

Binary Classification of Internet Traffic

Identifying Distributed Denial of Service Attacks

Chris Soyars

Faculty Advisor: Professor Michael Frankel

‡ Distributed Denial of Service (DDOS) cant variables.

METHODS

- ‡ Examine data set to gather comprehensive information on variable properties. Eliminate variables that cannot be used due to poor or lacking information. Identify and impute missing or erroneous data if possible.
 - ‡ ~50000 observations had values NA, infinity, or implausibly negative for certain variables due to lack of precision and accuracy in time-related information. Values were recalculated to restore information.
- ‡ Cluster variables to identify variables of greatest interest.
- ‡ Discretize variables and calculate odds of benign vs ddos traffic for each bin to extract additional information and trends from selected variables. Natural log of odds also calculated for each bin.
- ‡ Eliminate non -significant variables from selection for logistic model.
- ‡ Create logistic regression model using selected variables. Model is trained on 80% of the data set.
- ‡ Refine logistic regression model to eliminate redundancy and select for highly significant variables.

Table 2. Confusion Matrix for Logistic Regression Model with 17 Variables

	Actual Benign	Actual DDOS
Predicted Benign	98.7%	0.3%
Predicted DDOS	1.3%	98.7%

RESULTS

- ‡ Logistic regression model with 17 selected variables has 99.9% concordant pair rate (C=0.999). Model predicts a lower probability of possible DDOS traffic for benign cases than actual DDOS attacks in most observations.
- ‡ Maximal KS statistic of .954
- ‡ Model can be simplified depending on needs and resources. C=.995 with as few as 4 variables.
 - ‡ Initial measurements only: C=.965 with 3 variables, with marginal gains for additional initial statistics.
- ‡ Sensitivity and specificity are maximized with probability threshold of 0.188.
 - ‡ 17 variable model correctly identifies 98.7% of benign traffic and 98.2% of DDOS traffic.

- Traffic patterns, attack methods, and security practices can evolve rapidly.
- ‡ Model likely to benefit from periodic re-evaluation.
- ‡ Unknown success rate against novel attack vectors of the same general class.
- ‡ Limitations & Improvements
 - ‡ No analysis of trends by time or location (IP address)
 - ‡ Lack of precision on some variables
 - ‡ Possible errors from traffic logging software used

SAS CODE

```
%IMPV5 (DSN=class.test, VARS=&varlist, EXCLUDE=Label,
PCTREM=1,MSTD=);

PROC SQL;
    SELECT NAME INTO: VARNAME SEPARATED BY ' '
    FROM DICTIONARY.COLUMNS
    WHERE UPCASE(LIBNAME)="DDOS" AND
    UPCASE(MEMNAME)="DDOS" AND NAME NOT IN("Label");
QUIT;
PROC VARCLUS DATA=import OUTTREE=tree MAXCLUSTERS=71;
    VAR &varname;
RUN;

PROC GLMSELECT DATA=ddos.disc2;
    MODEL label=&mvar / DETAILS=all SELECTION=lasso
STATS=all;
RUN;

PROC LOGISTIC DATA=train DESC OUTEST=betas
    OUTMODEL=scoringdata;
    MODEL label=&mvarsn /SELECTION=BACKWARD
CTABLE pprob=(0.16 to 0.21 by 0.001)
LACKFIT RISKLIMITS;
    OUTPUT OUT=output p=predicted;
    SCORE DATA=valid OUT=ddos.score;
RUN;
```