从大规模网络流量中构建重要特征，实现轻量级入侵检测

Outline

Outline

① Background

② Main Method

③ Research Process

④ Conclusion

Background: Intrusion Detection

- 入侵检测是当今网络环境下实现信息保障的防御深度框架中的一项重要技术。

- **基于签名的检测**

  识别不良模式，例如恶意软件

- **基于异常的检测**

  检测与“良好”流量模型的偏差，这往往依赖于机器学习

Anomaly-based Intrusion detection

网络异常入侵检测通常包括三个步骤：特征构建、模型建立和异常检测。

- **Feature Construction**

从网络流量构建的特征对检测至关重要，因为它们描述了网络行为的特征。

Background: KDD’99 data

- 数据挖掘和知识探索

Raw data → Feature extraction analysis / Data preprocessing → KDD’99 data

- Training 5,000,000
- Testing 2,000,000

A network connection with 41 features and a label		
Features(41)	Normal	
单个TCP连接的基本特征（9）	Attack (39)	PROBING
连接中的内容特征（13）		DOS
基于时间的网络流量统计（9）		R2L
基于主机的网络流量统计（10）		U2L

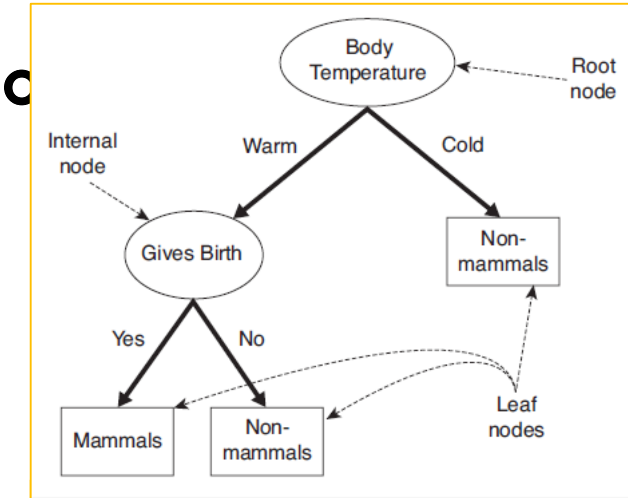
Background: KDD'99 data

- Problem:
- Using all 41 features to detect 4 attack modes, some features for specific attack mode is useless.
- More importantly, 41 features will bring in a lot more parameters, which will spend more time on training and testing

- Solution:
- Reduce feature dimensions

Outline

Information Gain (IG):Feature Selectio



- Definition
  - 减少具有特征X的物品的类别Y的不确定性
- Calculation
  - uncertainty for Y

$$H(Y) = -\sum_i P(y_i)\log_2(P(y_i))$$

  - uncertainty for Y after observing X

$$H(Y|X) = -\sum_j P(x_j)\sum_i P(y_i|x_j)\log_2(P(y_i|x_j))$$

  - the IG of a feature X with respect to Y

$$\mathrm{IG}(Y|X) = H(Y) - H(Y|X)$$

$$\boxed{\mathrm{IG}(Y|X) > \mathrm{IG}(Y|Z)}$$

一个特征X比特征Z与Y类的相关性更高

Decision Trees Classifier (C4.5): intrusion detection

- Definition
  - An extension of Quinlan's earlier ID3 algorithm
- Calculation
  - 给定一个包含标记样本的学习数据集S和一个数据特征X。
  - Si是S的子集，其中特征X有一个值，|S|表示S中的样本数。

$$\mathrm{NIG(S|X)} = \frac{IG(\mathrm{S|X})}{-\sum_i |S_i|/|S| \log_2 (|S_i|/S)}$$

Bayesian networks (BN)

贝叶斯网络是一种概率图形模型，它可以通过有向无环图（DAG）来表示一组随机变量和它们的条件依赖关系 .

例如，贝叶斯网络可以表示数据特征和类（即正常或个别攻击）之间的概率关系。

给定一组特征，网络可以用来计算各种异常行为存在的概率。

Outline

① Background

② Main Method

③ **Research Process**

④ Conclusion

1. Feature Selection

- Using IG to choose 10 features from 41 features

#	Data features	#	Data features	#	Data features	#	Data features
1	duration	12	logged_in	23	count	34	dst_host_same_srv_rate
2	protocol_type	13	num_compromised	24	srv_count	35	dst_host_diff_srv_rate
3	service	14	root_shell	25	serror_rate	36	dst_host_same_src_port_rate
4	flag	15	su_attempted	26	srv_serror_rate	37	dst_host_srv_diff_host_rate
5	src_bytes	16	num_root	27	rerror_rate	38	dst_host_serror_rate
6	dst_bytes	17	num_file_creations	28	srv_rerror_rate	39	dst_host_srv_serror_rate
7	land	18	num_shells	29	same_srv_rate	40	dst_host_rerror_rate
8	wrong_fragment	19	num_access_files	30	diff_srv_rate	41	dst_host_srv_rerror_rate
9	urgent	20	num_outbound_cmds	31	srv_diff_host_rate		
10	hot	21	is_host_login	32	dst_host_count		
11	num_failed_logins	22	is_guest_login	33	dst_host_srv_count		

IG

Table 4 Important features selected for detecting four categories of attacks

Attacks	Features selected
DoS	3, 4, 5, 6, 8, 10, 13, 23, 24, 37
Probe	3, 4, 5, 6, 29, 30, 32, 35, 39, 40
R2L	1, 3, 5, 6, 12, 22, 23, 31, 32, 33
U2R	1, 2, 3, 5, 10, 13, 14, 32, 33, 36

2. Intrusion Detection Schemes

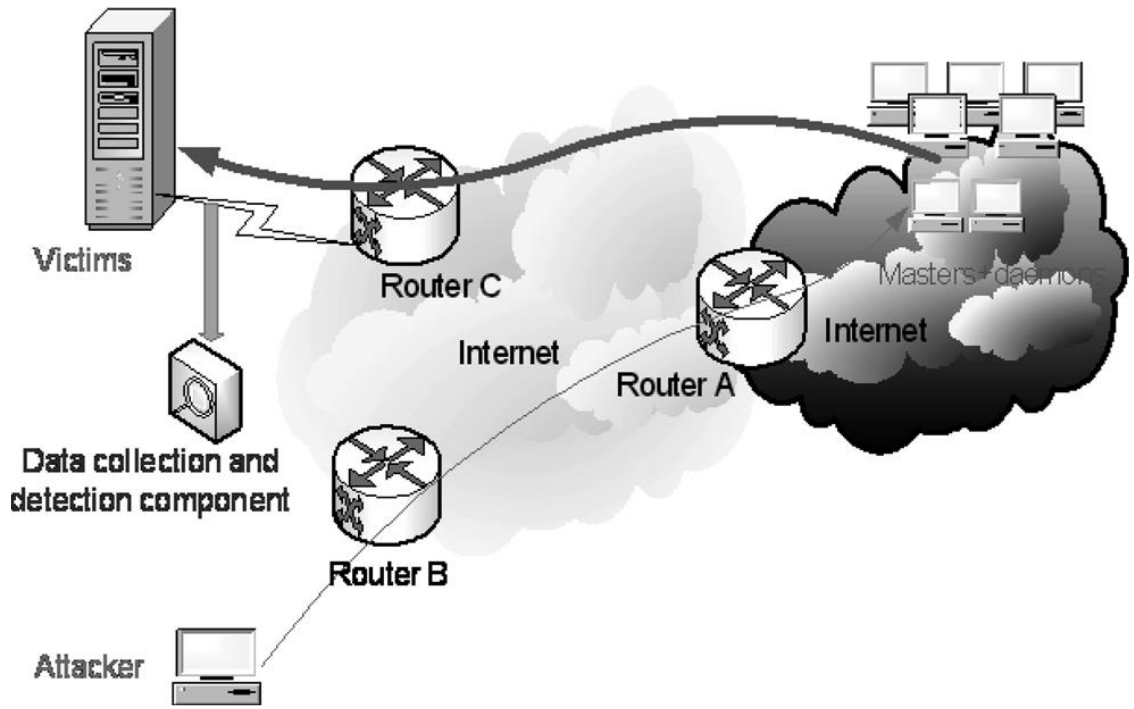
- **Machine learning**
  - Training
  - Testing
- **Main steps**
  - Identify class attributes (features) and classes from training data
  - Identify a subset of the attributes necessary for classification
  - Learn the model using training data
  - Use the trained model to classify the unknown data.

  - Training 4(Attact mode) x 4(2 x BN+2 x C4.5)=16

41 features 10 features

3. Experiments and Analysis

- KDD’99 Data
- Detecting DDoS attacks in real networking



4. Preliminary Results

- Result
  - With fewer features computing time during training and detection can be largely saved

Attack	Method	Using 41 features		Using 10 features	
		Training, s	Test, s	Training, s	Test, s
DoS	BN	4.7	2.1	**0.8**	**0.6**
	C4.5	16.3	1.2	**4.6**	**0.5**
Probe	BN	3.1	2.8	**0.5**	**0.4**
	C4.5	14.5	1.1	**1.2**	**0.3**
R2L	BN	2.6	1.8	**0.5**	**0.4**
	C4.5	10.5	0.8	**0.5**	**0.2**
U2R	BN	2.6	1.8	**0.4**	**0.4**
	C4.5	9.9	0.7	**0.6**	**0.2**

  - The attack detection with 10 important features has the same or even better performance than that with all the 41 features

Attacks	Methods	with 41 features			with 10 features		
		DR, %	FPR, %	F-measure	DR, %	FPR, %	F-measure
DoS	BN	98.73	0.08	0.9927	**99.88**	**0**	**0.9994**
	C4.5	99.96	0.15	0.9980	99.87	**0.14**	0.9977
Probe	BN	92.89	6.08	0.6015	82.93	**3.06**	**0.6874**
	C4.5	82.59	0.04	0.9009	**82.88**	0.05	**0.9017**
R2L	BN	92.22	0.33	0.8535	89.33	**0.32**	0.8408
	C4.5	80.29	0.02	0.8836	**87.34**	**0.01**	**0.9288**
U2R	BN	75.86	0.29	0.2635	65.5	**0.12**	**0.3597**
	C4.5	24.14	0	0.3889	**24.14**	**0**	0.3889

Outline

Conclusion

Selecting 10 features from 41 by Information Gan, the detection efficiency is significantly improved as well as the performance.

It turns out that selecting important features from massive features could help to adapt massive network traffic enviroment.

Thank you !