Feature Selection Using a Genetic Algorithm for Intrusion Detection

Guy Helmer
Dept. of Computer Science
Iowa State University
Ames, IA 50011

Johnny Wong
Dept. of Computer Science
Iowa State University
Ames, IA 50011

Vasant Honavar
Dept. of Computer Science
Iowa State University
Ames, IA 50011

Les Miller
Dept. of Computer Science
Iowa State University
Ames, IA 50011

Abstract

We show the use of a genetic algorithm for feature subset selection over feature vectors that describe the system calls executed by privileged processes. Genetic feature subset selection significantly reduces the number of features used without adversely affecting the accuracy of the predictions.

1 OUR APPROACH

The Computer Immunology project at the University of New Mexico (Forrest et al. 1996) developed databases of system calls from normal and anomalous uses of sendmail. We obtained the system call data from their web site at http://www.cs.unm.edu/~immsec.

We designed a feature vector representation to encode an entire process' system call sequence data into a single feature vector. With this encoding, if at least one of the processes involved in a trace is flagged as abnormal, we can identify that an intrusion is taking place. With one minor exception, our feature vector approach identified the intrusions in the sendmail data (Helmer, Wong, Honavar, and Miller, 1998).

Feature subset selection is a possible means of improving the performance of machine learning algorithms for intrusion detection. Genetic algorithms and related approaches are an attractive search algorithm (Yang and Honavar, 1998). The genetic algorithm in our project used standard mutation, crossover operators with 0.001 probability of mutation, 0.6 probability of crossover with rank-based selection, 0.6 probability of selecting the best individual. A population size of 100 was used with 5 generations. The RIPPER learning algorithm was used.

Table 1 shows the abilities of the classifiers to detect intrusive attacks in five feature selection trials. Boldface attack names indicate the attacks that were used in training. 80% of the normal data was also used for training, and the remaining data was used for testing. The "Normal" line is the 20% of normal testing data in which no attacks should be detected. Despite using only about half the features in the original data set, the performance under genetic feature selection was the same or better than when the entire set of features was used for learning.

Table 1: Detection of Attacks

Attack	All Features, 1 Trial	Times Detected in 5 F. S. Trials	Attack	All Features, 1 Trial	Times Detected in 5 F. S. Trials
chasin	1	5	sm5x	1	5
decode1	1	5	smdhole	1	5
decode2	1	5	**sscp-1**	1	5
fwd-lps-1	1	5	**sscp-2**	1	5
fwd-lps-2	0	1	sscp-3	1	5
fwd-lps-3	1	5	**syslog-l1**	1	5
fwd-lps-4	1	5	syslog-l2	1	5
fwd-lps-5	1	5	**syslog-r1**	1	5
recursive	1	5	syslog-r2	1	5
sm565a	1	5	Normal	1	0

2 CONCLUSIONS

Feature subset selection reduced the number of features in the data, which should result in less data required for training due to the smaller search space. Feature selection also gave equivalent accuracy with a smaller set of features.

References

Forrest, S., Hofmeyr, S. A., Somayaji, A., and Longstaff, T. A. 1996. A Sense of Self for UNIX Processes. In Proceedings of the 1996 IEEE Symposium on Security and Privacy, 120-128, Los Alamitos, CA: IEEE Computer Society Press.

Helmer, G., Wong, S. K., Honavar, V., and Miller, L. 1998. Intelligent Agents for Intrusion Detection. In Proceedings of the 1998 IEEE Information Technology Conference, 121-124, Syracuse, NY.

Yang, J., and Honavar, V. 1998. Feature Subset Selection Using a Genetic Algorithm. Feature Extraction, Construction, and Selection—A Data Mining Perspective, Liu and Motoda, eds. Boston: Kluwer Academic Publishers.