Machine-Learning-Based Feature Selection Techniques for Large-Scale Network Intrusion Detection

Session 1 , 9:00am-10:00am

O. Y. Al-Jarrah, A. Siddiqui, M. Elsalamouny, P. D. Yoo, S. Muhaidat, K. Kim

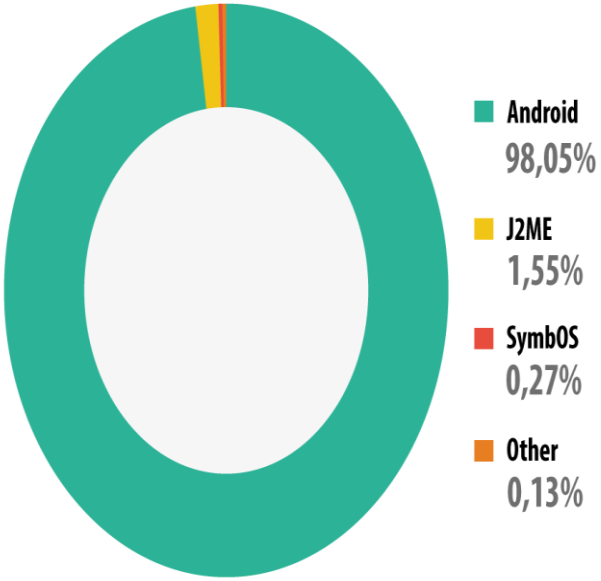
ECE Department, College of Engineering
Khalifa University of Science, Technology and Research (KUSTAR)

Outline

- Motivation
- Methods: FSR/BER Feature Selection
- Experiment & Results
- Conclusion & Future Work

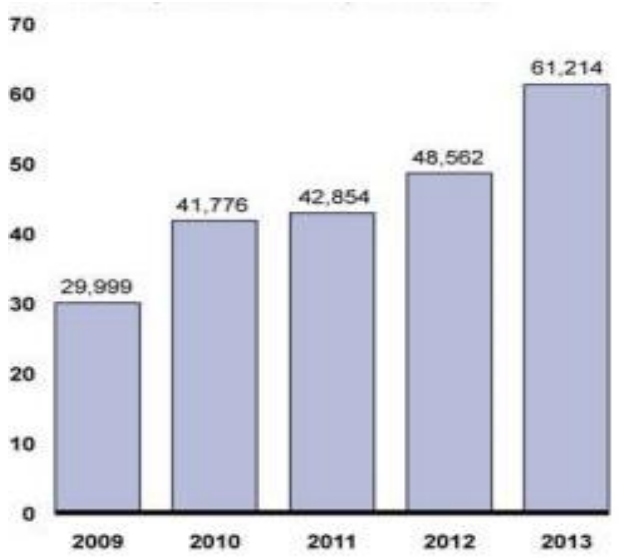
Cyber Attacks Statistics

Mobile malware distribution by platform, 2013



[Source: Kaspersky]

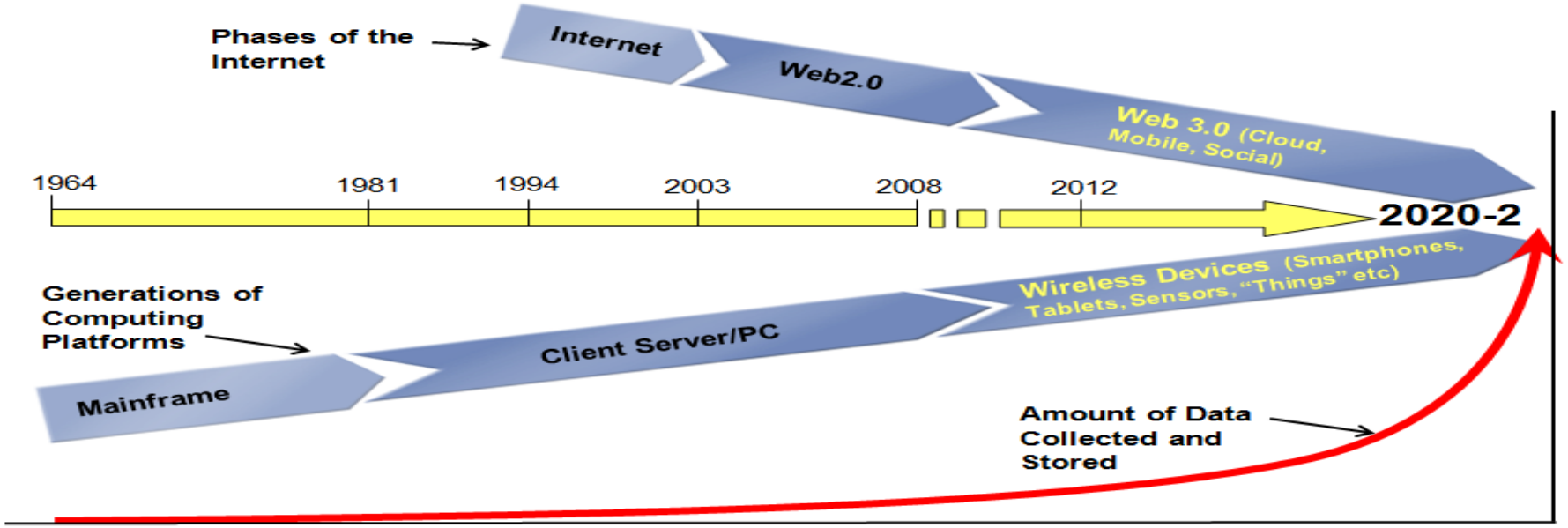
Cyber attacks Incidents Reported by Federal Agencies
No. of Incidents in 1000s



[Source: GAO,US-CERT data]

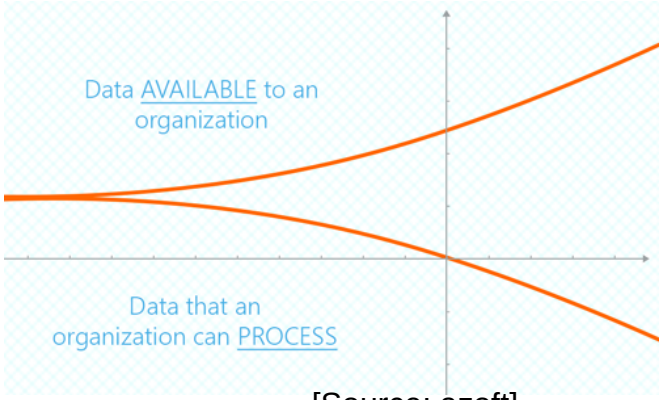
Malicious Cyber Attacks Could Cost U.S. $100B Annually [Source: McAfee]

Large-Scale Data



[Source: Bill Chamberlin]

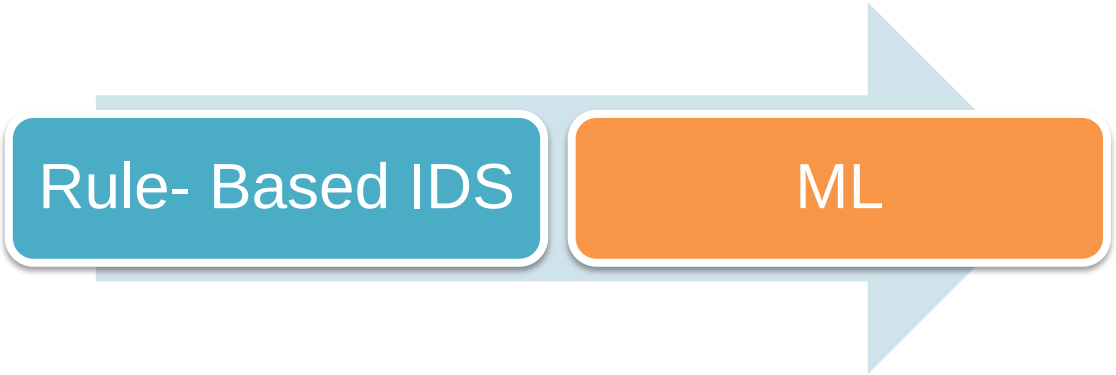
[Source: Thomsonreuters]



[Source: azoft]

Large-Scale IDS

- What is IDS?
- Misuse-based IDS Vs. Anomaly-based IDS



- Supervised Vs. Unsupervised Machine Learning algorithms.
- Issues:
  - Efficiency
  - Scalability
  - Real-time detection

Outline

- Motivation
- Methods: FSR/BER Feature Selection
- Experiment & Results
- Conclusion & Future Work

Dataset

- KDD99 Dataset Vs. NSL-KDD99

Basic characteristics of the KDD 99 intrusion detection dataset in terms of number of samples

Dataset	DoS	Prob	U2R	R2L	Normal
KDD	3,883,370	41,102	52	1126	972,780
10% KDD	391,458	4,107	52	1126	97,277
Corrected KDD	229,853	4,166	70	16,347	60,593

H. G. Kayacik, et al, "Selecting features for intrusion detection: a feature relevance analysis on kdd99 intrusion detection datasets," in Proceedings of the third annual conference on privacy, security and trust, 2005.

Redundant records in the KDD 99 train set

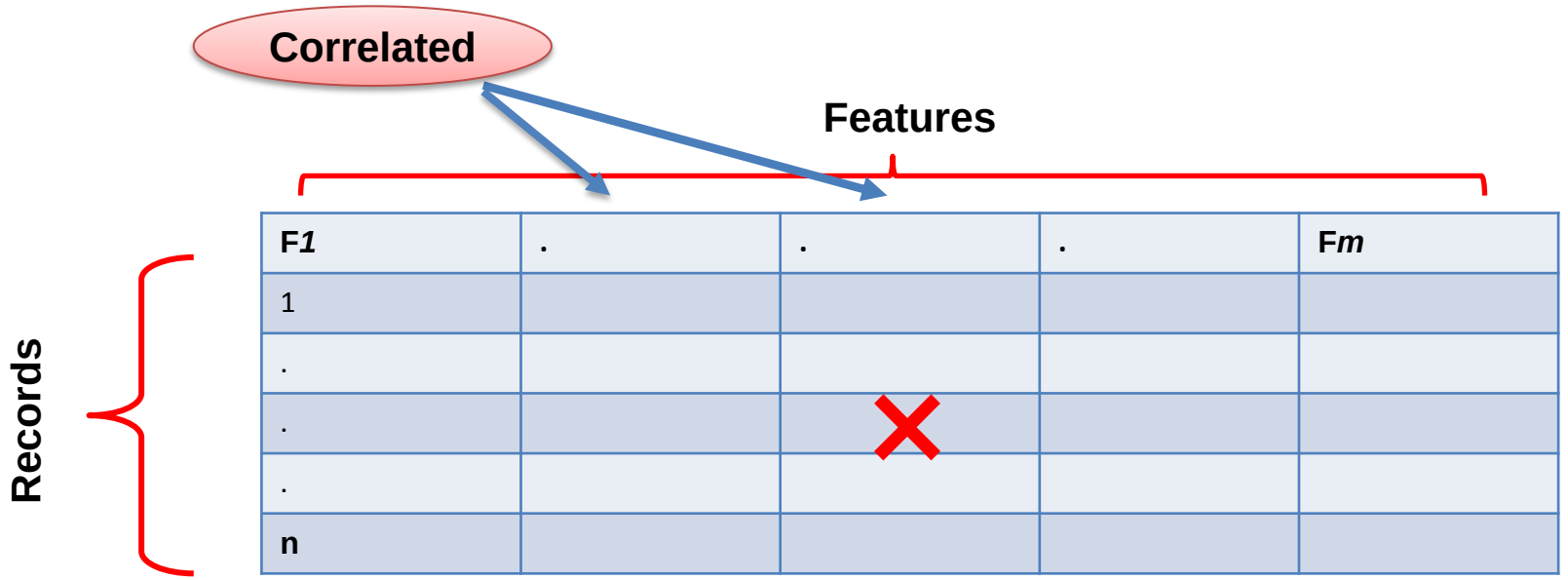
	Original Records	Distinct Records	Reduction Rate
Attacks	3,925,650	262,178	93.32%
Normal	972,780	812,814	16.44%
Total	4,898,431	1,074,992	78.05%

M. Tavallaee , et.al, "A Detailed analysis of the kdd cup 99 data set," in Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defence Applications, 2009.

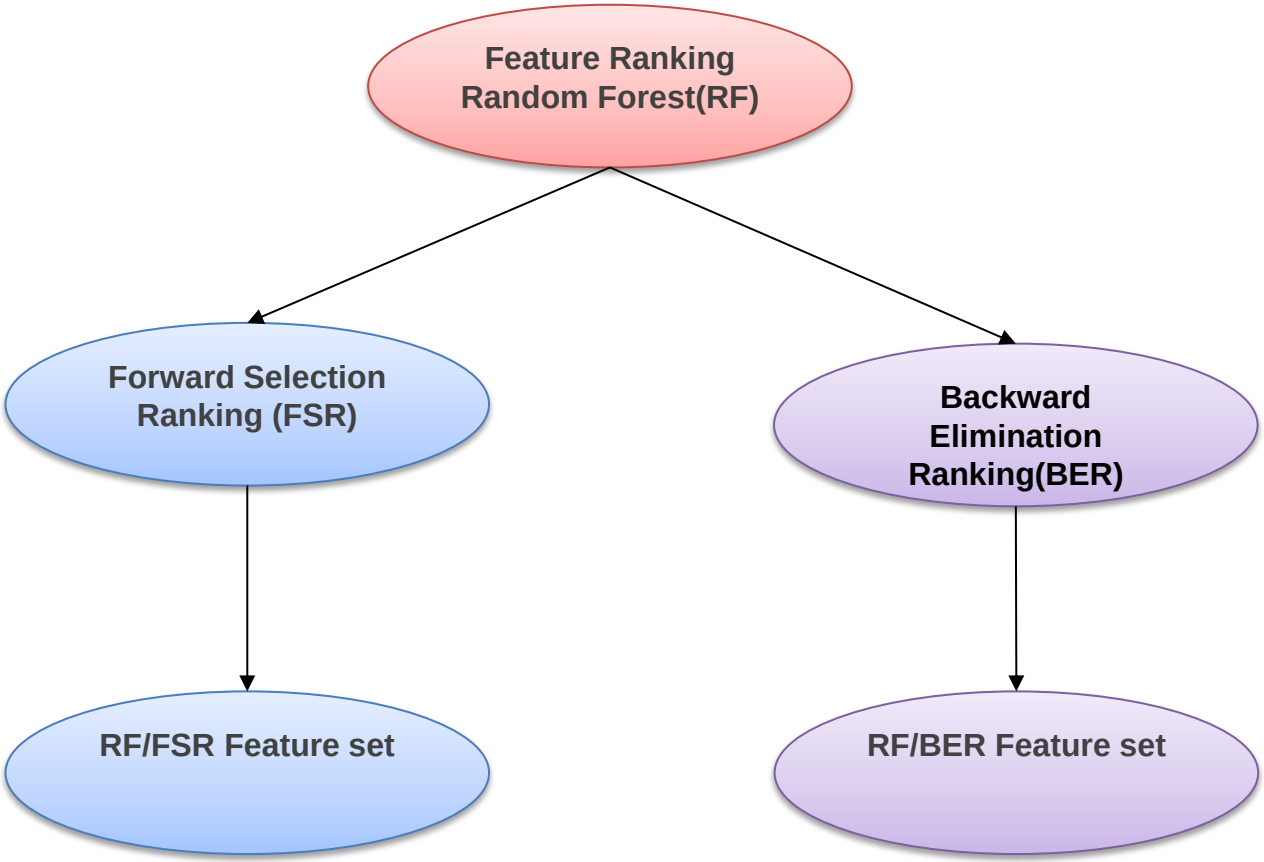
Feature Selection

- Why feature selection?

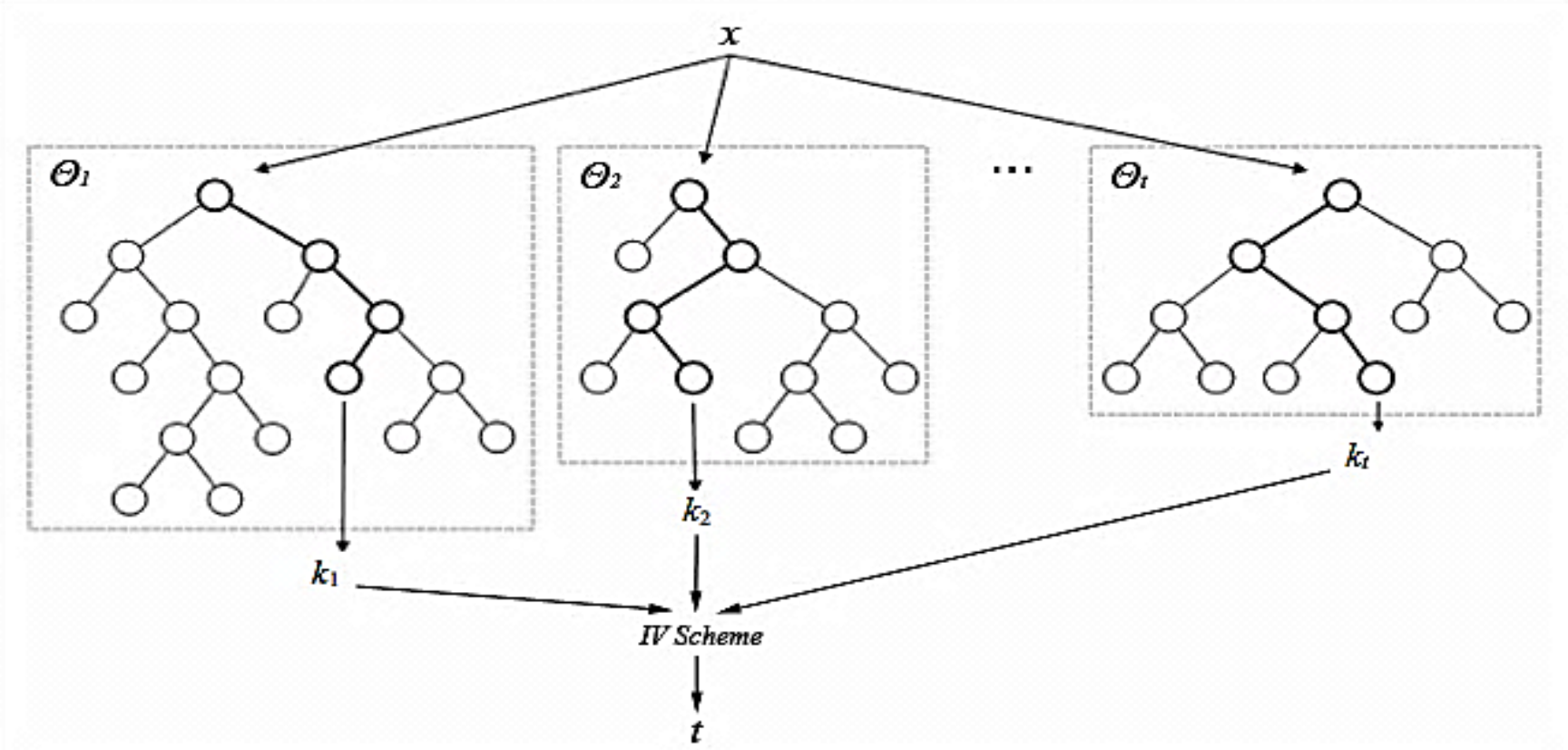
Using all the features of a dataset does not necessarily guarantee the best performances from the IDS. It might increase the computational cost as well as the error rate of the system.



Proposed Methods



Random Forest (RF)



The collection of decision trees (DTs) {h(x, Θ_k), k = 1...}, where the Θ_k are independently, identically distributed random DTs, and each DT casts "a unit vote" for the final classification of input x.

Random Forest (RF) cont.

1. Bagging

- Given a training set $X = x_1, \ldots, x_n$ with responses $Y = y_1$ through y_n,
- For b = 1 through B *(# trees)*: Sample, with replacement, n training examples from X, Y; call these X_b, Y_b.Train a decision or regression tree f_b on X_b, Y_b.

2. Random Subspace Selection

- At each candidate split in the learning process, a random subset of the features is used.

3. Voting

$$P(x') = \frac{1}{B}\sum_{b=1}^{B} P_b(x')$$

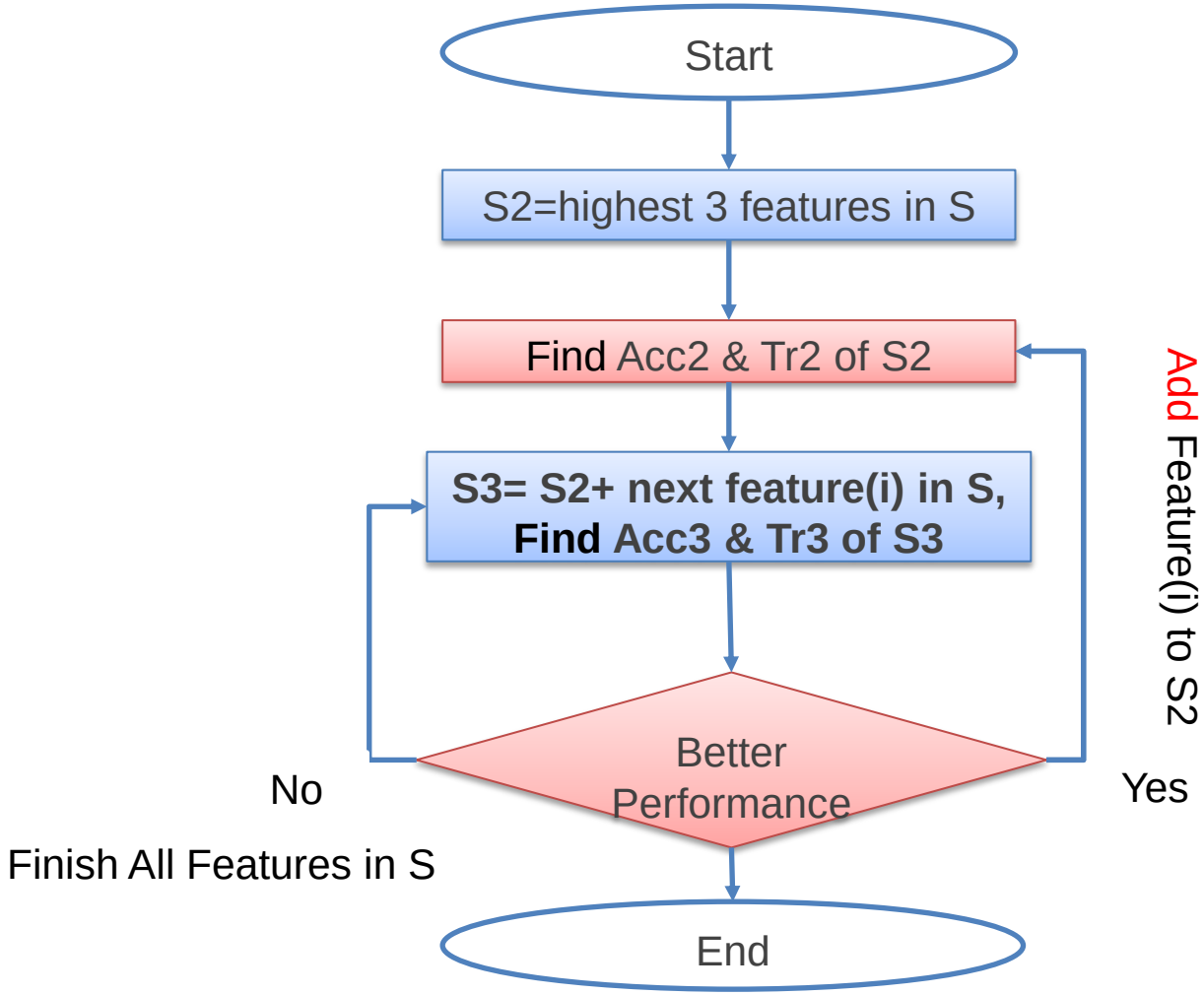
Ranked Features

Feature Importance in Descending Order			
1 .	3	22 .	2
2 .	23	23 .	29
3 .	10	24 .	31
4 .	35	25 .	38
5 .	33	26 .	37
6 .	17	27 .	30
7 .	8	28 .	18
8 .	6	29 .	19
9 .	32	30 .	41
10 .	14	31 .	27
11 .	24	32 .	9
12 .	5	33 .	26
13 .	36	34 .	11
14 .	40	35 .	28
15 .	13	36 .	25
16 .	12	37 .	39
17 .	4	38 .	15
18 .	16	39 .	7
19 .	34	40 .	20
20 .	22	41 .	21
21 .	1		

Feature #	Description
3. Service	Destination service (e.g. telnet, ftp)
23. Count	Number of connections to the same host as the current connection in the past two seconds
10. hot	Numbers of “hot” indicators
20.# outbound cmds	Number of outbound command in ftp session
21.Is hot login	1 if the login belongs to the “hot”, 0 otherwise

J. Zhang, *et al , “Random-Forests-Based network intrusion detection systems,” IEEE Transactions on Systems, Man, and Cybernetics,*2008.

Forward Selection Ranking (FSR)



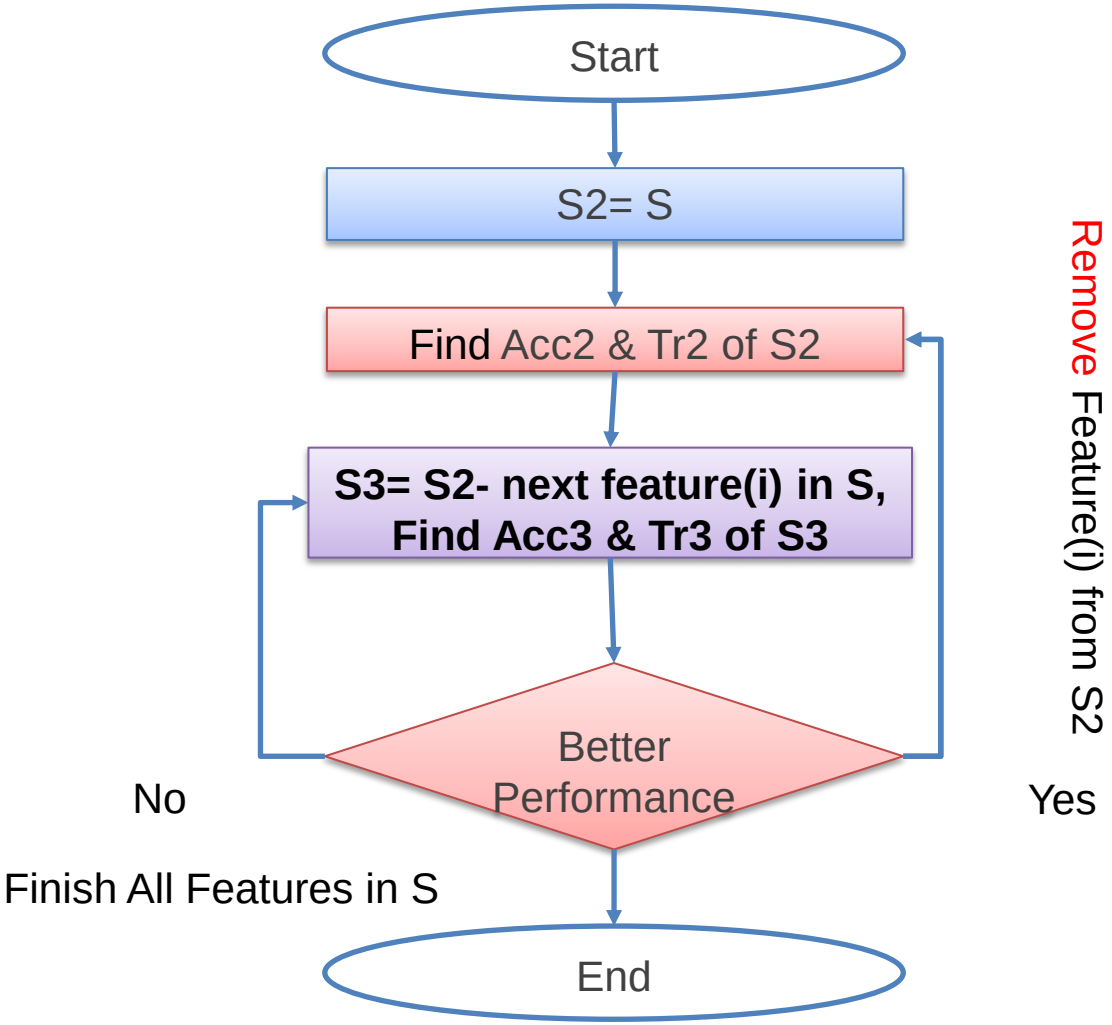
Example: FSR

Feature	Accuracy	Training Time (S)	Remarks
3,10,23	96.38732	4.68	
3,10,23,35	98.37425	6.19	
3,10,23,35,33	98.81165	7.12	
3,10,23,35,33,17	98.80451	7.53	*
3,10,23,35,33,8	98.82276	7.1	**

* Feature 17 is removed because it reduced the accuracy.

** Feature 18 improved the accuracy and the training time.

Backward Elimination Ranking (BER)



Outline

- Motivation
- Methods: FSR/BER Feature Selection
- Experiment & Results
- Conclusion & Future work

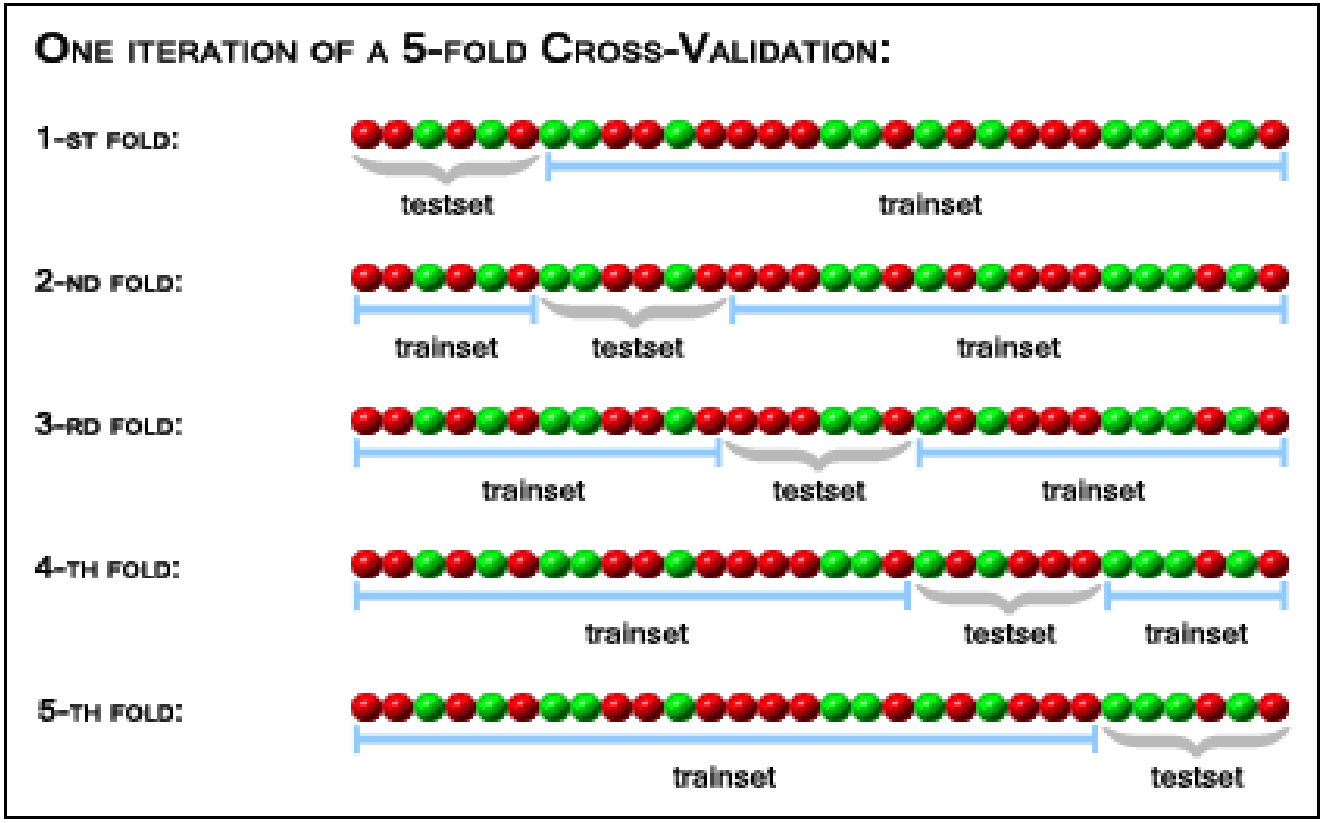
Evaluation

We aim to have a high Acc, Sn, and Mcc while low in Tr and Far.

- $Far = \frac{FP}{TN+FP}$,
- $Acc = \frac{TP+TN}{TP+TN+FP+FN}$,
- $Sn = \frac{TP}{TP+FN}$,
- $Mcc = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$.

Validation

- 10 folds cross-validation technique
- *CVParameterSelection (Weka)*



[Source: genome.turgaz.at]

Feature Set

Feature Sets

Method	Features
RF-FSR	1, 3, 4, 5, 6, 8, 10, 13, 16, 23, 24, 32, 33, 35, 36
RF-BER	1, 2, 3, 5, 6, 10, 14, 16, 32, 33, 36, 37, 38, 41
Kaycik [1]	1, 2, 3, 4, 5, 6, 8, 11, 12, 16, 23, 24, 26, 32, 33
Araújo [2]	2, 3, 5, 6, 9, 11, 12, 14, 22, 30,31, 32, 35, 37
Kantor [3]	1, 2, 3, 4, 5, 6
KDD-99	1–41

1. H. G. Kayacik, et al, "Selecting features for intrusion detection: a feature relevance analysis on kdd99 intrusion detection datasets," in Proceedings of the third annual conference on privacy, security and trust, 2005.
2. N. Araujo, et al, "Identifying important characteristics in the kdd99 intrusion detection dataset by feature selection using a hybrid approach," in IEEE 17th International Conference on Telecommunications (ICT),2010.
3. P. Kantor, et al, "Analysis of three intrusion detection system benchmark datasets using machine learning algorithms," in Intelligence and Security Informatics, Springer – Verlag,2005.

Proposed Methods Performances Vs. Others

Experimental Results

Method	Tr	Sn (DR)	Acc	Mcc	Far
RF/FSR	12.75	**99.857**	**99.901**	**0.99801**	0.000609
RF/BER	11.52	99.833	99.881	0.99761	0.000772
Kaycik [1]	9.76	99.732	99.809	0.99616	0.001247
Araújo [2]	12.23	99.840	99.891	0.99781	0.000639
Kantor [3]	**4.77**	99.499	99.354	0.98702	0.007722
KDD-99	22.09	99.830	99.895	0.99790	**0.000505**

1. H. G. Kayacik, et al, "Selecting features for intrusion detection: a feature relevance analysis on kdd99 intrusion detection datasets," in Proceedings of the third annual conference on privacy, security and trust, 2005.

2. N. Araujo, et al, "Identifying important characteristics in the kdd99 intrusion detection dataset by feature selection using a hybrid approach," in IEEE 17th International Conference on Telecommunications (ICT),2010.

3. P. Kantor, et al, "Analysis of three intrusion detection system benchmark datasets using machine learning algorithms," in Intelligence and Security Informatics, Springer – Verlag,2005.

Outline

- Motivation
- Methods: FSR/BER Feature Selection
- Experiment & Results
- Conclusion & Future work

Conclusion

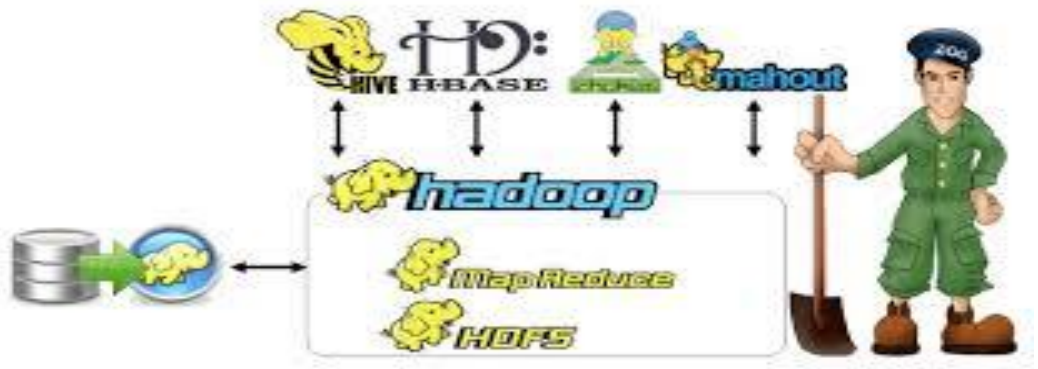
- Two features selection methods, namely, RF-FSR and RF-BER.
- The features selected by the proposed methods were compared with other three popular feature sets on widely known KDD-99 datasets.
- The proposed feature set outperformed other feature sets in the literature.

Future Work

Semi-Supervised Model

Distributed Computing model

Efficiency & Scalability



[Source: bdisys]

Thanks..

Omar.aljarrah@kustar.ac.ae