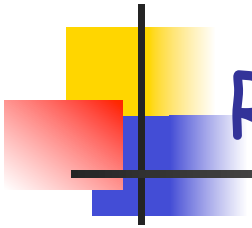
Classification of BGP Anomalies Using Decision Trees and Fuzzy Rough Sets

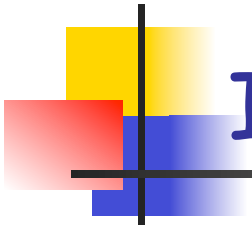
Yan Li, Hong-Jie Xing, Qiang Hua,
Xi-Zhao Wang, Prerna Batta,
Soroush Haeri, and Ljiljana Trajković

Hebei University, Baoding, Hebei, China
Email: {ly, hjxing, huaq, wangxz}@hbu.cn
Simon Fraser University, Vancouver, British Columbia, Canada
Email: {pbatta, shaeri, ljilja}@sfu.ca



Roadmap

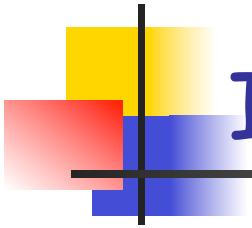
- Introduction
- BGP
  - Update messages
  - Anomalies
  - Features
- Feature selection
- Anomaly classifiers
- Conclusion
- References



Introduction

Anomaly detection techniques are employed to detect Internet traffic anomalies

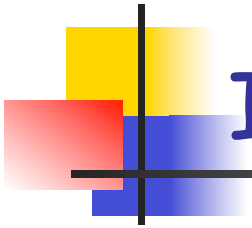
- Classification problem:
  - assigning an "anomaly" or "regular" label to a data point
- Accuracy of a classifier depends on:
  - extracted features
  - combination of selected features
  - underlying model



Introduction

Goal:

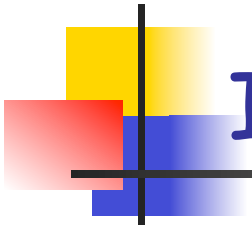
- Detect Internet routing anomalies using the Border Gateway Protocol (BGP) update messages



Introduction

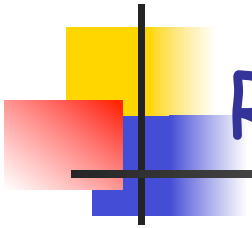
Approach:

- Define a set of 37 features based on BGP update messages
- Extract the features from available BGP update messages that are collected during the time period when the Internet experienced anomalies:
  - Slammer
  - Nimda
  - Code Red I



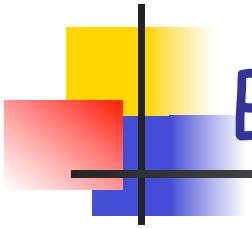
Introduction

- Select the most relevant features for classification using:
  - Decision Tree
  - Fuzzy Rough Sets
- Train classifiers for BGP anomaly detection using:
  - Decision Tree
  - Extreme Learning Machine (ELM)



Roadmap

- Introduction
- BGP
  - Update messages
  - Anomalies
  - Features
- Feature selection
- Anomaly classifiers
- Conclusion
- References



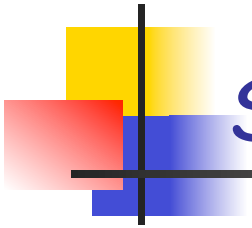
BGP: update messages

- Border Gateway Protocol (BGP) enables exchange of routing information between gateway routers using update messages
- BGP update message collections:
  - Réseaux IP Européens (RIPE) under the Routing Information Service (RIS) project
  - Route Views
  - Available in multi-threaded routing toolkit (MRT) binary format

BGP: anomalies

Anomaly	Date	Duration (h)
Slammer	January 25, 2003	16
Nimda	September 18, 2001	59
Code Red I	July 19, 2001	10

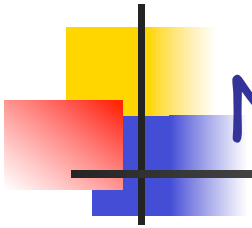
Training Data	Dataset
Slammer + Nimda	Dataset 1
Slammer + Code Red I	Dataset 2
Code Red I + Nimda	Dataset 3
Slammer	Dataset 4
Nimda	Dataset 5
Code Red I	Dataset 6



Slammer worm

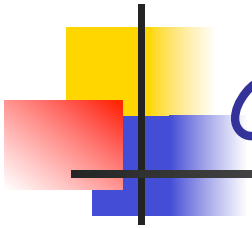
- Sends its replica to randomly generated IP addresses
- Destination IP address gets infected if:
  - it is a Microsoft SQL server

  or
  - a Personal Computer with the Microsoft SQL Server Data Engine (MSDE)



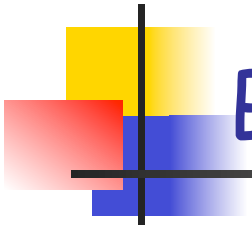
Nimda worm

- Propagates through email messages, web browsers, and file systems
- Viewing the email message triggers the worm payload
- The worm modifies the content of the web document files in the infected hosts and copies itself in all local host directories



Code Red I worm

- Takes advantage of vulnerability in the Microsoft Internet Information Services (IIS) indexing software
- Triggers a buffer overflow in the infected hosts by writing to the buffers without checking their limit



BGP: features

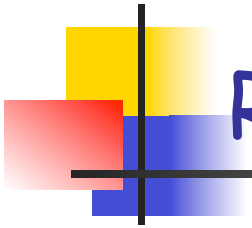
- Define 37 features
- Sample every minute during a five-day period
  - the peak day of an anomaly
  - two days prior and two days after the peak day
- 7,200 samples for each anomalous event
  - 5,760 regular samples (non-anomalous)
  - 1,440 anomalous samples
  - imbalanced dataset

BGP features

Feature	Definition	Category
1	Number of announcements	Volume
2	Number of withdrawals	Volume
3	Number of announced NLRI prefixes	Volume
4	Number of withdrawn NLRI prefixes	Volume
5	Average AS-PATH length	AS-path
6	Maximum AS-PATH length	AS-path
7	Average unique AS-PATH length	AS-path
8	Number of duplicate announcements	Volume
9	Number of duplicate withdrawals	Volume
10	Number of implicit withdrawals	Volume

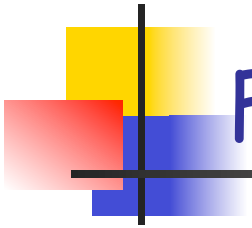
BGP features

Feature	Definition	Category
11	Average edit distance	AS-path
12	Maximum edit distance	AS-path
13	Inter-arrival time	Volume
14–24	Maximum edit distance = n, where n = (7, ..., 17)	AS-path
25–33	Maximum AS-path length = n, where n = (7, ..., 15)	AS-path
34	Number of IGP packets	Volume
35	Number of EGP packets	Volume
36	Number of incomplete packets	Volume
37	Packet size (B)	Volume



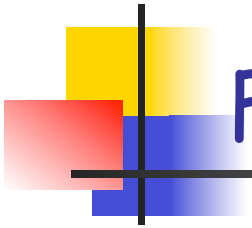
Roadmap

- Introduction
- BGP
  - Update messages
  - Anomalies
  - Features
- Feature selection
- Anomaly classifiers
- Conclusion
- References



Feature selection algorithms

- Misclassification occurs due to:
  - Redundancy or noise contained in datasets
- Feature selection algorithms may be employed to select most relevant features:
  - Fisher
  - Minimum Redundancy Maximum Relevance (mRMR)
  - Odds Ratio
  - Decision Tree
  - Fuzzy Rough Sets



Feature selection: decision tree

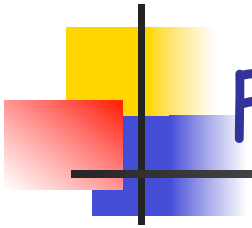
- Commonly used algorithm in data mining
- Generates a model that predicts the value of a target variable based on several input variables
- A top-down approach is commonly used for constructing decision trees:
  - an appropriate variable is chosen to best split the set of items based on homogeneity of the target variable within subsets
- C5 software tool was used to generate decision trees

C5 [Online]. Available: http://www.rulequest.com/see5-info.html.

Feature selection: decision tree

Dataset	Training data	Selected Features
Dataset 1	Slammer + Nimda	1-21, 23-29, 34-37
Dataset 2	Slammer + Code Red I	1-22, 24-29, 34-37
Dataset 3	Code Red I + Nimda	1-29, 34-37

- Either four (30, 31, 32, 33) or five (22, 30, 31, 32, 33) features are removed in the constructed trees mainly because:
  - features are numerical and some are used repeatedly



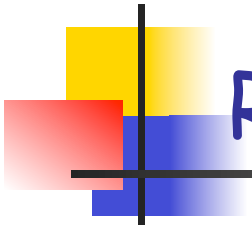
Feature selection: fuzzy rough sets

- Deal with the approximation of fuzzy sets in a fuzzy approximation space defined by a fuzzy similarity relation or by a fuzzy partition
- The fuzzy similarity relation *Sim(C)* is:
  - an *nxn* matrix that describes similarities between any two samples
  - computed by the min operator
- Computational complexity: $O(n^2m)$
  - *n* is the number of samples
  - *m* is the number of features

Feature selection: fuzzy rough sets

Dataset	Training data	Selected Features
Dataset 4	Slammer	1, 3–6, 9, 10, 13–32, 35
Dataset 5	Nimda	1, 3–4, 8–10, 12, 14–32, 35, 36
Dataset 6	Code Red I	3–4, 8–10, 12, 14–32, 35, 36

- Using combination of datasets, for example Slammer + Nimda for training leads to higher computational load
- Each dataset was used individually



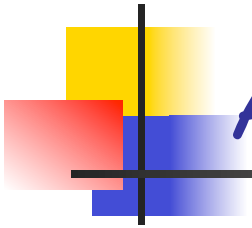
Roadmap

- Introduction
- BGP
  - Update messages
  - Anomalies
  - Features
- Feature selection
- Anomaly classifiers
- Conclusion
- References

Anomaly classifiers: decision tree

Dataset	Testing data	Acc_{train}	Acc_{test}	Training time (s)
Dataset 1	Code Red I	90.7	78.8	1.8
Dataset 2	Nimda	92.3	72.8	2.1
Dataset 3	Slammer	87.1	23.8	2.3

- Each path from the root node to a leaf node may be transformed into a decision rule
- A set of rules that are obtained from a trained decision tree may be used for classifying unseen samples



Anomaly classifier: ELM

- Used for learning with a single hidden layer feed forward neural network
- Weights connecting the input and hidden layers with the bias terms are initialized randomly
- Weights connecting the hidden and output layers are analytically determined
- Learns faster than SVMs by a factor of thousands
- Suitable for online applications
- We use all the features (37), all continuous features (17), features selected by fuzzy rough sets (28 or 27), and continuous features selected by fuzzy rough sets (9 or 8)

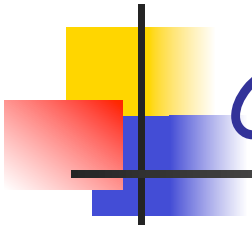
Anomaly classifiers: ELM

No. of features	Dataset	Acc_{train}	Acc_{test}	Training time (s)
37	Dataset 1	83.57 ± 0.11	80.01 ± 0.07	2.3043
	Dataset 2	83.53 ± 0.12	79.75 ± 0.08	2.2756
	Dataset 3	80.82 ± 0.09	21.65 ± 1.93	2.2747
17	Dataset 1	84.50 ± 0.07	79.91 ± 0.01	1.9268
	Dataset 2	84.43 ± 0.12	79.53 ± 0.10	1.5928
	Dataset 3	83.06 ± 0.07	51.56 ± 16.38	1.8882

- 195 hidden units
- The binary features 14–33 are removed to form a set of 17 features

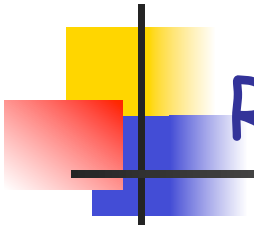
Anomaly classifiers: ELM

No. of features	Dataset	Acc_{train}	Acc_{test}
28	Dataset 4	83.08 ± 0.11	80.03 ± 0.06
28 (from 37)	Dataset 5	83.08 ± 0.09	79.78 ± 0.07
27	Dataset 6	80.05 ± 0.00	81.00 ± 1.41
9	Dataset 4	84.59 ± 0.07	80.00 ± 0.05
9 (from 17)	Dataset 5	84.25 ± 0.11	79.79 ± 0.12
8	Dataset 6	83.38 ± 0.04	49.24 ± 12.90



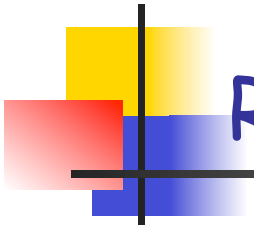
Conclusions

- Machine learning algorithms (feature selection and classification algorithms) are used for detecting BGP anomalies
- Performance of classifiers greatly depended on the employed datasets
- Feature selection algorithms were used to improve the performance of classifiers
- For smaller datasets, performance of the ELM classifier was improved by using fuzzy rough sets
- Both decision tree and ELM are relatively fast classifiers with satisfactory accuracy



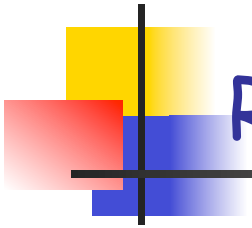
References

- N. Al-Rousan, S. Haeri, and Lj. Trajković, "Feature selection for classification of BGP anomalies using Bayesian models," in *Proc. ICMLC 2012*, Xi'an, China, July 2012, pp. 140-147.
- N. Al-Rousan and Lj. Trajkovic, "Machine learning models for classification of BGP anomalies," in *Proc. IEEE Conf. High Performance Switching and Routing, HPSR 2012*, Belgrade, Serbia, June 2012, pp. 103-108.
- S. Deshpande, M. Thottan, T. K. Ho, and B. Sikdar, "An online mechanism for BGP instability detection and analysis," *IEEE Trans. Computers*, vol. 58, no. 11. 1470–1484, Nov. 2009.
- T. Ahmed, B. Oreshkin, and M. Coates, "Machine learning approaches to network anomaly detection," in Proc. *USENIX Workshop on Tackling Computer Systems Problems with Machine Learning Techniques*, Cambridge, MA, USA, May 2007, pp. 1–6.
- J. Li, D. Dou, Z. Wu, S. Kim, and V. Agarwal, "An Internet routing forensics framework for discovering rules of abnormal BGP events," *SIGCOMM Comput. Commun. Rev.*, vol. 35, no. 4, pp. 55–66, Oct. 2005.



References

- RIPE RIS raw data [Online]. Available: http://www.ripe.net/data-tools/.
- University of Oregon Route Views project [Online]. Available: http://www.routeviews.org/.



References

- J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, Mar. 1986.
- Z. Pawlak, "Rough sets," *Int. J. Inform. and Comput. Sciences*, vol. 11, no. 5, pp. 341–356, Oct. 1982.
- G. B. Huang, Q. Y. Zhu, and C. K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, pp. 489–501, Dec. 2006.
- G. B. Huang, X. J. Ding, and H. M. Zhou, "Optimization method based extreme learning machine for classification," *Neurocomputing*, vol. 74, no. 1-3, pp. 155–163, Dec. 2010.