Description of Kyoto University Benchmark Data

Our benchmark data consist of the following 17 statistical features; 14 conventional features and 3 additional features. Among them, the first 14 features were extracted based on KDD Cup 99 data set, which is a very popular and widely used performance evaluation data in intrusion detection research field[1]. KDD Cup 99 data set consists of 41 features, but the existing researches verified that many of them are duplicate and redundant[2]. Among 41 original features of KDD Cup 99 data set, we have extracted only 14 significant and essential features from the raw traffic data obtained by honeypot systems[3] that are deployed in Kyoto University. Addition to those 14 features, we have also extracted additional 3 features which may enable us to investigate more effectively what happens on our networks. Of course, they also can be utilized for training and testing our data with 14 convention features. Note that the order of the below features is exactly the same to that of the actual data.

============14 conventional features=============

1. Duration: the length (number of seconds) of the connection
2. Service: the connection's service type, e.g., http, telnet, etc
3. Source bytes: the number of data bytes sent by the source IP address
4. Destination bytes: the number of data bytes sent by the destination IP address
5. Count: the number of connections whose source IP address and destination IP address are the same to those of the current connection in the past two seconds
6. Same_srv_rate: % of connections to the same service in Count feature
7. Serror_rate: % of connections that have "SYN" errors in Count feature
8. Srv_serror_rate: % of connections that have "SYN" errors in Srv_count(the number of connections whose service type is the same to that of the current connection in the past two seconds) feature
9. Dst_host_count: among the past 100 connections whose destination IP address is the same to that of the current connection, the number of connections whose source IP address is also the same to that of the current connection
10. Dst_host_srv_count: among the past 100 connections whose destination IP address is the same to that of the current connection, the number of connections whose service type is also the same to that of the current connection
11. Dst_host_same_src_port_rate: % of connections whose source port is the same to that of the current connection in Dst_host_count feature
12. Dst_host_serror_rate: % of connections that have "SYN" errors in Dst_host _count feature

13. Dst_host_srv_serror_rate: % of connections that "SYN" errors in Dst_host_srv_count feature
14. Flag: the state of the connection at the time the summary was written (which is usually when the connection terminated). The different states are summarized in the below section.

=============3 additional features=============

1. IDS_detection: reflects whether IDS(Intrusion Detection System) triggered an alert for the connection; '0' means any alerts were not triggered, and an arabic numeral(except '0') means the different kinds of the alerts. Parenthesis indicates the number of the same alert observed during the connection. We used Symantec IDS[4] to extract this feature.
2. Malware_detection: indicates whether malware, also known as malicious software, was observed in the connection; '0' means no malware was observed, and a string indicates the corresponding malware observed at the connection. We used 'clamav' software to detect malwares. Parenthesis indicates the number of the same malware observed during the connection.
3. Ashula_detection: means whether shellcodes and exploit codes were used in the connection by using the dedicated software[5]; '0' means no shellcodes and exploit codes were observed, and an arabic numeral(except '0') means the different kinds of the shellcodes or exploit codes. Parenthesis indicates the number of the same shellcode or exploit code observed during the connection.

Connection State Summaries

- S0: Connection attempt seen, no reply.
- S1: Connection established, not terminated.
- SF: Normal establishment and termination.
- REJ: Connection attempt rejected.
- S2: Connection established and close attempt by originator seen (but no reply from responder).
- S3: Connection established and close attempt by responder seen (but no reply from originator).
- RSTO: Connection established, originator aborted (sent a RST).
- RSTR: Established, responder aborted.
- RSTOS0: Originator sent a SYN followed by a RST, we never saw a SYN ACK from the responder.
- RSTRH: Responder sent a SYN ACK followed by a RST, we never saw a SYN from the (purported) originator.
- SH: Originator sent a SYN followed by a FIN, we never saw a SYN ACK from the responder (hence the connection was "half" open).
- SHR: Responder sent a SYN ACK followed by a FIN, we never saw a SYN from the originator.
- OTH: No SYN seen, just midstream traffic (a "partial connection" that was not later closed).

References

1. The third international knowledge discovery and data mining tools competition dataset KDD99-Cup http://kdd.ics.uc i.edu/databases/kddcup99/kddcup99.html, 1999.
2. Srinivas Mukkamala, Andrew H.Sung., “Identifying Significant Features for Network Forensic Analysis Using Artificial Intelligent Techniques”, International Journal of Digital Evidence, Volume 1, Issue 4, 2003.
3. Jungsuk Song, Hiroki Takakura and Yasuo Okabe, “Cooperation of Intelligent Honeypots to Detect Unknown Malicious Codes”, WOMBAT Workshop on Information Security Threat Data Exchange (WISTDE 2008), The IEEE CS Press, Amsterdam, Netherlands, 21-22 April 2008.
4. Symantec Network Security 7100 Series.
5. http://www.secure-ware.com/contents/product/ashula.html