Comparison of Machine Learning Models for Classification of BGP Anomalies

Nabil M. Al-Rousan
Simon Fraser University
Vancouver, British Columbia, Canada
Email: nalrousa@sfu.ca
http://www.sfu.ca/~ljilja/cnl

August 7, 2012

Roadmap

Introduction

- Slammer, Nimda, and Code Red I anomalies affect performance of the Internet Border Gateway Protocol (BGP)
- BGP anomalies also include: Internet Protocol (IP) prefix hijacks, miss-configurations, and electrical failures
- BGP anomalies often occur
- Techniques for BGP anomalies detection have recently gained visible attention and importance

Contribution

- We introduce new BGP features to design anomaly detection mechanisms by applying:
  - Support Vector Machine (SVM) models
  - Hidden Markov Models (HMMs)
  - Naive Bayes (NB)
- The proposed models are tested with collected BGP traffic traces and are employed to successfully classify and detect various BGP anomalies
- We apply multi-classification models to correctly classify test datasets and identify the correct anomaly types
- Graphical user interface tool (BGPAD) is built to classify BGP anomalies for BGP datasets

Roadmap

Datasets sources

- The RIPE and Route Views BGP update messages: multi-threaded routing toolkit (MRT) binary format
- Validity of the proposed models was checked by also using BGP traffic trace collected from the BCNET

	Class	Date	Duration (h)
Slammer	Anomaly	January 25, 2003	16
Nimda	Anomaly	September 18, 2001	59
Code Red I	Anomaly	July 19, 2001	10
RIPE regular	Regular	July 14, 2001	24
BCNET	Regular	December 20, 2011	24

References

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

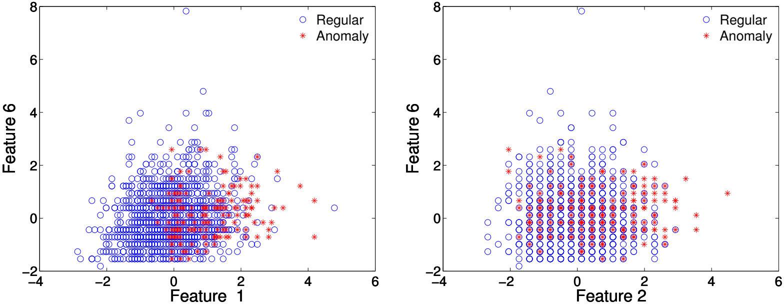
List of extracted features

- Extracted features: volume and AS-path 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4	Number of withdrawn NLRI prefixes	volume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
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

Normalized scattering graphs

- Feature 1, feature 2, and feature 6:



- Selecting appropriate combination of features is essential for a an accurate classification

Feature selection algorithms

- Features scoring algorithms:
  - Fisher
  - minimum Redundancy Maximum Relevance (mRMR)
  - odds Ratio
- These algorithms measure the correlation and relevancy among features
- The top ten features were selected

References

- Y.-W. Chen and C.-J. Lin, "Combining SVMs with various feature selection strategies," *Strategies*, vol. 324, no. 1, pp. 1–10, Nov. 2006.
- H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

Fisher algorithm

- Training datasets: a real matrix $\mathbf{X}_{7200\times 37}$.
- Column vector $\mathbf{X}_k, k = 1, ..., 37$ corresponds to one feature
- The Fisher score for $\mathbf{X}_k$:

$$\text{F-score} = \frac{m_a^2 - m_r^2}{s_a^2 + s_r^2}$$

$$= \frac{\frac{1}{N_a}\sum\limits_{i\in anomaly} x_{ik}{}^2 - \frac{1}{N_r}\sum\limits_{i\in regular} x_{ik}{}^2}{\frac{1}{N_a}\sum\limits_{i\in anomaly} (x_{ik} - m_a)^2 + \frac{1}{N_r}\sum\limits_{i\in regular} (x_{ik} - m_r)^2}$$

- N_a and N_r: number of anomaly and regular data points
- m_a and s_a^2 (m_r and s_r^2): the mean and the variance of anomaly (regular) class

Fisher algorithm

- Fisher algorithm: maximizes the inter-class separation $m_a^2 - m_r^2$ and minimizes the intra-class variances s_a^2 and s_r^2
- mRMR algorithm: maximizes the relevance of features with respect to the target class while minimizing the redundancy among features
- Variants of the mRMR algorithm:
  - Mutual Information Difference (MID)
  - Mutual Information Quotient (MIQ)
  - Mutual Information Base (MIBASE)

mRMR algorithm

- mRMR relevance between a feature set $S = \{\mathbf{X}_1, ..., \mathbf{X}_k, \mathbf{X}_l, ..., \mathbf{X}_{37}\}$ and a class vector $\mathbf{Y}$ is based on the mutual information function $\mathcal{I}$:

$$\mathcal{I}(\mathbf{X}_k, \mathbf{X}_l) = \sum_{k,l} p(\mathbf{X}_k, \mathbf{X}_l) log \frac{p(\mathbf{X}_k, \mathbf{X}_l)}{p(\mathbf{X}_k)p(\mathbf{X}_l)}$$

- Criteria for mRMR variants:
  - MID: $\max\left[V(\mathcal{I}) - W(\mathcal{I})\right]$
  - MIQ: $\max\left[V(\mathcal{I})/W(\mathcal{I})\right]$

$$V(\mathcal{I}) = \frac{1}{|S|} \sum_{\mathbf{X}_k \in S} \mathcal{I}(\mathbf{X}_k, \mathbf{Y}), \qquad W(\mathcal{I}) = \frac{1}{|S|^2} \sum_{\mathbf{X}_k, \mathbf{X}_l \in S} \mathcal{I}(\mathbf{X}_k, \mathbf{X}_l)$$

  - MIBASE: ordered based on the $\mathcal{I}(X_k, X_l)$ function

Odds ratio algorithm

- Performs well for feature selection in binary classification with NB classifiers
- Computed as:

$$OR(\mathbf{X}_k) = \log \frac{\Pr(\mathbf{X}_k|c)\big(1 - \Pr(\mathbf{X}_k|\bar{c})\big)}{\Pr(\mathbf{X}_k|\bar{c})\big(1 - \Pr(\mathbf{X}_k|c)\big)},$$

where $\Pr(\mathbf{X}_k|c)$ and $\Pr(\mathbf{X}_k|\bar{c})$ are the probabilities of feature $\mathbf{X}_k$ being in classes c and $\bar{c}$, respectively.

EOR, WOR, MOR, and CDM Algorithms

- The odds ratio (OR), extended odds ratio (EOR), weighted odds ratio (WOR), multi-class odds ratio (MOR), and class discriminating measure (CDM) are variants that enable feature selection for multi-class problems:

$$EOR(\mathbf{X}_k) = \sum_{j=1}^{J} \log \frac{\Pr(\mathbf{X}_k|c_j)(1 - \Pr(\mathbf{X}_k|\bar{c}_j))}{\Pr(\mathbf{X}_k|\bar{c}_j)(1 - \Pr(\mathbf{X}_k|c_j))}$$

$$WOR(\mathbf{X}_k) = \sum_{j=1}^{J} \Pr(c_j) \times \log \frac{\Pr(\mathbf{X}_k|c_j)(1 - \Pr(\mathbf{X}_k|\bar{c}_j))}{\Pr(\mathbf{X}_k|\bar{c}_j)(1 - \Pr(\mathbf{X}_k|c_j))}$$

$$MOR(\mathbf{X}_k) = \sum_{j=1}^{J} \left| \log \frac{\Pr(\mathbf{X}_k|c_j)(1 - \Pr(\mathbf{X}_k|\bar{c}_j))}{\Pr(\mathbf{X}_k|\bar{c}_j)(1 - \Pr(\mathbf{X}_k|c_j))} \right|$$

$$CDM(\mathbf{X}_k) = \sum_{j=1}^{J} \left| \log \frac{\Pr(\mathbf{X}_k|c_j)}{\Pr(\mathbf{X}_k|\bar{c}_j)} \right|$$

where

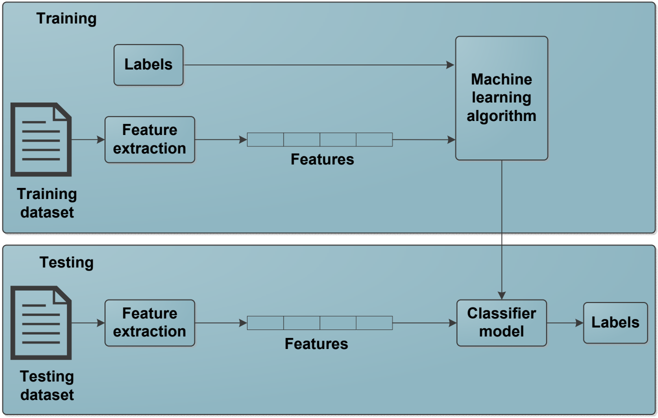
- $\Pr(\mathbf{X}_k|c_j)$ is the conditional probability of $\mathbf{X}_k$ given the class c_j
- $\Pr(c_j)$ is the probability of occurrence of the j^{th} class

The top ten selected features

Fisher		mRMR						Odds Ratio variants									
		MID		MIQ		MIBASE		OR		EOR		WOR		MOR		CMD	
$\mathcal{F}$	Score	$\mathcal{F}$	Score	$\mathcal{F}$	Score	$\mathcal{F}$	Score	$\mathcal{F}$	Score	$\mathcal{F}$	Score	$\mathcal{F}$	Score	$\mathcal{F}$	Score	$\mathcal{F}$	Score
11	0.397758	15	0.94	15	0.94	15	0.94	10	1.3602	5	2.1645	5	1.3963	6	2.3588	5	8.5959
6	0.354740	5	0.12	12	0.36	17	0.63	4	1.3085	7	2.1512	7	1.3762	5	2.3486	11	6.9743
9	0.271961	12	0.11	3	0.35	2	0.47	1	1.1088	6	2.1438	6	1.3648	11	2.3465	9	3.0844
2	0.185844	7	0.10	8	0.34	8	0.34	14	1.1080	11	2.1340	11	1.3495	17	2.3350	2	2.3485
16	0.123742	4	0.07	1	0.32	6	0.27	12	1.0973	10	2.0954	13	1.1963	16	2.3247	8	2.2402
17	0.121633	10	0.07	6	0.30	3	0.13	3	1.0797	4	2.0954	9	1.0921	14	2.1228	16	2.0985
8	0.116092	8	0.04	4	0.27	1	0.13	15	1.0465	13	2.0502	2	1.0198	1	2.1109	3	2.0606
3	0.086124	13	0.04	17	0.26	9	0.10	8	1.0342	9	2.0127	16	0.9850	2	2.1017	14	2.0506
1	0.081760	2	0.03	9	0.25	12	0.08	17	1.0304	1	2.0107	17	0.9778	7	2.0968	1	2.0417
14	0.081751	14	0.03	2	0.24	11	0.06	16	1.0202	14	2.0105	8	0.9751	3	2.0897	17	2.0213

Roadmap

Supervised classification process



Performance Evaluation

- We considered: accuracy, balanced accuracy, and F-score
- Definitions:
  - True positive (TP): is number of anomalous training data points that are classified as anomaly
  - True negative (TN): is number of regular training data points that are classified as regular
  - False positive (FP): is number of regular training data points that are classified as anomaly
  - False negative (FN): is number of anomalous training data points that are classified as regular

		Actual class	
		True (anomaly)	False (regular)
Anomaly test outcome	Positive	TP	FP
	Negative	FN	TN

Performance measures and indices

- Performance measures:

$$\text{sensitivity} = \frac{TP}{TP + FN}$$

$$\text{precision} = \frac{TP}{TP + FP}$$

- Performance indices:

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{balanced accuracy} = \frac{\text{sensitivity} + \text{precision}}{2}$$

$$\text{F-score} = 2 \times \frac{\text{precision} \times \text{sensitivity}}{\text{precision} + \text{sensitivity}}$$

Roadmap

Support Vector Machines

- Support vector machines were introduced by V. Vapnik in 1970s
- SVMs perform more accurately for datasets with high dimensional complexity
- For each training dataset $\mathbf{X}_{7200 \times 37}$, we target two classes: anomaly (true) and regular (false)
- Dimension of feature matrix: $7,200 \times 10$
- Each row contains the top ten selected features within the one-minute interval

References

- Support Vector Machine - The Book [Online]. Available: http://www.support-vector.net/chapter_6.html.
- Libsvm–a library for support vector machines [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

SVM two-way datasets

NB	Training dataset	Test dataset
SVM_1	Slammer and Nimda	Code Red I
SVM_2	Slammer and Code Red I	Nimda
SVM_3	Code Red I and Nimda	Slammer

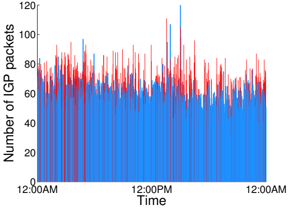
Two-way classification: performance

- All anomalies are treated as one class

SVM	Feature	Performance index: Accuracy (%) Test dataset (anomaly)	Accuracy (%) RIPE (regular)	Accuracy (%) BCNET (regular)	F-score (%) Test dataset (anomaly)
SVM_1	All features	64.1	55.0	62.0	63.2
SVM_1	Fisher	72.6	63.2	58.5	73.4
SVM_1	MID	63.1	52.2	59.4	61.2
SVM_1	MIQ	60.7	47.9	61.7	57.8
SVM_1	MIBASE	79.1	74.3	60.9	80.1
SVM_2	All features	68.6	97.7	79.2	22.2
SVM_2	Fisher	67.4	96.6	74.8	16.3
SVM_2	MID	67.9	97.4	72.5	19.3
SVM_2	MIQ	67.7	97.5	76.2	15.3
SVM_2	MIBASE	67.5	96.8	78.8	17.8
SVM_3	All features	81.5	92.0	69.2	84.6
SVM_3	Fisher	89.3	93.8	68.4	75.2
SVM_3	MID	75.4	92.8	71.7	79.2
SVM_3	MIQ	85.1	92.2	73.2	86.1
SVM_3	MIBASE	89.3	89.7	69.7	80.1

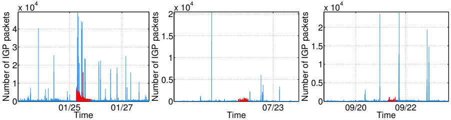
Classification results

- SVM_3 achieves the best F-score (86.1%) using features selected by MIQ
- BCNET and RIPE test datasets contain no anomalies and have low F-scores:
    - Performance measure: accuracy
    - SVM_2: the best overall two-way classifier
- Incorrectly classified (anomaly) BCNET traffic collected on December 20, 2011 (red):

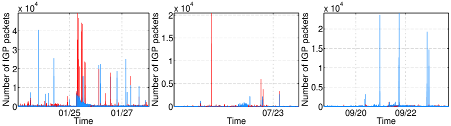


Classification results

- Correctly classified anomaly traffic (red):
  - Slammer (left)
  - Code Red I (middle)
  - Nimda (right)



- Incorrectly classified regular and anomaly traffic (red):
  - Slammer (left)
  - Code Red I (middle)
  - Nimda (right)



Four-way classification: performance

- Multi-class SVMs are used on training datasets: Slammer, Nimda, Code Red I, and RIPE regular/BCNET

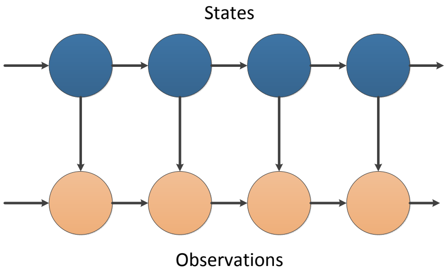
Feature	Average accuracy (%) (3 anomalies and 1 regular)	
	RIPE regular	BCNET
All features	77.1	91.4
Fisher	82.8	85.7
MID	67.8	78.7
MIQ	71.3	89.1
MIBASE	72.8	90.2

References
C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Networks*, vol. 13, no. 2, pp. 415–425, Mar. 2002.

Roadmap

Hidden Markov Models

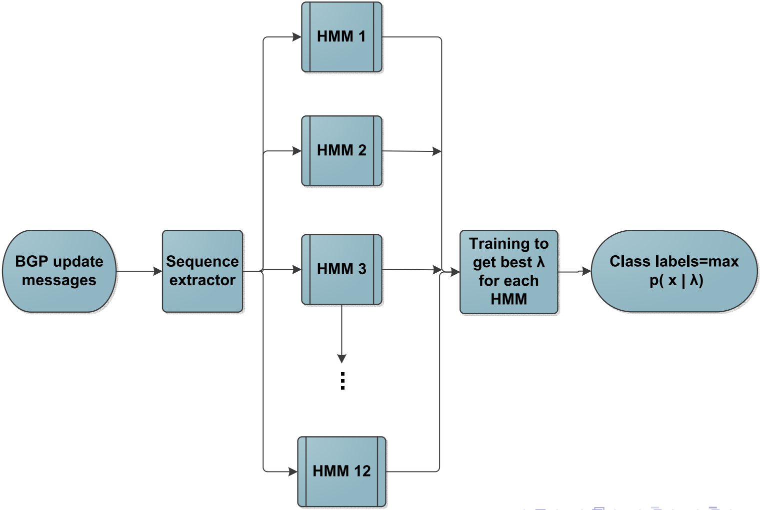
- First order HMMs are used to model stochastic processes that consist of two embedded processes:
  - observable process that maps BGP features
  - unobserved hidden process that has the Markov property
- Assumption: observations are independent and identically distributed



HMM classification stages

- HMM model is specified by a tuple $\lambda = (N, M, \alpha, \beta, \pi)$:
  N = number of hidden states (cross-validated)
  M = number of observations (11)
  α = transition probability distribution $N \times N$ matrix
  β = emission probability distribution $N \times M$ matrix
  π = initial state probability distribution matrix
- The proposed detection model consists of three stages:
  - *Observation vector extractor and mapping*: all features are mapped to 1-D observation vector
  - *Training*: two HMMs for two-way classification and four HMMs for four-way classification are trained to identify the best α and β for each class
  - *Classification*: maximum likelihood probability $p(x|\lambda)$ is used to classify the test observation vectors

HMM classification process



Classification

- HMMs with the same number of hidden states are compared
- Example: HMM_1, HMM_4, HMM_7, and HMM_{10} correspond to HMMs with two hidden states for various training datasets
- HMM accuracy:

$$\frac{Number\ of\ correctly\ classified\ observation\ vectors}{Total\ number\ of\ observation\ vectors}$$

Two-way classification: performance

		Performance index			
		Accuracy (%) anomaly concatenated with regular		F-score (%) anomaly concatenated with regular	
N	Feature set	RIPE regular	BCNET	RIPE regular	BCNET
2	(1,2)	86.0	94.0	84.4	93.8
2	(6,12)	79.0	71.0	76.2	60.7
4	(1,2)	78.0	87.0	72.2	85.0
4	(6,12)	64.0	60.0	48.0	35.9
6	(1,2)	85.0	91.0	84.3	90.1
6	(6,12)	81.0	65.0	80.1	50.2

- HMMs have better F-score using set (1, 2) than set (6, 12)

Four-way classification: performance

- Similar tests are applied using RIPE and BCNET datasets with four-way HMM classification.
- The classification accuracies are averaged over four HMMs for each dataset

		Average accuracy (%) 3 anomalies concatenated with 1 regular	
N	Feature set	RIPE regular	BCNET
2	(1,2)	72.50	77.50
2	(6,12)	38.75	41.25
4	(1,2)	66.25	76.25
4	(6,12)	26.25	33.75
6	(1,2)	70.00	76.25
6	(6,12)	43.75	42.50

Roadmap

Naive Bayes

- One of the most efficient machine learning classifiers
- Naivety: to assume that features are independent conditioned on a given class:

$$\Pr(\mathbf{X}_k = \boldsymbol{x}_k, \mathbf{X}_l = \boldsymbol{x}_l | c_j) = \Pr(\mathbf{X}_k = \boldsymbol{x}_k | c_j) \Pr(\mathbf{X}_l = \boldsymbol{x}_l | c_j)$$

  - $\boldsymbol{x}_k$ is realization of feature vector $\mathbf{X}_k$
  - $\boldsymbol{x}_l$ is realization of feature vector $\mathbf{X}_l$
- Advantages:
  - in some applications, it performs better than other classifiers
  - low complexity
  - may be trained effectively with smaller datasets

NB posterior

- Posterior of a data point represented as a row vector $\boldsymbol{x}_i$ is calculated using the Bayes rule:

$$\Pr(c_j|\mathbf{X}_i = \boldsymbol{x}_i) = \frac{\Pr(\mathbf{X}_i = \boldsymbol{x}_i|c_j)\Pr(c_j)}{\Pr(\mathbf{X}_i = \boldsymbol{x}_i)} \approx \Pr(\mathbf{X}_i = \boldsymbol{x}_i|c_j)\Pr(c_j)$$

- Naive Bayes:
  - Bayes rule: allows calculation of posterior distributions
  - Independence (naive): helps calculate the likelihood of a data point:

$$\Pr(\mathbf{X}_i = \boldsymbol{x}_i|c_j) = \prod_{k=1}^{K} \Pr(X_{ik} = x_{ik}|c_j)$$

Likelihoods and priors

- Priors correspond to the relative frequencies of the training data for each class c_j:

$$\Pr(c_j) = \frac{N_j}{N}$$

  - N_j is the number of training data that belong to the j^{th} class
  - N is the total number of training data points

- Gaussian distribution is used to generate the likelihood distributions (continuous features):

$$\Pr(X_{ik} = x_{ik} | c_j, \mu_k, \sigma_k) = \mathcal{N}(X_{ik} = x_{ik} | c_j, \mu_k, \sigma_k)$$

- Parameters $\{\mu_{c_j}, \sigma_{c_j}\}$ are validated for each class

NB classification

- Classification:
  - two-way classification:
    $\max\{\Pr(c_1|\mathbf{X}_i = \boldsymbol{x}_i), \Pr(c_2|\mathbf{X}_i = \boldsymbol{x}_i)\}$
  - four-way classification:
    $\max\{\Pr(c_1|\mathbf{X}_i = \boldsymbol{x}_i)), \Pr(c_2|\mathbf{X}_i = \boldsymbol{x}_i), \Pr(c_3|\mathbf{X}_i = \boldsymbol{x}_i), \Pr(c_4|\mathbf{X}_i = \boldsymbol{x}_i)\}$
- Example (two-way classification):
  an arbitrary training data point $\boldsymbol{x}_i$ is classified as anomalous if $\Pr(c_1|\mathbf{X}_i = \boldsymbol{x}_i) > \Pr(c_2|\mathbf{X}_i = \boldsymbol{x}_i)$

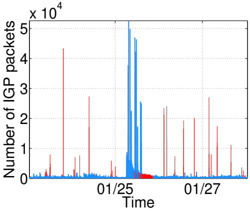
Two-way classification: performance

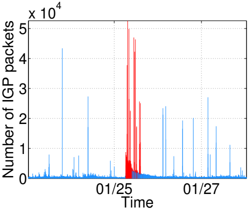
No.	NB	Feature	Performance index: Accuracy (%) Test dataset (anomaly)	Accuracy (%) RIPE (regular)	Accuracy (%) BCNET (regular)	F-score (%) Test dataset (anomaly)
1	NB1	All features	69.1	91.1	77.3	38.8
2	NB1	Fisher	72.1	92.3	76.3	46.1
3	NB1	MID	66.0	94.7	78.2	25.4
4	NB1	MIQ	70.8	89.9	80.9	44.7
5	NB1	MIBASE	71.2	88.2	81.3	46.9
6	NB1	OR	66.5	77.9	94.7	26.2
7	NB1	EOR	70.4	78.3	92.7	42.0
8	NB1	WOR	74.1	77.2	89.3	52.8
9	NB1	MOR	72.1	80.8	90.9	46.8
10	NB1	CDM	71.8	80.8	92.6	45.3
11	NB2	All features	68.1	92.1	87.1	21.4
12	NB2	Fisher	68.2	93.4	89.0	22.6
13	NB2	MID	65.2	**95.8**	90.7	6.4
14	NB2	MIQ	68.0	91.5	88.9	22.3
15	NB2	MIBASE	68.5	90.7	89.3	24.8
16	NB2	OR	65.2	87.9	96.0	6.2
17	NB2	EOR	69.0	90.4	93.6	26.5
18	NB2	WOR	70.1	90.9	91.6	32.1
19	NB2	MOR	68.2	91.2	93.8	22.0
20	NB2	CDM	70.1	91.5	90.9	32.1
21	NB3	All features	83.4	91.3	85.9	57.8
22	NB3	Fisher	88.1	90.7	85.9	68.5
23	NB3	MID	80.5	95.8	90.9	43.6
24	NB3	MIQ	84.4	91.2	89.1	58.1
25	NB3	MIBASE	85.1	89.8	89.1	61.4
26	NB3	OR	82.3	88.6	**95.5**	46.7
27	NB3	EOR	84.8	85.1	92.4	58.9
28	NB3	WOR	87.4	84.3	90.1	69.7
29	NB3	MOR	87.3	84.4	89.1	69.2
30	NB3	CDM	87.9	84.4	91.4	67.0

Four-way classification: performance

		Average accuracy (%) 3 anomalies concatenated with 1 regular	
No.	Feature set	RIPE regular	BCNET
1	All features	74.3	67.6
2	Fisher	24.7	34.3
3	MID	74.9	33.1
4	MIQ	24.6	34.8
5	MIBASE	75.4	33.1
6	OR	25.5	36.7
7	EOR	75.3	68.1
8	WOR	75.8	53.2
9	MOR	77.7	68.7
10	CDM	24.8	34.5

Classification results: Slammer worm (January 25, 2003)

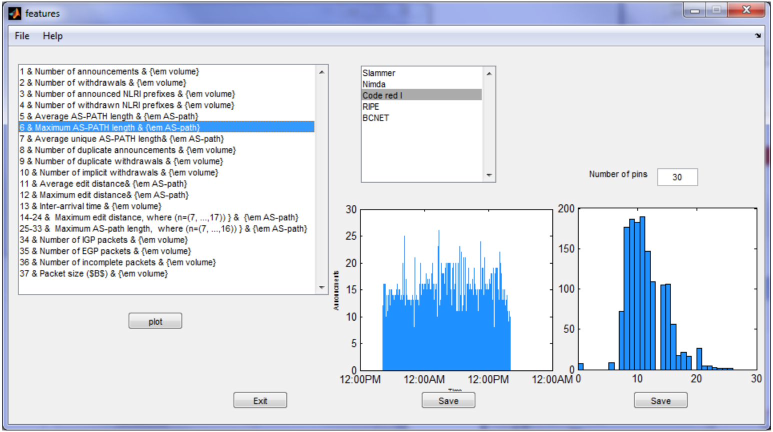




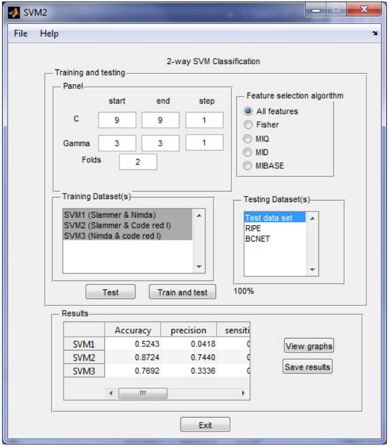
- Left: incorrectly classified (red) regular (false positives) and anomaly (false negatives) data points
- Right: correctly classified (red) anomaly (true positives) data points
- Correctly classified regular (true negatives) data points are not shown
- All anomalous data points that have large number of IGP packets (volume feature) are correctly classified

Roadmap

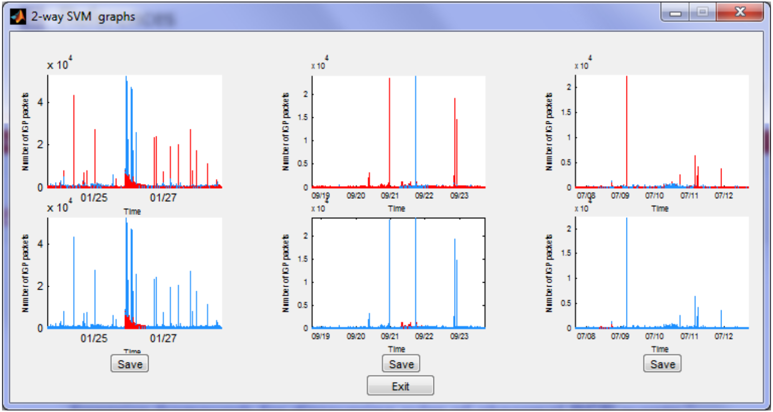
BGPAD tool:
Inspects BGP pcap and MRT files for anomalies



BGPAD tool:
Provides test performance indices



BGPAD tool: Displays anomalous traffic



Roadmap

Discussion: feature extraction and selection

- The trust relationship among BGP peers is vulnerable during anomaly attacks
- Example: during BGP hijacks, a BGP peer may announce unauthorized prefixes that indicate to other peers that it is the originating peer
- Effect of anomalies on volume features:
  - False announcements propagate across the Internet and affect the number of BGP announcements (updates and withdrawals)

Discussion: feature extraction and selection

- Effect of anomalies on AS-path features:
  - large length of the AS-PATH BGP attribute implies that the packet is routed via a longer path to its destination
  - very short lengths of AS-PATH attributes occur during BGP hijacks when the new (false) originator usually gains a preferred or shorter path to the destination
  - edit distance and AS-PATH length of the BGP announcements tend to have a very high or a very low value (large variance)
- The top selected AS-path features appear on the boundaries of the distributions: AS-path features 25, 32, and 24 have the highest Fisher, MID, and MIQ scores

Discussion: classification

- SVM models exhibited better performance than the HMMs and NB in two-way and four-way classifications
- SVM and NB models based on Code Red I and Nimda datasets
- HMMs have the highest accuracies
  - with two hidden states
  - using the number of announcements and number of withdrawals (feature 1 and feature 2) than the than models with the maximum number of AS-PATH length (feature 6) and the maximum edit distance (feature 12)
- SVM, HMM, and NB two-way classifications produced better results than four-way classifications because of the common semantics among BGP anomalies

Discussion: classification

- RIPE regular and BCNET test datasets contain no anomalies and have low F-scores. For example, In two-way NB:
  - Performance measure (accuracy):
    - RIPE regular: 95.8%
    - BCNET: 95.5%
- OR algorithms often achieve better performance:
  - feature score is calculated using the probability distribution that the NB classifiers use for posterior calculations
  - features selected by the OR variants are expected to have stronger influence on the posteriors

Discussion: classification

- WOR feature selection algorithm achieves the best F-score for all NB classifiers
- Performance of the NB classifiers is often inferior to the SVM and HMM classifiers
- NB2 classifier trained on Slammer and Code Red I datasets performs better than the SVM classifier

References

- N. Al-Rousan, S. Haeri, and Lj. Trajkovic, "Feature selection for classification of BGP anomalies using Bayesian models," *in Proc. ICMLC 2012*, Xi'an, China, July 2012.
- N. Al-Rousan and Lj. Trajkovic, "Machine learning models for classification of BGP anomalies," *Proc. IEEE Conf. High Performance Switching and Routing, HPSR 2012*, Belgrade, Serbia, June 2012, pp. 103-108.
- D. Mladenic and M. Grobelnik, "Feature selection for unbalanced class distribution and naive bayes," in *Proc. Int. Conf. Machine Learning*, Bled, Slovenia, June 1999, pp. 258–267.

Discussion: comparison of features category performance

- The volume features accounted for 65% of selected features
- We applied two-way SVM classification with only *volume* and again with *AS-path* features
- Performance of SVM using *volume* features was superior to *AS-path*

		Performance index					
SVM	category	accuracy	precision	sensitivity	specificity	balanced accuracy	f-score
SVM1	*volume*	68.5	53.6	16.6	73.2	44.9	27.1
SVM1	*AS-path*	56.4	6.12	29.5	58.8	44.1	3.93
SVM2	*volume*	87.0	69.6	12.5	99.1	55.8	22.3
SVM2	*AS-path*	86.0	38.7	1.19	99.6	50.4	2.36
SVM3	*volume*	94.8	79.7	76.4	97.3	86.8	85.0
SVM3	*AS-path*	56.9	19.1	79.4	53.8	66.6	64.1

Discussion: performance comparison

- Performance comparison:
  - Rule based techniques: better results in two out of three datasets
  - Behavioural techniques: worse results in all the three datasets

Dataset	Proposed models SVM (two-way)	SVM (four-way)	HMM (two-way)	HMM (four-way)	NB (two-way)	NB (four-way)	Rule based techniques	Behavioural techniques
Slammer	89.3	82.8	86.0	70.0	87.4	77.7	94.4	74.0
Nimda	68.6	82.8	86.0	70.0	70.1	77.7	84.1	74.0
Code Red I	79.1	82.8	86.0	70.0	74.1	77.7	74.9	74.0

References

- R. Moskovitch, Y. Elovici, and L. Rokach, "Detection of unknown computer worms based on behavioral classification of the host abstract." 2008.
- D. Dou, J. Li, H. Qin, and S. Kim, "S.: Understanding and utilizing the hierarchy of abnormal bgp events," in *In: SIAM International Conference on Data Mining*, 2007, pp. 457–462.

Conclusions

- Anomalies in BGP traffic traces were successfully classified using SVM, HMM, and NB models
- Various feature selection algorithms and machine learning models were employed to design BGP anomaly detectors
- Volume features are more relevant to the anomaly class than the AS-path features
- The OR algorithms often achieved higher F-scores in the two-way and four-way classifications with various training datasets

Conclusions

- The best achieved F-scores: SVM (86.1%), HMM (84.4%), and NB (69.7%)
- Using the BGP volume features is a viable approach for detecting possible worm attacks
- The proposed models may be used as online mechanisms to predict new BGP anomalies and detect the onset of worm attacks

Acknowledgements

- Chair:
  - Professor Glenn H. Chapman
- Senior Supervisor:
  - Professor Ljiljana Trajković
- Supervisor:
  - Associate Professor Jie Liang
- Examiner:
  - Professor Emeritus William A. Gruver

Roadmap

References: http://www.sfu.ca/~ljilja/cnl

N. Al-Rousan, S. Haeri, and Lj. Trajkovic, "Feature selection for classification of BGP anomalies using Bayesian models," *in Proc. ICMLC 2012*, Xi'an, China, July 2012.

N. Al-Rousan and Lj. Trajkovic, "Machine learning models for classification of BGP anomalies," *Proc. IEEE Conf. High Performance Switching and Routing, HPSR 2012*, Belgrade, Serbia, June 2012, pp. 103-108.

T. Farah, S. Lally, R. Gill, N. Al-Rousan, R. Paul, D. Xu, and Lj. Trajkovic, "Collection of BCNET BGP traffic," in *Proc. 23rd ITC*, San Francisco, CA, USA, Sept. 2011, pp. 322–323.

S. Lally, T. Farah, R. Gill, R. Paul, N. Al-Rousan, and Lj. Trajkovic, "Collection and characterization of BCNET BGP traffic," in *Proc. 2011 IEEE Pacific Rim Conf. Communications, Computers and Signal Processing*, Victoria, BC, Canada, Aug. 2011, pp. 830–835.

References: literature review

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, pp. 1470–1484, Nov. 2009.

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, pp. 55–66, Oct. 2005.

L. Wang, X. Zhao, D. Pei, R. Bush, D. Massey, A. Mankin, S. F. Wu, and L. Zhang, "Observation and analysis of BGP behavior under stress," in *Proc. 2nd ACM SIGCOMM Workshop on Internet Measurment*, New York, NY, USA, 2002, pp. 183–195.

H. Hajji, "Statistical analysis of network traffic for adaptive faults detection," *IEEE Trans. Neural Netw.*, vol. 16, no. 5, pp. 1053–1063, Sept. 2005.

M. Thottan and C. Ji, "Anomaly detection in IP networks," *IEEE Trans. Signal Process.*, vol. 51, no. 8, pp. 2191–2204, Aug. 2003.

A. Dainotti, A. Pescape, and K. Claffy, "Issues and future directions in traffic classification," *IEEE Network*, vol. 26, no. 1, pp. 35–40, Feb. 2012.

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. 5, pp. 55–66, Oct. 2005.

References: BGP

- T. Manderson, "Multi-threaded routing toolkit (MRT) border gateway protocol (BGP) routing information export format with geo-location extensions," RFC 6397, *IETF*, Oct. 2011 [Online]. Available: `http://www.ietf.org/rfc/rfc6397.txt`.
- D. Meyer, "BGP communities for data collection," RFC 4384, *IETF*, 2006 [Online]. Available: `http://www.ietf.org/rfc/rfc4384.txt`.
- Y. Rekhter and T. Li, "A Border Gateway Protocol 4 (BGP-4)," RFC 1771, *IETF*,Mar. 1995.
- RIPE RIS raw data [Online]. Available: `http://www.ripe.net/data-tools/stats/ris/ris-raw-data`.
- University of Oregon Route Views project [Online]. Available: `http://www.routeviews.org/`.
- Zebra BGP parser [Online]. Available: `http://www.linux.it/~md/software/zebra-dump-parser.tgz`.
- BCNET [Online]. Available: `http://www.bc.net`.
- YouTube Hijacking: A RIPE NCC RIS case study [Online]. Available: `http://www.ripe.net/internet-coordination/news/industry-developments/youtube-hijacking-a-ripe-ncc-ris-case-study`.

References: machine learning

C. M. Bishop, *Pattern Recognition and Machine Learning*. Secaucus, NJ, USA: Springer-Verlag, 2006.

Support Vector Machine - The Book [Online]. Available: http://www.support-vector.net/chapter_6.html.

Libsvm–a library for support vector machines [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

T. Ahmed, B. Oreshkin, and M. Coates, "Machine learning approaches to network anomaly detection," in *Proc. USENIX Workshop Tackling Computer Systems Problems with Machine Learning Techniques*, Cambridge, MA, 2007, pp. 1–6.

Y.-W. Chen and C.-J. Lin, "Combining SVMs with various feature selection strategies," *Strategies*, vol. 324, no. 1, pp. 1–10, Nov. 2006.

A. Munoz and J. Moguerza, "Estimation of high-density regions using one-class neighbor machines," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 3, pp. 476–480, Mar. 2006.

T. Ahmed, M. Coates, and A. Lakhina, "Multivariate online anomaly detection using kernel recursive least squares," in *Proc. 26th IEEE Int. Conf. Comput. Commun.*, Anchorage, AK, USA, May 2007, pp. 625–633.

J. Zhang, J. Rexford, and J. Feigenbaum, "Learning-based anomaly detection in BGP updates," in *Proc. Workshop Mining Network Data*, Philadelphia, PA, USA, Aug. 2005, pp. 219–220.

C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Networks*, vol. 13, no. 2, pp. 415–425, Mar. 2002.

References: feature selection

- H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.
- G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant features and the subset selection problem," in *Proc. Int. Conf. Machine Learning*, New Brunswick, NJ, USA, July 1994, pp. 121–129.
- Q. Gu, Z. Li, and J. Han, "Generalized Fisher score for feature selection," in *Proc. Conf. Uncertainty in Artificial Intelligence*, Barcelona, Spain, July 2011, pp. 266–273.
- J. Wang, X. Chen, and W. Gao, "Online selecting discriminative tracking features using particle filter," in *Proc. Computer Vision and Pattern Recognition*, San Diego, CA, USA, June 2005, vol. 2, pp. 1037–1042.
- H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.
- J. Chen, H. Huang, S. Tian, and Y. Qu, "Feature selection for text classification with naive Bayes," *Expert Systems with Applications*, vol. 36, no. 3, pp. 5432–5435, Apr. 2009.

References: naive Bayes

A. W. Moore and D. Zuev, "Internet traffic classification using Bayesian analysis techniques," in *Proc. Int. Conf. Measurement and Modeling of Comput. Syst.*, Banff, Alberta, Canada, June 2005, pp. 50–60.

K. El-Arini and K. Killourhy, "Bayesian detection of router configuration anomalies," in *Proc. Workshop Mining Network Data*, Philadelphia, PA, USA, Aug. 2005, pp. 221–222.

D. Mladenic and M. Grobelnik, "Feature selection for unbalanced class distribution and naive bayes," in *Proc. Int. Conf. Machine Learning*, Bled, Slovenia, June 1999, pp. 258–267.

Thank You