Intrusion Detection using KDD CUP 99 Dataset

Dataset description

KDD CUP 99 dataset is one of the most frequently used dataset for intrusion detection in academic literature.

The dataset was originally created in 1999 at MIT's Lincoln Laboratory during a a DARPA sponsored event, where an attack scenario to Air-Force base is simulated. First two weeks were attack free; therefore, it is suitable for training anomaly detection algorithms. Next five weeks, various attack was used against simulated air-force base. Resulting TCP dump files was filtered to extract features, and resulting dataset was given to Knowledge Discovery and Data Mining (KDD) yearly competition. Dataset consists of 2 main files, one for training and another for testing.

Size and redundancy

```r
dim(train)
```

```
[1] 4898431      42
```

```r
train.dedup <- unique(train)
nrow(train.dedup)
```

```
[1] 1074992
```

Test dataset has 311,029 patterns and no redundancy.

```
38
39 ```{r}
40 dim(test)
41 ```
```

```
[1] 311029     42
```

Features

There are 41 variables. However, some of these variables are highly correlated.

V1 <chr>	V2 <chr>	correlation <dbl>
serror_rate	srv_serror_rate	1.00
serror_rate	dst_host_serror_rate	1.00
serror_rate	dst_host_srv_serror_rate	1.00
srv_serror_rate	dst_host_serror_rate	1.00
srv_serror_rate	dst_host_srv_serror_rate	1.00
dst_host_serror_rate	dst_host_srv_serror_rate	1.00
num_compromised	num_root	0.99
rerror_rate	srv_rerror_rate	0.99
rerror_rate	dst_host_rerror_rate	0.99
rerror_rate	dst_host_srv_rerror_rate	0.99

1–10 of 10 rows

Index	Feature name	Description
1	duration	Length of connection
2	protocol type	Type of protocol (TCP, UDP...)
3	service	Destination service (ftp, telnet...)
4	flag	Status of connection
5	source bytes	No. of B from source to destination
6	destination bytes	No. of B from destination to source
7	land	If the source and destination address are the same land=1/if not, then 0
8	wrong fragments	No. of wrong fragments
9	urgent	No. of urgent packets
10	hot	No. of hot indicators
11	failed logins	No. of unsuccessful attempts at login

Target classes

Dataset has five classes for output variable:

- DOS: denial-of-service, e.g. syn flood;
- R2L: unauthorized access from a remote machine, e.g. guessing password;
- U2R: unauthorized access to local superuser (root) privileges, e.g., various ``buffer overflow'' attacks;
- Probing: surveillance and other probing, e.g., port scanning.

Each intrusion is sub-classified into specific type of attack: 24 attack types for training data, and an additional 14 types in the test data only. Further, test data is not from the same probability distribution as the training data, and it includes specific attack types not in the training data.

Skew

The patterns are highly skewed training dataset. From the table below it is evident that 99% of the data belongs to either the Normal or DoS categories.

Class <chr>	Training Set <int>	Percentage <dbl>	Test Set <int>	Percentage (Test) <dbl>
DoS	3883370	79.278	229855	73.901
normal	972781	19.859	60593	19.481
Probe	41102	0.839	4166	1.339
R2L	1126	0.023	16345	5.255
U2R	52	0.001	70	0.023

5 rows

Methodology

- Feature selection
- Training
- Analysis metrics
- Results
- Issues

Preprocessing and Feature Selection

- Preprocessing
  - Convert strings to integers (factors)
  - Remove 2 variables that are all 0’s
- Redundancy
  - redundant examples from training set removed (being pragmatic - computation is slow)
- Irrelevant features removed
  - principal component included where features are highly correlated (dimension reduction)
  - backward stepwise selection used
  - 24 selected, 17 removed
- Mean standardization
  - mean is 0, SD is 1

Training

Objective is to implement machine learning models and establish benchmarks.

Following machine learning models are implemented:

- Support vector machines (radial kernel)
- K-means (500 clusters, majority vote)
- Bayes classifier (Quadratic Discriminant Analysis)
- Logistic Regression

Tuning parameters selected using K-fold cross-validation (k=4)

Multiclass classification

All pairs (one vs one) approach:

Test example is classified using binary classifier on each pair of target class ($_kC_2$ such pairs).

Analysis Metrics

Balanced F-Score (F1-Score):

$$F_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

Weighted F1-Score:

Class <chr>	Training Set <int>	Percentage <dbl>	Test Set <int>	Percentage (Test) <dbl>
DoS	3883370	79.278	229855	73.901
normal	972781	19.859	60593	19.481
Probe	41102	0.839	4166	1.339
R2L	1126	0.023	16345	5.255
U2R	52	0.001	70	0.023

5 rows

Results

F1 Scores

Model	**DoS**	**Normal**	**Probe**	**R2L**	**U2R**	**Weighted F1-Score**
SVM	98	83	75	7	1	90
K-Means	97	85	70	4	0	89
Naive Bayes (QDA)	39	36	76	6	1	37
Logistic Regression	35	35	69	5	1	34

Issues

- Decision boundary not linear (logistic, LDA poor), QDA sightly better
- Resampling needed (ex. bootstrap)