



Data Mining Techniques for Intrusion Detection

Jérôme Froment-Curtil
Bertrand Portier

05/02/2000

Outline

- Introduction
- The KDD Cup 99 Data Set
- The Intrusion Detection Model
- Different Data Mining Methods
- Dynamic Approach
- Conclusion

Introduction

- Increasing numbers of network attacks on the Internet
- Financial and economic well-being of companies are at stake
- Need for a proactive approach to network security instead of current post-attack cleanup

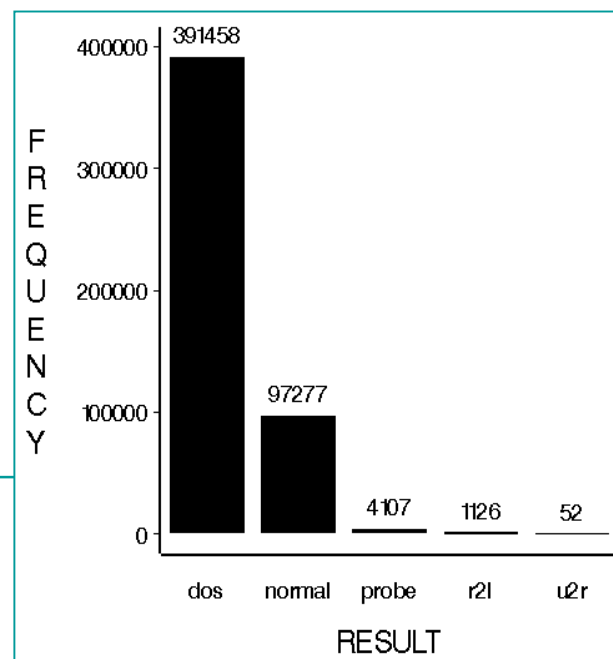
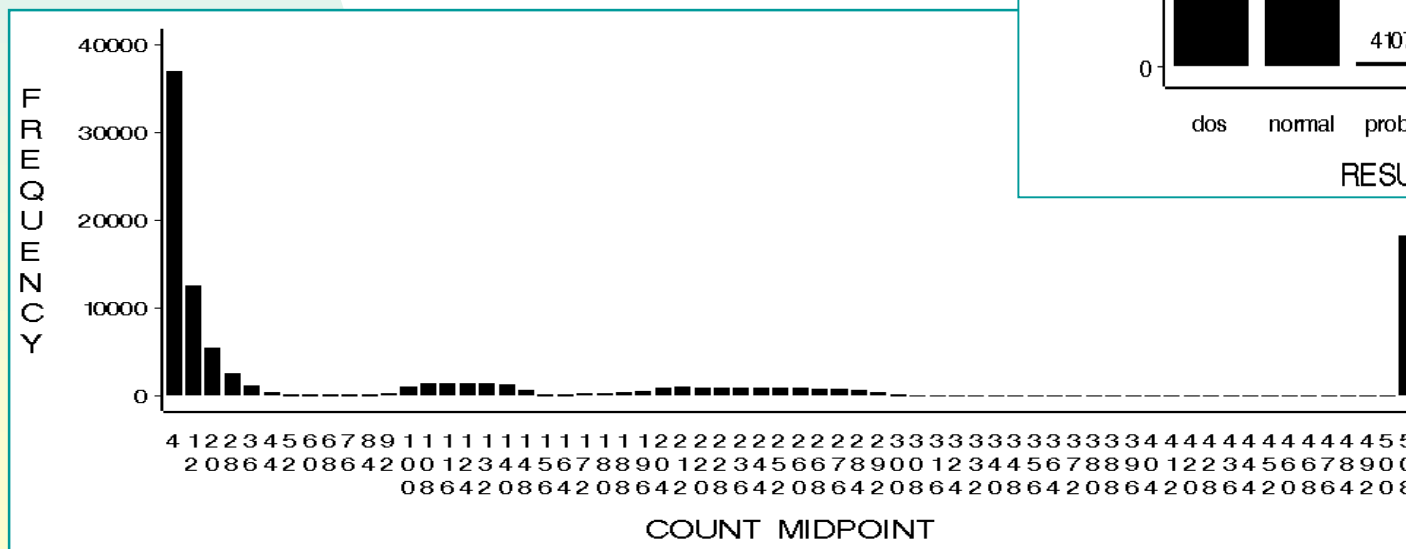
The KDD Cup 99 Data Set

- 744 MB
- 4,940,000 records
- Hardware limitations \Rightarrow work on 10%

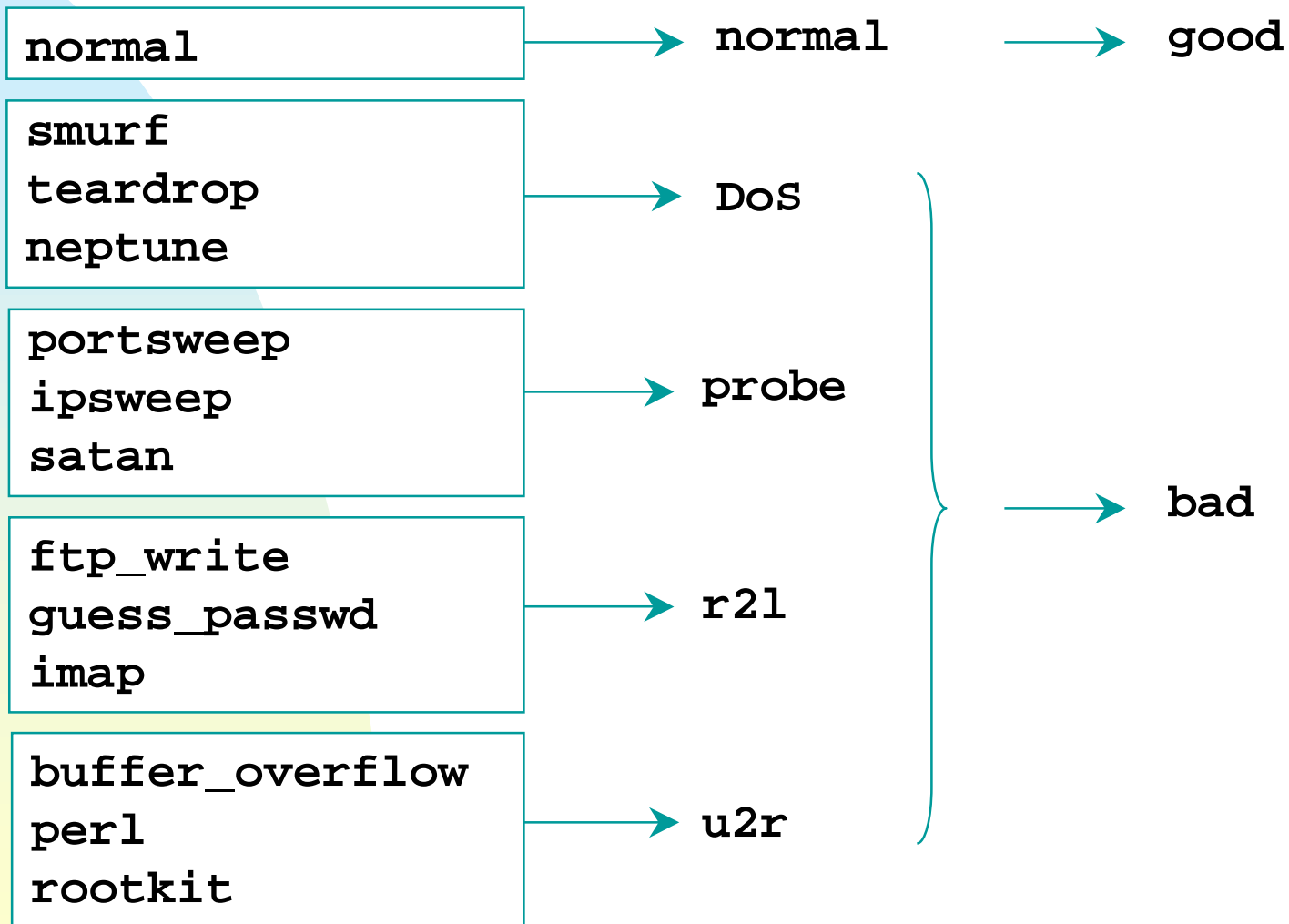
- 41 features per record
 - ◆ `0,tcp,http,SF,181,5450,0,...,1.00,0.00,normal`
- 3 types of features: binary, ordinal, nominal

The KDD Cup 99 Data Set (Cntd.)

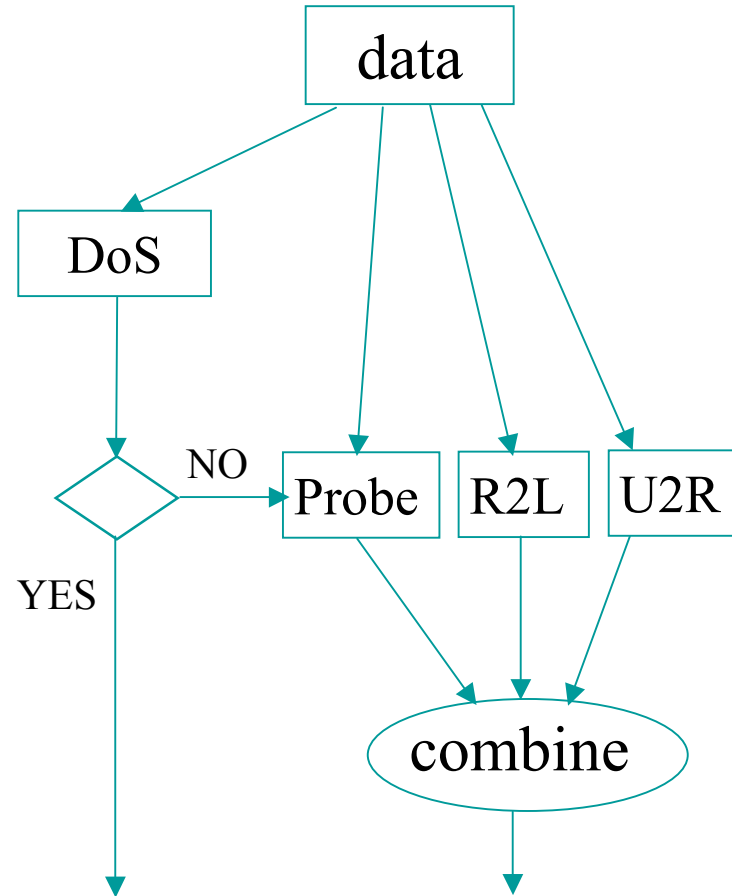
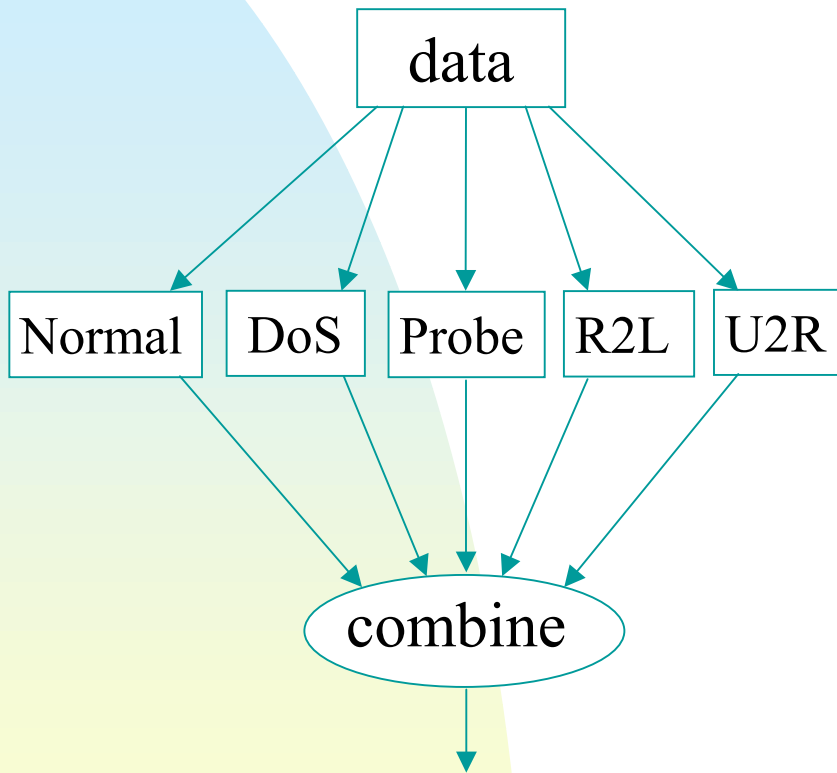
- Sparse data set
- Irregular feature distributions
- Distributions vary with time



Attack Types



The Model

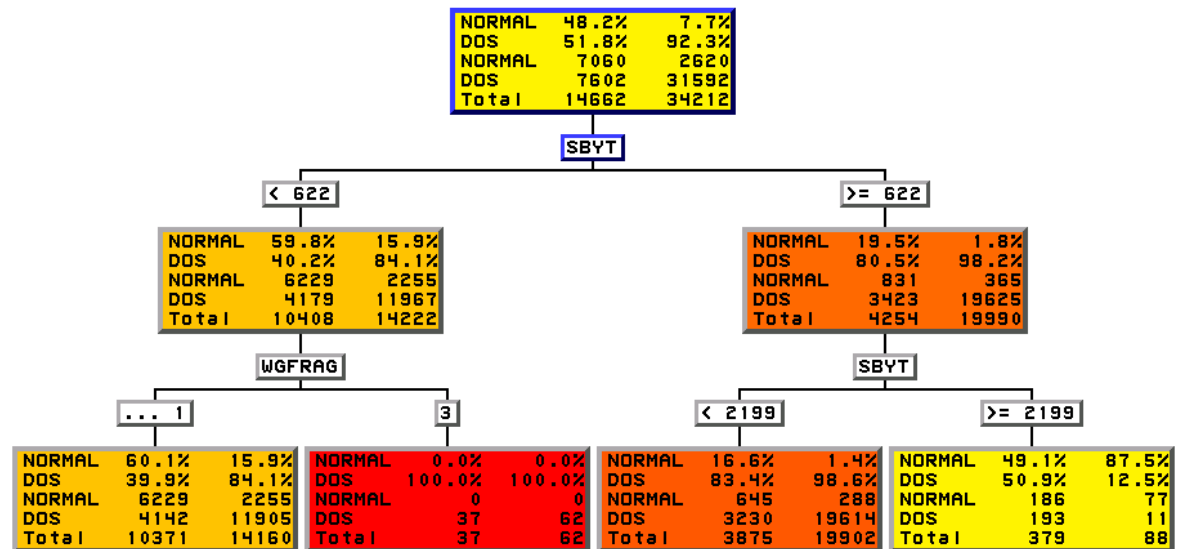


Association Rule Mining

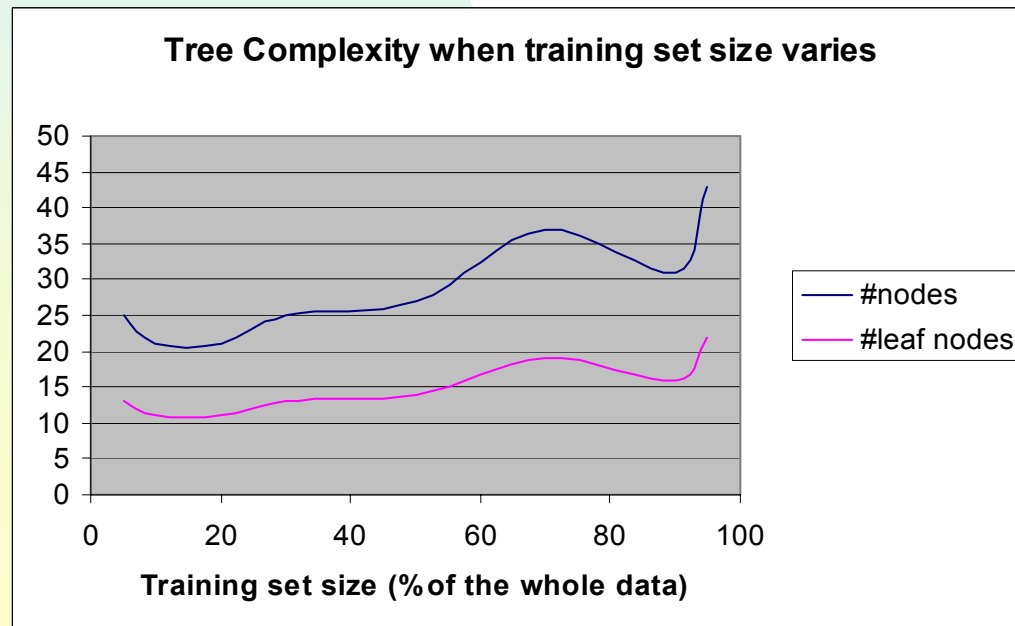
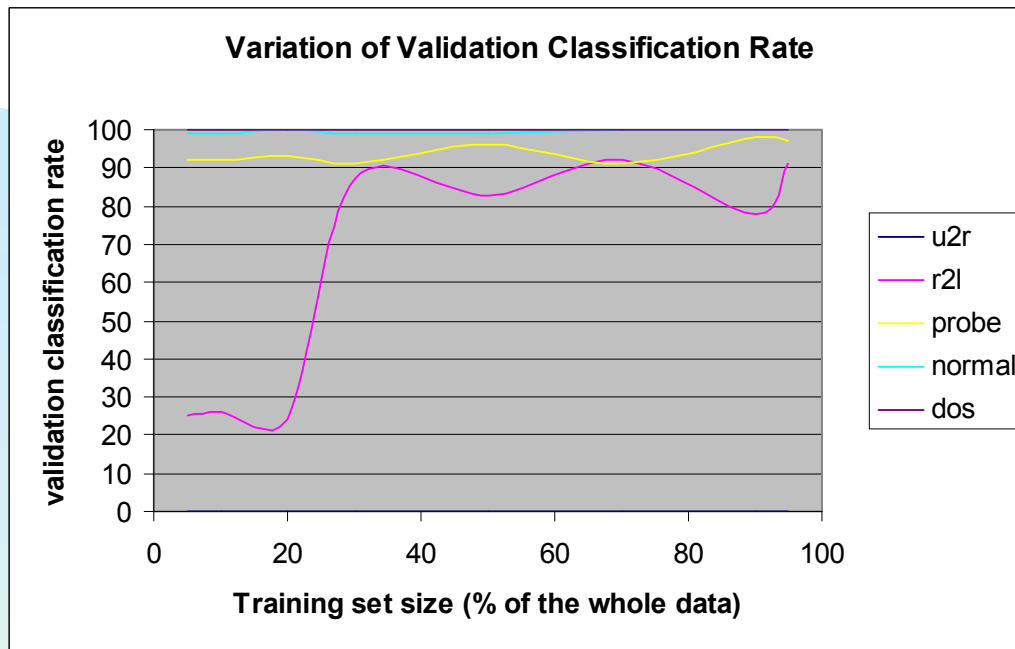
- No timestamp, no client, no ID
- rules:
 - ◆ `sourceByte < 622, service = HTTP ⇒ result = DoS`
- preprocessing to import to SAS EM:
 - ◆ each data record is broken into 41 records, where each record has 3 features: `ID, feature, target`
- 0.05%, reduced feature set, PIII Xeon 256MB

Decision Trees

- Decision Trees, the best method to classify the KDD Cup 99 data.



Decision Trees (Cntd.)

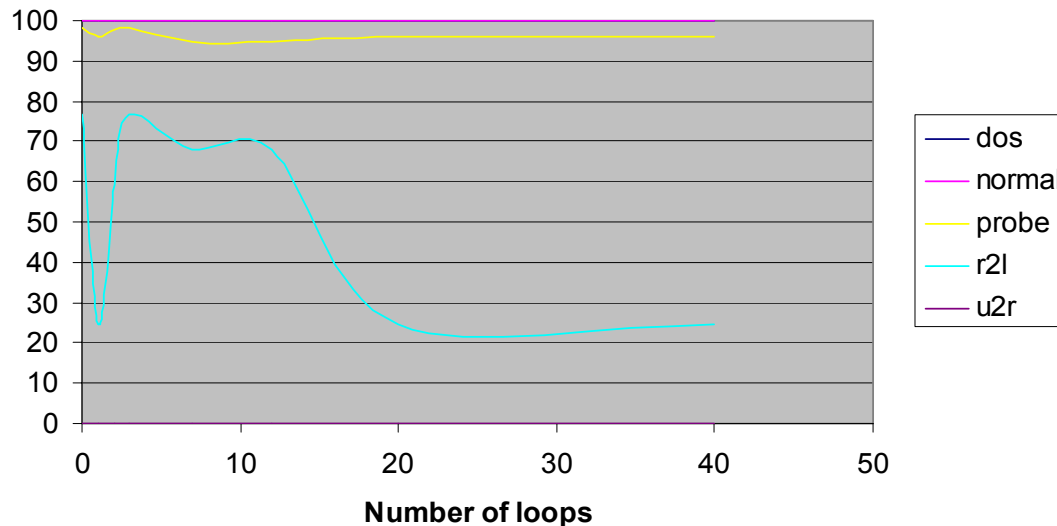


- Good classification rate obtained with a small training set
- No need to increase tree complexity
- Training 40%
Validation 60%

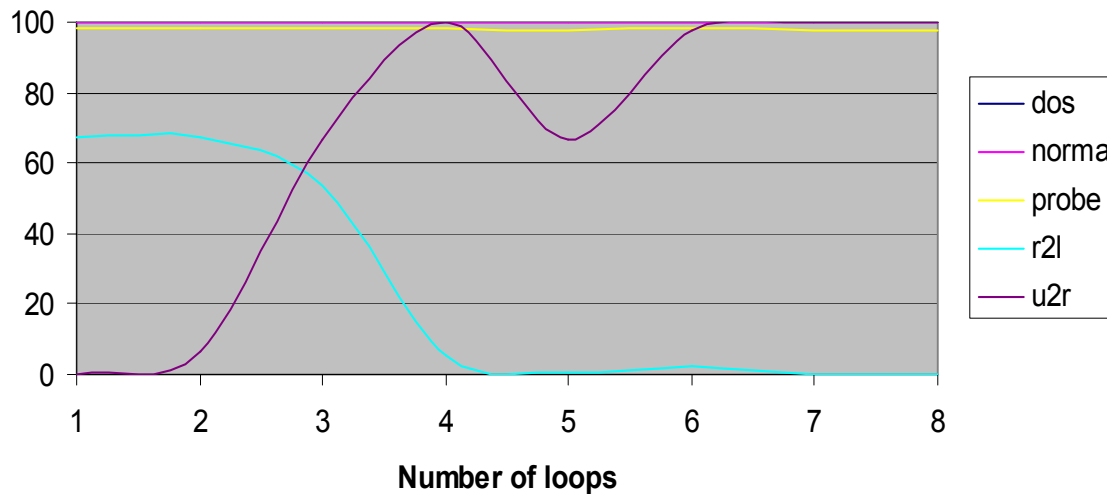
Bagging & Boosting

- Need to modify priors
- U2R = 0.0105%
- Boosting is the only method to properly classify U2R attacks

Bagging Validation Classification Rate



Boosting Validation Classification Rate



Dynamic Approach

- Sliding window technique
 - ◆ Choice of the right window size
 - ◆ Weighting the samples based on time
 - ◆ Dynamically updating the detector
- Detecting new attack types
 - ◆ Experiments on new types of DoS attacks
 - ◆ Classification rates decreases by 19%
 - ◆ More complex trees will not boost the classification rate back to its initial value

Conclusion

- Boosting is the only viable method to detect U2R attacks
- SAS EM is a first step
 - ◆ user has no control of resources
 - ◆ other tools are needed to overcome current limitations
- Next logical step is to integrate our models into one unified detection model