connections occurring in the test file. The reported graphs show similar detection performances, between the models acquired by the systems, for Probe and Remote-To-Local (R2l) attacks types. Instead, REGAL's model performs slightly better on DOS type attacks but worst on User-To-Root (U2r) attacks.

Let consider, now, the modeling approaches exploited by the two systems. Lee and Stolfo [11] run RIPPER over an extended data representation of the tcp connection including, in addition to the basic tcp features, derived information such as: the number of connections to the same host in the past two seconds ('count'), the number of connections to the same service, as the current connection, in the past two seconds ('srv-count'). These features have been chosen on the basis of the authors expertise. A preprocessing of the raw network logs is required in order to exploits this features. Several classifiers (rule sets) for each attack type have been obtained. Eventually meta-learning, i.e. learning at the classifier level, has been applied to produce the reported performances.

REGAL, on the contrary, has been run after applying a compression mapping to the feature values, as described in Appendix B. Only the basic features of a TCP connection have been considered such as: 'duration', stating the length (number of seconds) of the connection, 'protocol-type', stating the type of the

protocol (e.g. tcp, udp, etc.), or 'src-bytes', stating the number of data bytes from source to destination. No additional meta-learning phase is necessary.

5 Conclusions

We investigated the potentiality of a distributed genetic learner and of a heuristic based learner in modeling network traffic to assist in detecting anomaly in data traffic. Two different applicative contexts to detect intrusions have been explored.

We analyzed the effect of exploiting a compressed representation for the network data packet values in modeling pattern of traffic. We are confident that a compression of the values of the packet's features may result in an abstract representation that, on one hand, could allow better recognition performances and, on the other one, could reduce the complexity of acquiring model of the traffic. We believe that discovering the right representation be an important prequisite for the automatic modeling and the on-line deployment of intrusion detection system.

A Appendix. The Information Exploration Shootout Raw Data Representation

The IES data (available on line at http://iris.cs.uml.edu) have been collected by means of the TCPDUMP utility. Taking into account privacy concerns, the data portiong of each packet has been dropped. For each packet in the datasets the following attributes are available:
time - converted to floating pt seconds .. hr*3600+min*60+secs.
addr and port - (just get rid of x.y.256.256.port) The first two fields of the src and dest address make up the fake address, so the converted address was made as: x + y*256.
flag - added a "U" for udp data (only has ulen) X - means packet was a DNS name server request or response. The ID# and rest of data is in the "op" field. (see tcpdump descrip.) XPE - means there were no ports... from "fragmented packets".
seq1 - the data sequence number of the packet.
seq2 - the data sequence number of the data expected in return.
buf - the number of bytes of receive buffer space available.
ack - the sequence number of the next data expected from the other direction on this connection.
win - the number of bytes of receive buffer space available from the other direction on this connection.
ulen - if a udp packet , the length.
op - optional info such as (df) ... do not fragment.
Particular attention has to be taken when dealing with fields like 'op' that contains a large amount of values.

Table 4. Compression mapping applied when dealing with IES network data

Original Value	New Value
0≤srcport<50	srcport=0
50≤srcport<100	srcport=0
<... skipped test ... >	<... skipped text ... >
srcport>20000	srcport=10
<... skipped text ... >	<... skipped text ... >
op contains "DF"	op=1
op contains "NXDomain"	op=2
op contains ANY OTHER VALUE	op=3

B Appendix. The compressed Feature Representation of IES Data

Some features of the IES data may assume a large set of values either continuous or discrete. These large sets do impact over classification performances of the learned models because of the intrinsic difficulty of acquiring rule having a general scope. Then, a reduction of the range of potential values is desirable to increase both the generality of the learned model and to reduce the learning computational complexity.

An alternative approach to this problem consists in adding/building more complex features, combining the basic ones, to the original data representation. We do not follow this approach in this work, because we believe that the previous approach is simpler and should be the first to be analyzed.

As an instance of reducing the range of the feature values, considers that the feature 'srcport' (see Appendix A for a description) may virtually assume any integer number from 0 to 65536. Also, the feature 'op' may assume hundreds of discrete values. Taking into account basic knowledge about the domain, we manually developed the reduction mapping shown in Table 4. This mapping is not to be considered as the best one but as a proof that a simple reduction of the feature values may positively impact over the recognition capabilities.

References

1. W. Cohen. Fast effective rule induction. In *Proceedings of International Machine Learning Conference 1995*, Lake Tahoe, CA, 1995. Morgan Kaufmann. 314, 315
2. D. Denning. An intrusion detection model. *IEEE Transaction on Software Engineering*, SE-13(2):222–232, 1987. 313
3. S. Forrest, S. A. Hofmeyr, A. Somayaji, and T. A. Longstaff. A sense of self for unix processes. In *Proceedings of 1996 IEEE Symposium on Computer Security and Privacy*, 1996. 313

4. A. Ghosh, A. Schwartzbard, and M. Schatz. Learning program behavior profiles for intrusion detection. In *USENIX Workshop on Intrusion Detection and Network Monitoring*. USENIX Association, 1999. 313
5. A. Giordana and F. Neri. Search-intensive concept induction. *Evolutionary Computation*, 3 (4):375–416, 1995. 314, 315
6. D. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, Reading, Ma, 1989. 315
7. S. A. Hofmeyr, A. Somayaji, and S. Forrest. Intrusion detection using sequences of system calls. *Journal of Computer Security*, 6:151–180, 1998. 313
8. S. Kumar and E. Spafford. A pattern matching model for misuse detection. In *National Computer Security Conference*, pages 11–21, Baltimore, 1994. 313
9. T. Lane and C. Brodley. An application of machine learning to anomaly detection. In *National Information Systems Security Conference*, Baltimore, 1997. 313
10. T. Lane and C. Brodley. Approaches to online learning and conceptual drift for user identification in computer security. Technical report, ECE and the COAST Laboratory , Purdue University, Coast TR 98-12, 1998. 314
11. W. Lee, S. Stolfo, and K. Mok. Mining in a data-flow environment: experience in network intrusion detection. In *Knowledge Discovery and Data Mining KDD'99*, pages 114–124. ACM Press, 1999. 313, 314, 316, 317, 319
12. R. Lippmann, R. Cunningham, D. Fried, I. Graf, K. Kendall, S. Webster, and M. Zissmann. Results of the DARPA 1998 offline intrusion detection evaluation. In *Recent Advances in Intrusion Detection 99, RAID'99*, W. Lafayette, IN, 1999. Purdue University. 317
13. R.S. Michalski. A theory and methodology of inductive learning. In R. Michalski, J. Carbonell, and T. Mitchell, editors, *Machine Learning, an Artificial Intelligence Approach*, volume I, pages 83–134. Morgan Kaufmann, Los Altos, CA, 1983. 314
14. F. Neri and L. Saitta. Exploring the power of genetic search in learning symbolic classifiers. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, PAMI-18:1135–1142, 1996. 314
15. M. A. Potter, K. A. De Jong, and J. J. Grefenstette. A coevolutionary approach to learning sequential decision rules. In *Sixth International Conference on Genetic Algorithms*, pages 366–372, Pittsburgh, PA, 1995. Morgan Kaufmann. 315
16. J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, California, 1993. 315