**DERIVED FEATURES**

Stolfo et al. defined higher-level features that help in distinguishing normal connections from attacks.  There are several categories of derived features.

The ``same host'' features examine only the connections in the past two seconds that have the same destination host as the current connection, and calculate statistics related to protocol behavior, service, etc.

The similar ``same service'' features examine only the connections in the past two seconds that have the same service as the current connection.

"Same host" and "same service" features are together called  time-based traffic features of the connection records.

Some probing attacks scan the hosts (or ports) using a much larger time interval than two seconds, for example once per minute.  Therefore, connection records were also sorted by destination host, and features were constructed using a window of 100 connections to the same host instead of a time window.  This yields a set of so-called host-based traffic features.

Unlike most of the DOS and probing attacks, there appear to be no sequential patterns that are frequent in records of R2L and U2R attacks. This is because the DOS and probing attacks involve many connections to some host(s) in a very short period of time, but the R2L and U2R attacks are embedded in the data portions   
of packets, and normally involve only a single connection.

Useful algorithms for mining the unstructured data portions of packets automatically are an open research question.  Stolfo et al. used domain knowledge to add features that look for suspicious behavior in the data portions, such as the number of failed login attempts.  These features are called ``content'' features.

A complete listing of the set of features defined for the connection records is given in the three tables below.  The data schema of the contest dataset is available in [machine-readable form](http://kdd.ics.uci.edu/databases/kddcup99/kddcup.names) .   
  

|  |  |  |  |
| --- | --- | --- | --- |
| *feature name* | *description* | *type* |  |
| duration | length (number of seconds) of the connection | continuous | OK |
| protocol\_type | type of the protocol, e.g. tcp, udp, etc. | discrete | OK |
| service | network service on the destination, e.g., http, telnet, etc. | discrete | OK PORT |
| src\_bytes | number of data bytes from source to destination | continuous | OK |
| dst\_bytes | number of data bytes from destination to source | continuous | OK |
| flag | normal or error status of the connection | discrete | ? |
| land | 1 if connection is from/to the same host/port; 0 otherwise | discrete | ? |
| wrong\_fragment | number of ``wrong'' fragments | continuous | OK |
| urgent | number of urgent packets | continuous | OK |
| Table 1: Basic features of individual TCP connections. | | |  |

|  |  |  |  |
| --- | --- | --- | --- |
| *feature name* | *description* | *type* |  |
| hot | number of ``hot'' indicators | continuous |  |
| num\_failed\_logins | number of failed login attempts | continuous |  |
| logged\_in | 1 if successfully logged in; 0 otherwise | discrete |  |
| num\_compromised | number of ``compromised'' conditions | continuous |  |
| root\_shell | 1 if root shell is obtained; 0 otherwise | discrete |  |
| su\_attempted | 1 if ``su root'' command attempted; 0 otherwise | discrete |  |
| num\_root | number of ``root'' accesses | continuous |  |
| num\_file\_creations | number of file creation operations | continuous |  |
| num\_shells | number of shell prompts | continuous |  |
| num\_access\_files | number of operations on access control files | continuous |  |
| num\_outbound\_cmds | number of outbound commands in an ftp session | continuous |  |
| is\_hot\_login | 1 if the login belongs to the ``hot'' list; 0 otherwise | discrete |  |
| is\_guest\_login | 1 if the login is a ``guest''login; 0 otherwise | discrete |  |
| Table 2: Content features within a connection suggested by domain knowledge. | | |  |

|  |  |  |  |
| --- | --- | --- | --- |
| *feature name* | *description* | *type* |  |
| count | number of connections to the same host as the current connection in the past two seconds | continuous | OK |
|  | *Note: The following  features refer to these same-host connections.* |  |  |
| serror\_rate | % of connections that have ``SYN'' errors | continuous | OK |
| rerror\_rate | % of connections that have ``REJ'' errors | continuous | OK |
| same\_srv\_rate | % of connections to the same service | continuous | OK |
| diff\_srv\_rate | % of connections to different services | continuous | OK |
| srv\_count | number of connections to the same service as the current connection in the past two seconds | continuous | OK |
|  | *Note: The following features refer to these same-service connections.* |  |  |
| srv\_serror\_rate | % of connections that have ``SYN'' errors | continuous | OK |
| srv\_rerror\_rate | % of connections that have ``REJ'' errors | continuous | OK |
| srv\_diff\_host\_rate | % of connections to different hosts | continuous | OK |
| Table 3: Traffic features computed using a two-second time window. | | |  |