Learning Cost-Sensitve Classification Rules for Network Intrusion Detection using RIPPER

Technical Report CUCS-035-1999

Advanced Intelligent Systems CS4721

Spring 1999 Final Project

Matthew Miller
Computer Science Department
Columbia University, New York, NY 10027
Email: mlm46@cs.columbia.edu

Introduction

A system for automating the process of network intrusion detection is currently underway as part of the JAM Project. This system utilizes many data mining methods to build classifiers of network intrusions which can be used to test live network stream input in order to detect intrusions. This is done by useing Link Analysis and Sequence Analysis methods to determine statistical attributes of network connections to build a set of connection profile records that can be useful in detection [Lee 99]. These statistical attributes have various costs associated with their computation in a live environment. When building a rule-set for classification, it would be quite useful for that rule-set to be built with a sensitivity to the cost of computing each attribute. Low-cost attributes would be biased wherever possible, using high-cost attributes only when needed for reliable classification.

Another learning task performed on data which contains values for attributes which have associated costs is that of the medical diagnosis domain. In this domain, certain tests that may assist in diagnosis have a higher monetary cost than others or may cause more patient discomfort. For example, a CAT scan is more expensive to perform than a blood pressure test, whereas an enema is more uncomfortable than having one’s temperature taken.

Rule-sets for the classification of network intrusion are currently learned using RIPPER, a decision-tree learning algorithm developed by William Cohen of AT&T Laboratories. RIPPER offers a number of modifications to IREP, C4.5, and C4.5rules which have proven to yeild faster training times and lower error rates than these predececessors, but does not provide the capability of associating a cost metric with a data-set’s attributes [Cohen 95]. This work has extended RIPPER to allow for these cost metrics, enabling RIPPER to learn cost-sensitive hypotheses.

Approach

Like its predecessor decision-tree induction algorithms, RIPPER grows rules by adding a test of an attribute to that rule if using that attribute will result in a more accurate separation of the training data. The algorithm uses an information gain function which measures this expected reduction in entropy caused by adding the attribute.

The information gain function can be modified in a number of different ways to weight its value according to the cost of the attribute being tested. Two such ways are presented by Tom Mitchell in his Machine Learning textbook [Mitchell 97].

The first of these methods is that used by Tan and Schlimmer in a robot perception domain where the robot must learn to classify differnt objects based on different sensor input. The cost metric here is the delay time in receiving input from a particular sensor. They modified the information gain function to return the square of the standard information gain divided by the cost of the attribute (Gain^2 / Cost(A)).

The second method presented by Mitchell is that used by Nunez for learning medical diagnosis rules. His modification to the information gain function is as follows: (2^Gain - 1) / (Cost(A)) + 1)^w. Here, a cost-sensitivity constant (w) is used which must be on the interval [0, 1]. This constant defines the weight that cost should play in the computation of information gain.

The method chosen for this experiment was the second method. This was done in order to allow for different levels of cost-sensitivity to be defined in order to produce rule-sets that are customized to the level of importance of computational cost in the network environment for which the rule-set is being produced.

The Data

The JAM project group participated in the 1998 DARPA Intrusion Detection Evaluation Program, which was prepared and managed by MIT Lincoln Labs. The objective of this program was to survey and evaluate research in intrusion detection. A standard set of extensively gathered audit data, which inludes a wide variety of intrusions simulated in a military network environment over a period of 7 weeks, was provided by DARPA. This data totaled 4 gigabytes, and was processed (using the aforementioned Link Analysis and Sequence Analysis methods) into 500 megabytes of connection records.

The data used in this experiment constitued a small portion of the entire data-set totaling 1.25MB of connection records which are each 100bytes.

Classification was into one of nine different classes. Eight of these correspond to different types of known intrusions, and one corresponds to normal network activity.

The data contained 18 different attributes which correspond to different features of the network connection record. There are three different categories of attributes that were constructed by the Link Analysis and Sequence Analysis methods, each with different costs of computation associated with them:

- "Intrinsic" features are those features that can be extracted from packet headers without much computational overhead.
- "Same Host" features examine only connections in the past 2 seconds which have the same destination host as the current connection, and calculate statistics related to protocol behavior, service, etc. These features have a higher cost associated with them, as they must search an in-memory connection record database in order to be computed.
- Similarly, "Same Service" features examine only connections in the past 2 seconds that have the same service as the current connection. These features also havea higher cost associated with them for the same reasons as "Same Host" attributes.

The Experiment

Of the data used in this experiment, 80% were used for training and 20% for testing. Eleven different training runs were made using evenly distributed cost sensitivity levels on the interval [0, 1].

Results

The results of the eleven training runs are printed below. The error rates for both the training and test data represent the total percentage of misclassified connection records. The total cost represents the sum of the cost metrics for each attribute tested in each rule of the hypothesis produced by RIPPER.

Cost Sensitivity	Train Error Rate (%)	Test Error Rate(%)	Total Cost
0.0	1.20	5.82	43
0.1	1.20	5.82	43
0.2	1.19	5.82	42
0.3	1.21	5.82	40
0.4	1.21	5.82	40
0.5	1.21	5.82	40
0.6	1.21	5.84	38
0.7	1.24	5.84	38
0.8	1.24	5.84	38
0.9	1.25	5.84	33
1.0	1.25	5.84	33

It is interesting to note that the error rates over both the training and test data do not change dramatically as the cost of each attribute is weighted more heavily in the information gain function. However, the total cost of execution decreases more than linearly as the cost sensitivity level rises. This behavior appears to be the result of additional classification rules not being included in the rule-set and a decrease in the number of attributes that are included in each rule.

Conclusions and Future Work

Adding the ability to learn cost-sensitive hypotheses to RIPPER has shown that it is possible to decrease the total hypothesis cost with little effect on the error rates. However, there is still work to be done before assuming the widespread applicability of the specific RIPPER modifications made in this experiment.

First, there are a wealth of possibilites for different information gain functions that allow for different degrees of weighted cost. Also, these different functions should be tested in multiple domains, over varied sets of data.

Currently, the cost metrics assigned to each attribute do not accurately reflect the actual cost of computation by a real-time intrusion detection system. We are in the process of implementing a real-time system using Network Flight Recorder [NFR 97], a powerful packet sniffing engine with the ability to program scripts in N-code (an interpreted traffic analysis language) in order to compute our connection attributes[Lee2 99]. As this work progresses, a more thorough cost analysis of attribute computation can be performed, and actual measures of performance gains offered by training RIPPER with cost-sensitivity can be determined.

References

[Cohen 95] Cohen, William W.: Fast Effective Rule Induction. From *Machine Learning: Proceedings of the Twelfth International Conference*, 1995.

[Cohen 96] Cohen, William W.: Learning Trees and Rules with Set-valued Features. From *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, 1996.

[Lee 99] Lee, Wenke; Stolfo, Salvatore J.; Mok, Kui W.: A Data Mining Framework for Building Intrusion Detection Models. To appear in the *Proc. 1999 IEEE Symposium on Security and Privacy*, 1999.

[Lee2 99] Lee, Wenke; Park, Christopher T.; Stolfo, Salvatore J.: Automated Intrusion Detection Using NFR: Methods and Experiences. From *USENIX Workshop on Intrusion Detection and Network Monitoring (ID ’99) Proceedings*, 1999.

[Mitchell 97] Mitchell, Tom: Machine Learning. McGraw-Hill, 1997.

[NFR 97] Network Flight Recorder. http://www.nfr.net, 1997.

[Stolfo 99] Stolfo, Salvatore J.; Fan, Wei; Lee, Wenke; Prodromidis, Andreas; Chan, Philip K.: Cost-based Modeling and Evaluation for Data Mining With Application to Fraud and Intrusion Detection: Results from the JAM Project. Yet to be published, April 21, 1999.

Created by Matthew Miller.
Last updated May 13, 1999
Please forward all comments to Matt.