Feature Selection for Classification of BGP Anomalies using Bayesian Models

Nabil Al-Rousan, Soroush Haeri, and Ljiljana Trajković
Simon Fraser University
Vancouver, British Columbia, Canada
Email: {nalrousa, shaeri, ljilja}@sfu.ca
http://www.sfu.ca/~ljilja/cnl

July 17, 2012

Roadmap

Motivation

- 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 protocol generates four types of messages:
  - open, update, keepalive, and notification
- Only BGP update messages were considered
- BGP update messages: announcements or withdrawals

Contributions

- We develop various Naive Bayes (NB) classifiers for detecting BGP anomalies
- The classifiers are trained on the feature sets selected by various feature selection algorithms
- Employed feature selection algorithms include:
  - Fisher
  - minimum redundancy maximum relevance (mRMR)
  - odds ratio (OR), extended/weighted/multi-class odds ratio (EOR/WOR/MOR), and class discriminating measure (CDM)
- The proposed models are tested with collected BGP traffic traces and are employed to successfully classify and detect various BGP anomalies

Detection techniques

- Techniques to detect BGP anomalies have recently gained visible attention and importance
- Statistical pattern recognition:
  - main disadvantage: difficulty in estimating distributions of high dimensions
- Rule-based:
  - require a priori knowledge of network conditions
  - example: the Internet Routing Forensics (IRF)
  - they are not adaptable learning mechanisms, slow, and have high degree of computational complexity

References

- 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.

Datasets sources

- The RIPE and Route Views BGP update messages: multi-threaded routing toolkit (MRT) binary format
- Proposed models were also validated by 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	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`.
- T. Manderson, "Multi-threaded routing toolkit (MRT) border gateway protocol (BGP) routing information export format with geo-location extensions,"
- Zebra BGP parser [Online]. Available: `http://www.linux.it/~md/software/zebra-dump-parser.tgz`.

Sampling and normalization of BGP features

- Features are sampled every minute over five days, producing 7,200 samples for each anomaly event
- Features are normalized to have zero mean and unit variance
- This normalization reduces the effect of the Internet growth between 2003 and 2011

References

- D. Meyer, "BGP communities for data collection," RFC 4384, *IETF*, 2006 [Online]. Available: `http://www.ietf.org/rfc/rfc4384.txt`.

List of extracted features

Feature ($\mathcal{F}$)	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	Number of Interior Gateway Protocol packets	*volume*
15	Number of Exterior Gateway Protocol packets	*volume*
16	Number of incomplete packets	*volume*
17	Packet size	*volume*

Feature selection algorithms

- Features scoring algorithms:
  - Fisher
  - minimum Redundancy Maximum Relevance (mRMR)
  - odds ratio/extended/weighted/multi-class odds ratio (OR/EOR/WOR/MOR)
  - class discriminating measure (CDM)

References

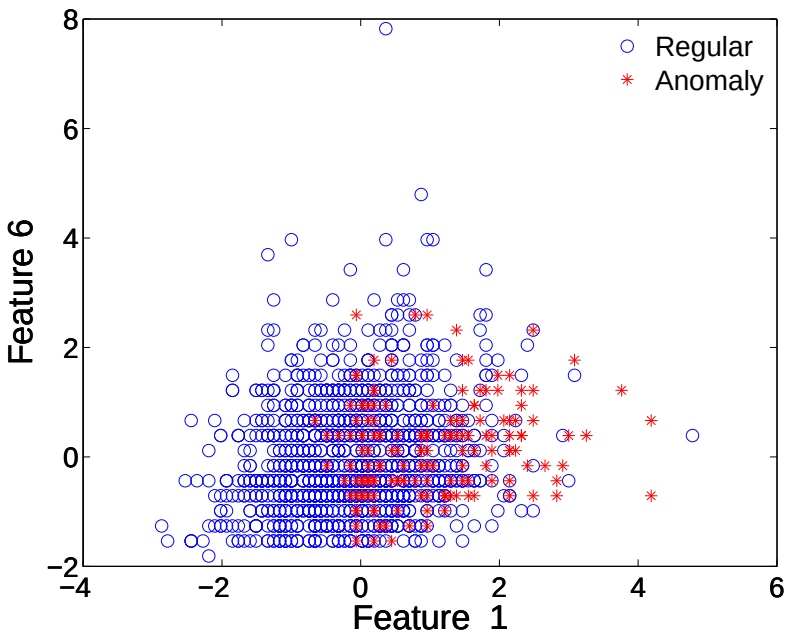
- H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.
- J. Wang, X. Chen, and W. Gao, "Online selecting discriminative tracking features using particle filter," in *Proc. Computer Vision and Pattern Recognition*, San Diego, CA, USA, June 2005, vol. 2, pp. 1037–1042.
- 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.

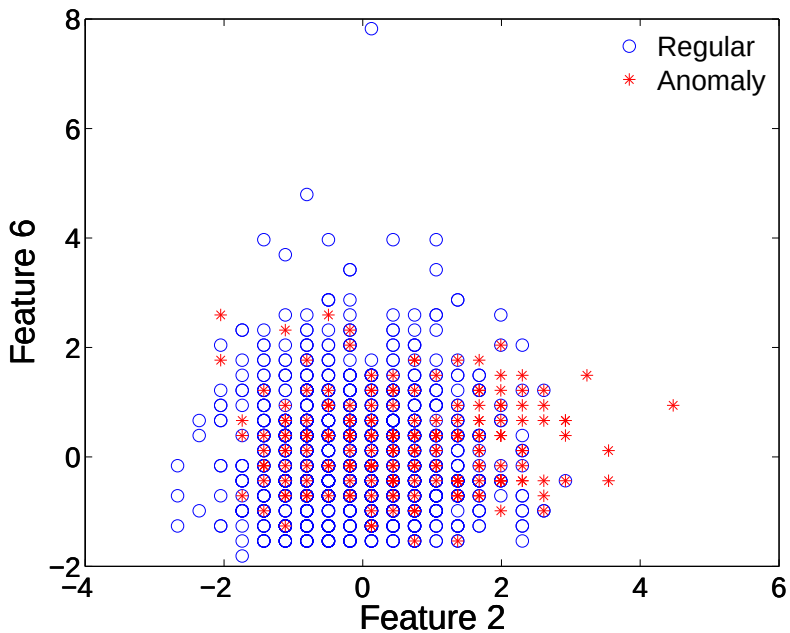
The top ten selected features

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

Normalized scattering graphs

- Feature 1, Feature 2, and Feature 6:





- Selecting appropriate combination of features is essential for accurate classification

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:

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

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

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

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 μ_k and σ_k are the mean and standard deviation of the k^{th} feature

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 $\Pr(c_1|\mathbf{X}_i = \boldsymbol{x}_i) > \Pr(c_2|\mathbf{X}_i = \boldsymbol{x}_i)$

NB performance evaluation

- We used MATLAB statistical toolbox
- The feature matrix for each dataset consist of:
  - 7,200 rows corresponding to the number of training data points
  - 17 columns representing the value of each feature for each data point
- Training datasets:
  - Two-way: each training data point as: anomalous and regular
  - Four-way: each training data point as: Slammer, Nimda, Code Red I, or Regular
- Training datasets for two-way classifiers:

NB	Training dataset	Test dataset
NB1	Slammer and Nimda	Code Red I
NB2	Slammer and Code Red I	Nimda
NB3	Code Red I and Nimda	Slammer

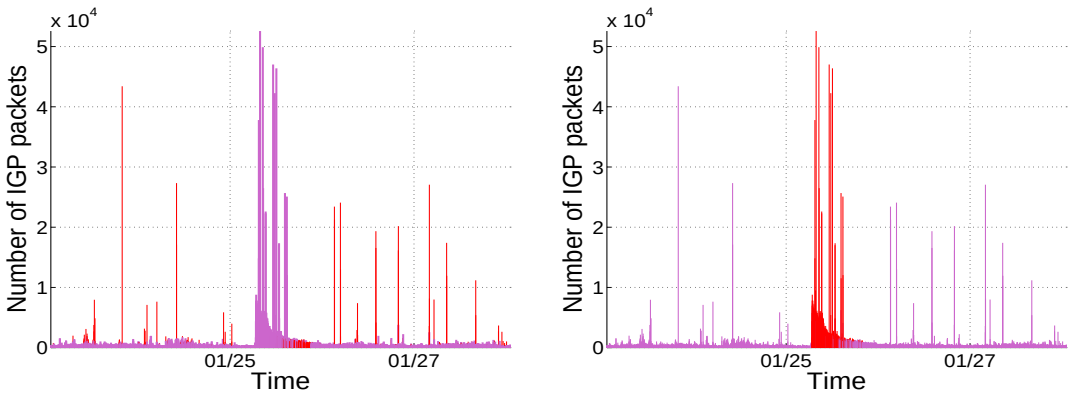
Two-way classification: performance

No.	NB	Feature	Performance index			
			Accuracy (%)			F-score (%)
			Test dataset (anomaly)	RIPE (regular)	BCNET (regular)	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

No.	Feature set	Average accuracy (%) 3 anomalies and 1 regular	
		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

Discussion

- Combination of Code Red I and Nimda training datasets (NB3) achieves the best classification results
- RIPE regular and BCNET test datasets contain no anomalies and have low F-scores:
  - Performance measure (accuracy):
    - RIPE regular: 95.8%
    - BCNET: 95.5%
  - F-score is not defined and the accuracy reduces to

$$\frac{TN}{TN + FP}$$

- 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

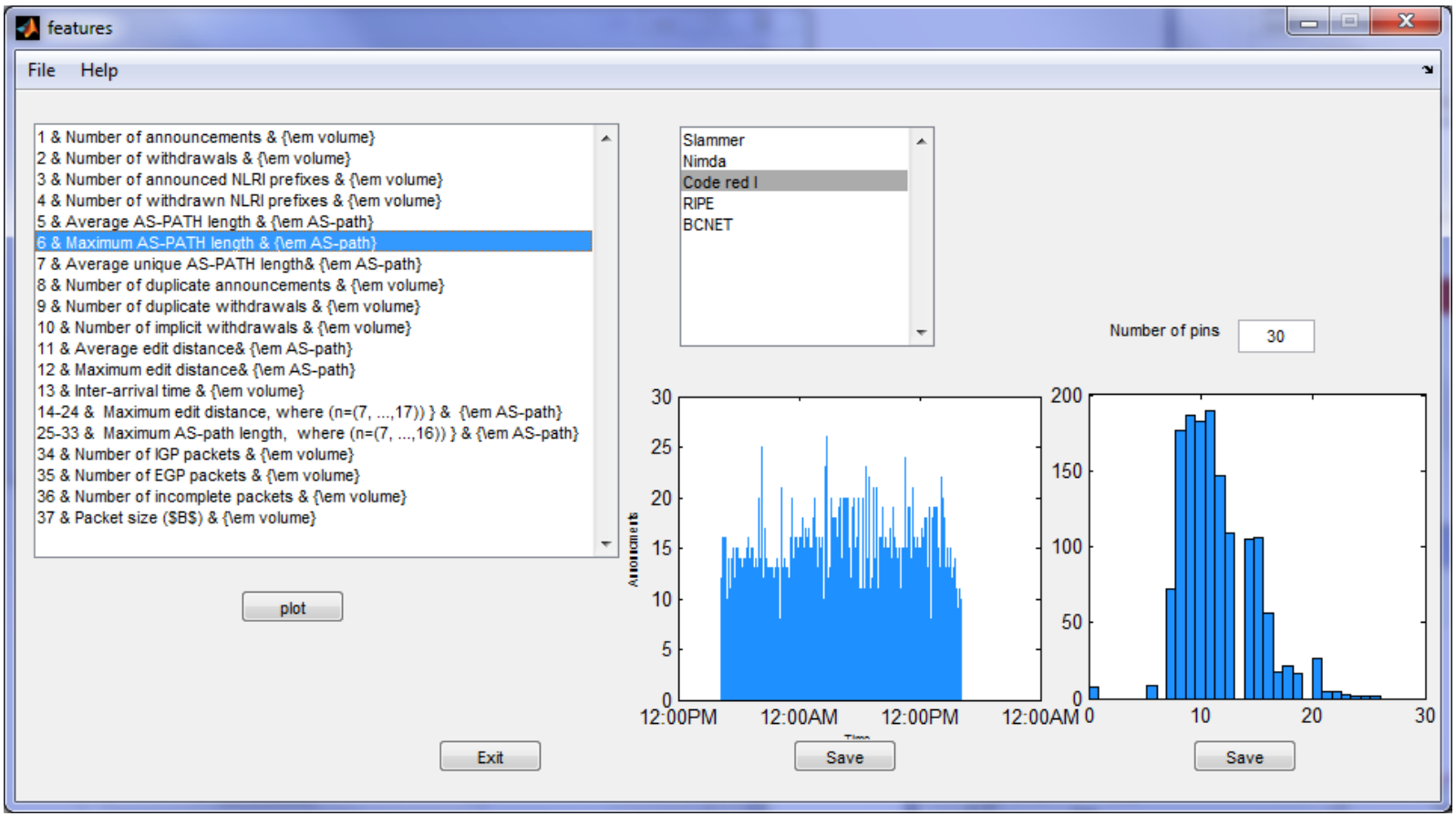
Discussion

- 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

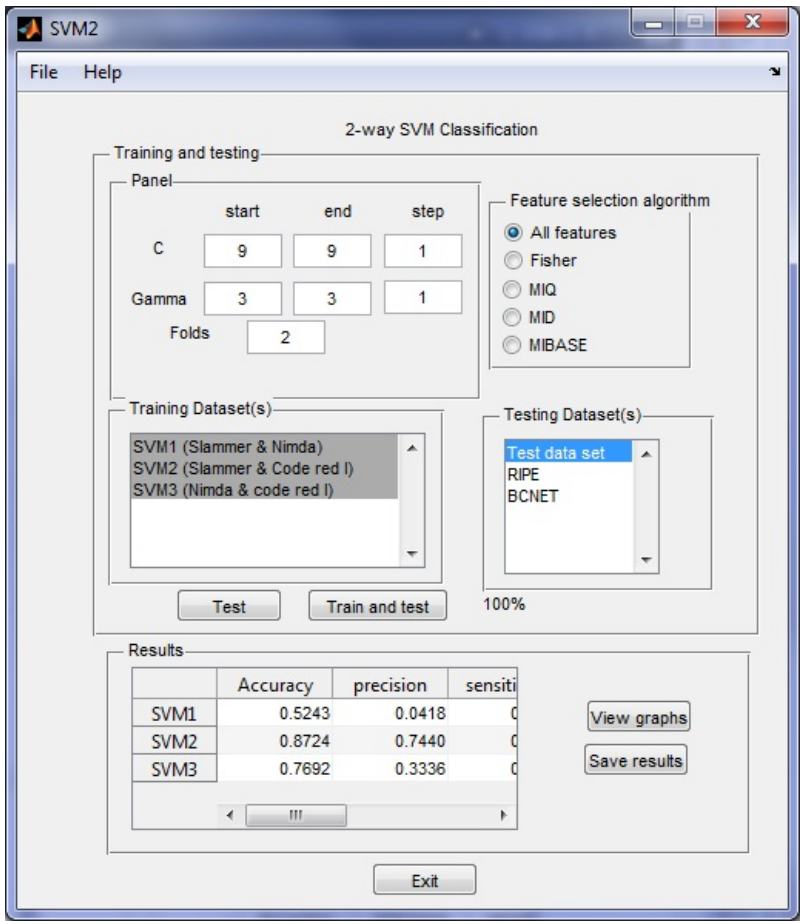
References

- 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.
- 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.

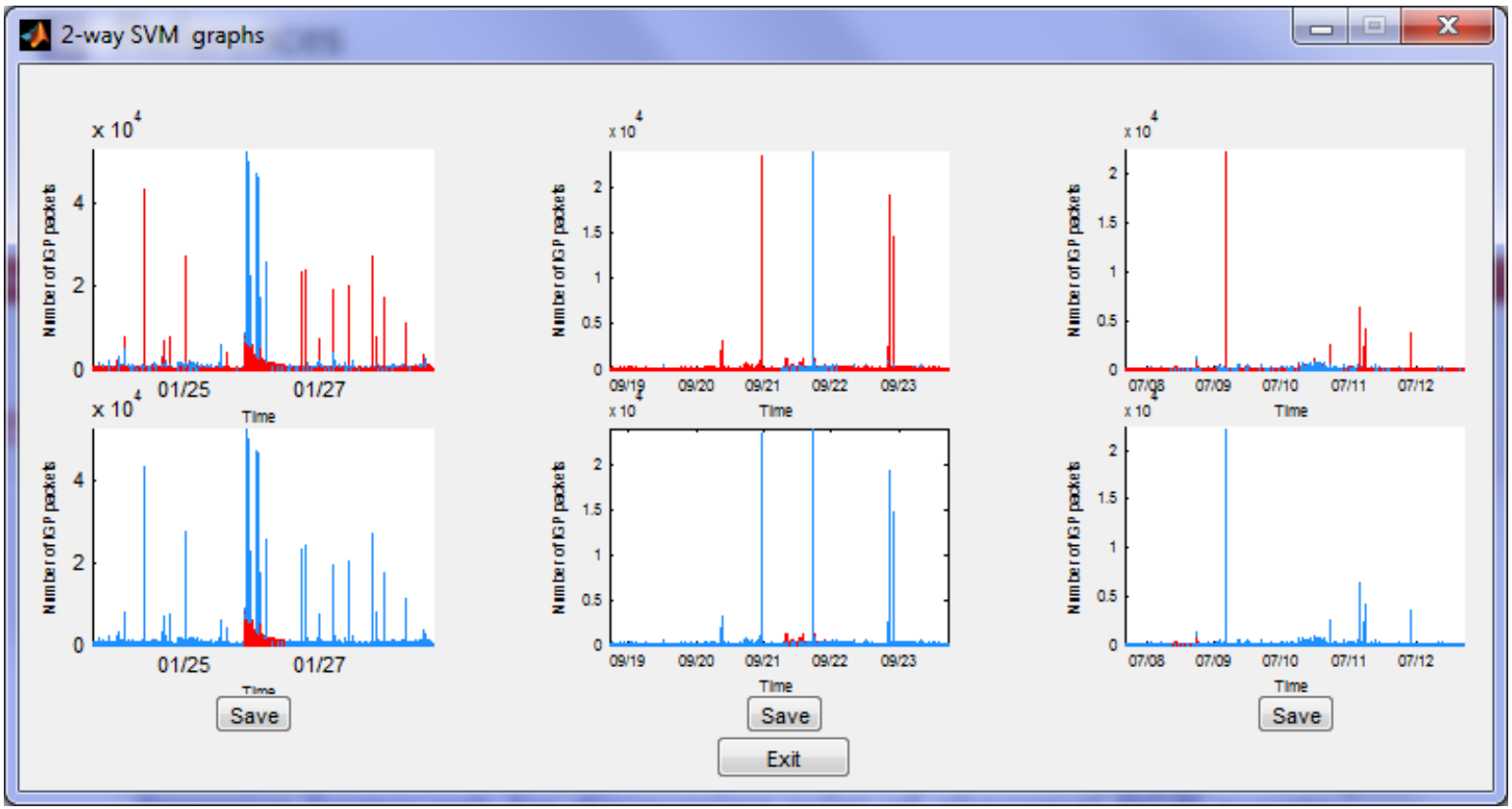
BGPAD tool: inspects BGP pcap and MRT files for anomalies



BGPAD tool: provides test performance indices



BGPAD tool: displays anomalous traffic



Conclusions

- Anomalies in BGP traffic traces were successfully classified using NB models
- Various feature selection algorithms and generative NB models were employed to design BGP anomaly detectors
- The usage of OR algorithms was extended from categorical to continuous features
- OR algorithms often achieved higher F-scores in the two-way and four-way classifications with various training datasets

Conclusions

- The NB classifiers may be used for online detection of anomalies:
  - they have low complexity
  - may be trained effectively on smaller datasets

References

- 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.

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


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.