Evaluation of Support Vector Machine Kernels for Detecting Network Anomalies

Prerna Batta
pbatta@sfu.ca

Communication Networks Laboratory
http://www.ensc.sfu.ca/~ljilja/cnl/
School of Engineering Science
Simon Fraser University

Roadmap

- Introduction:
  - Border Gateway Protocol (BGP)
  - Machine learning
- Feature extraction and selection
- Support vector machine and kernels
- Research contributions
- Experimental procedure and classification results
- Conclusions and future work
- References

Roadmap

- Introduction
  - Border Gateway Protocol (BGP)
  - Machine learning
- Feature extraction and selection
- Support vector machine and kernels
- Research contributions
- Experimental procedure and classification results
- Conclusions and future work
- References

Introduction: Border Gateway Protocol

- BGP's main function is to optimally route data between Autonomous Systems (ASes)
- AS: a collection of BGP routers (peers) within a single administrative domain
- Four types of BGP messages:
  - open, keepalive, update, and notification
- BGP anomalies:
  - Slammer, Nimda, Code Red I, routing misconfigurations

Introduction: Machine learning

- Machine learning models classify data using a feature matrix:
  - rows: data points
  - columns: feature values
- Algorithms:
  - Logistic Regression, Naïve Bayes, Support Vector Machine (SVM)
- SVM defines decision boundary to geometrically lie midway between the support vectors

Machine learning techniques

- Supervised learning:
  - input data is labelled
  - goal is to find specific connection among the input variable to predict the correct output
- Unsupervised learning:
  - input data is unlabeled
  - goal is to label the input data before determining the hidden patterns and structures

Roadmap

- Introduction
  - Border Gateway Protocol (BGP)
  - Machine learning
- **Feature extraction** and selection
- Support vector machine and kernels
- Research contribution
- Experimental procedure and classification results
- Conclusions and future work
- References

Feature extraction: BGP messages

- Extract 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

Feature extraction: BGP messages

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

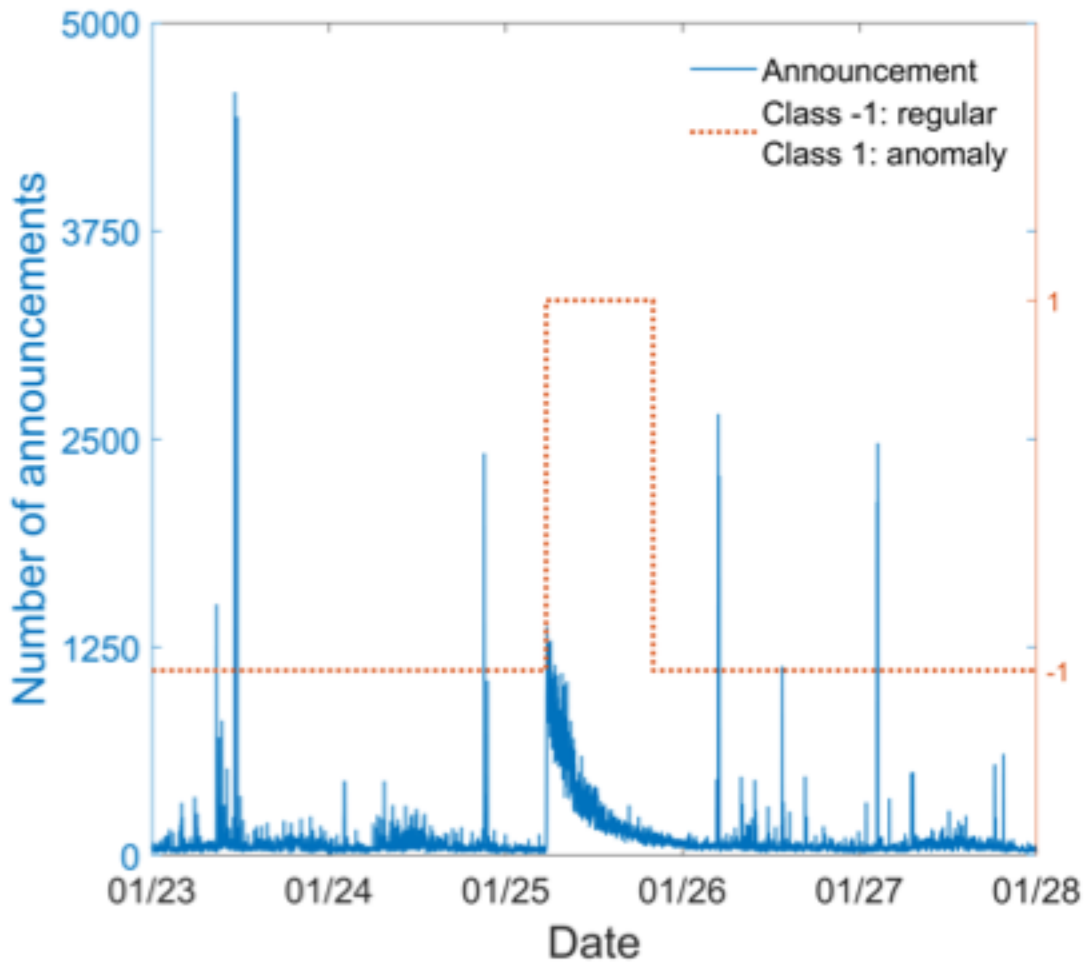
BGP anomalies

- Slammer:
  - infected Microsoft SQL servers through a small piece of code that generated IP addresses at random
- Nimda:
  - exploited vulnerabilities in the Microsoft Internet Information Services (IIS) web servers for Internet Explorer 5
- Code Red I:
  - attacked Microsoft IIS web servers by replicating itself through IIS server weaknesses

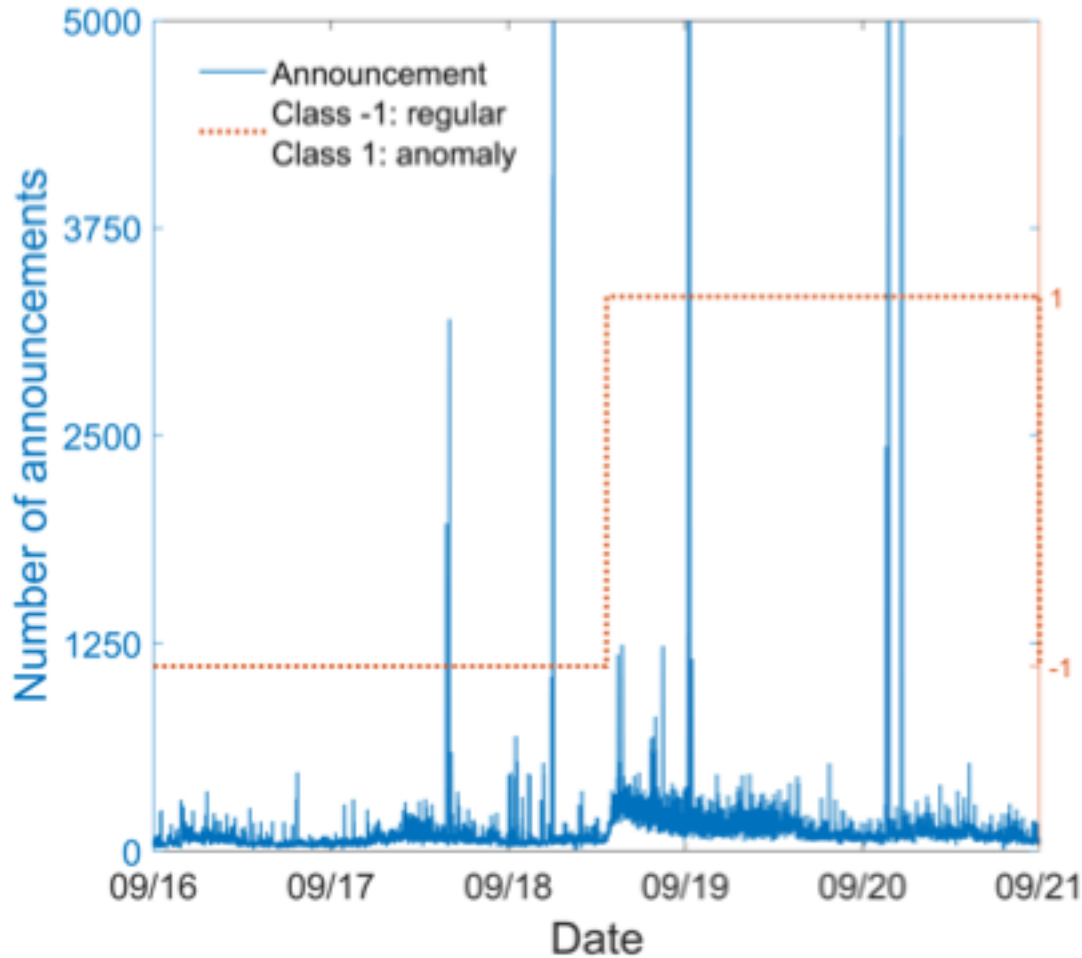
Duration of BGP events

Anomaly	Date	Anomaly (min)	Regular (min)
Slammer	January 25, 2003	869	6,331
Nimda	September 18-20, 2001	3,521	3,679
Code Red I	July 19, 2001	600	6,600

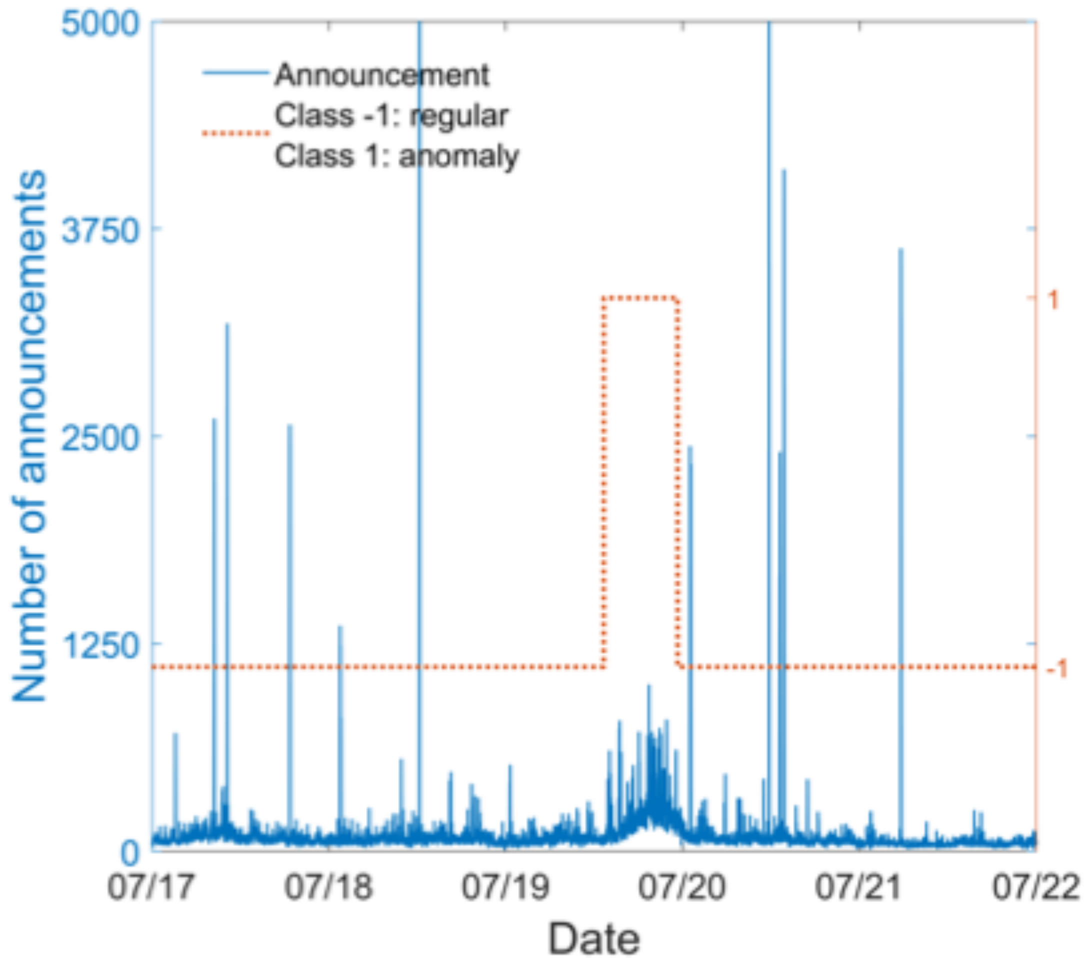
Number of BGP announcements: Slammer



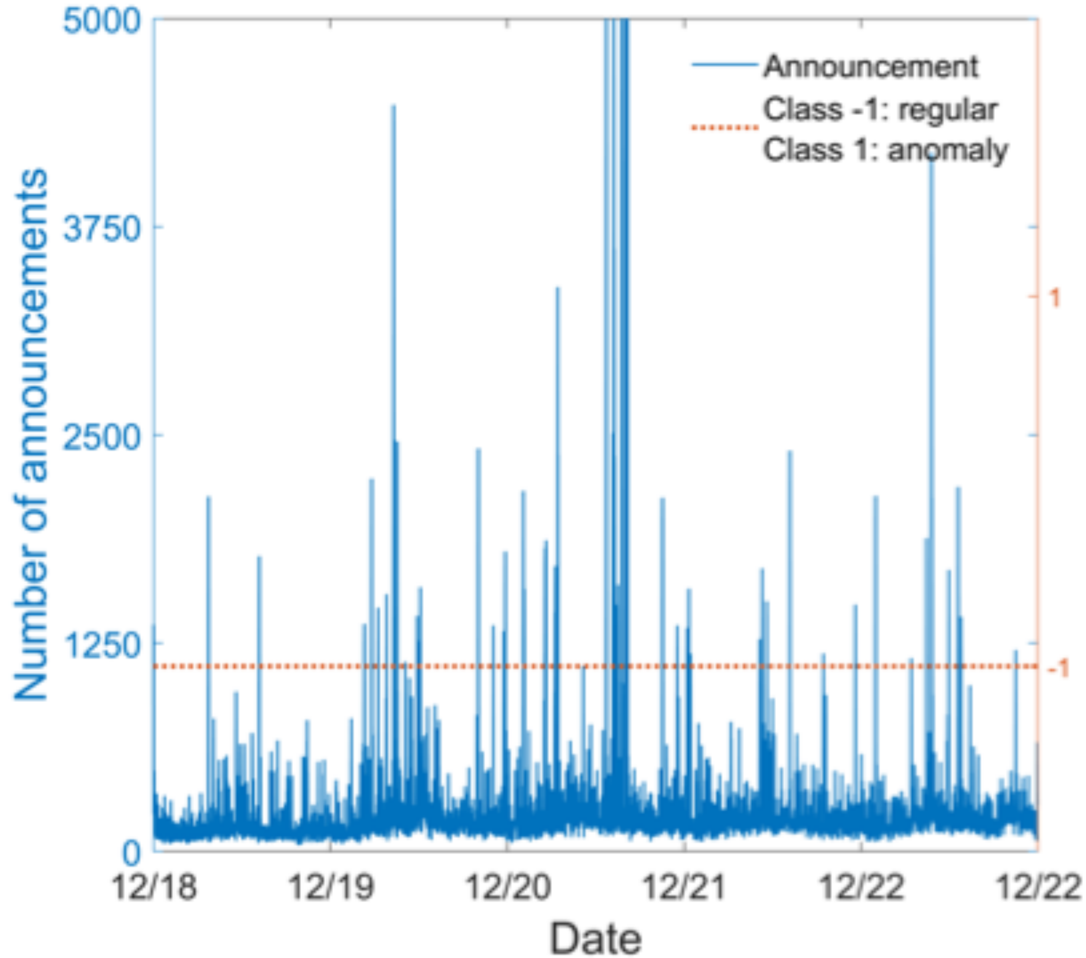
Number of BGP announcements: Nimda



Number of BGP announcements: Code Red I



Number of BGP announcements: Regular



Roadmap

- Introduction
  - Border Gateway Protocol (BGP)
  - Machine learning
- Feature extraction and selection
- Support vector machine and kernels
- Research contribution
- Experimental procedure and classification results
- Conclusions and future work
- References

Feature selection

- Reduces redundancy among features and improves the classification accuracy
- Decision tree algorithm was used for for feature selection:
  - one of the most successful techniques for supervised classification learning
- It can handle both numerical and categorical features
- Publicly available software tool: C5

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

Roadmap

- Introduction
  - Border Gateway Protocol (BGP)
  - Machine learning
- BGP Feature extraction and selection
- Support vector machine and kernels
- Research contributions
- Experimental procedure and classification results
- Conclusions and future work
- References

Support Vector Machine

- SVM defines a separating hyperplane in order to assign the target variables into distinct categories
- It is a non-probabilistic binary classifier
- Used for classification problems and in pattern recognition applications
- Modified version of logistic regression

Support Vector Machine

- For a given dataset **x** with *n* number of training data, SVM finds the maximum margin hyperplane separating different classes of data:

  $$\mathbf{x} = (\mathbf{x}_n, y_n), \mathbf{x}_n \in \boldsymbol{R}^p, y_n \in \{1, -1\}, \forall\, n = 1,2, \ldots, N$$

  - $\mathbf{x}_n$: p-dimensional input vector
  - y_n: output value (1 or -1)
- Decision vector separating two classes is given by:

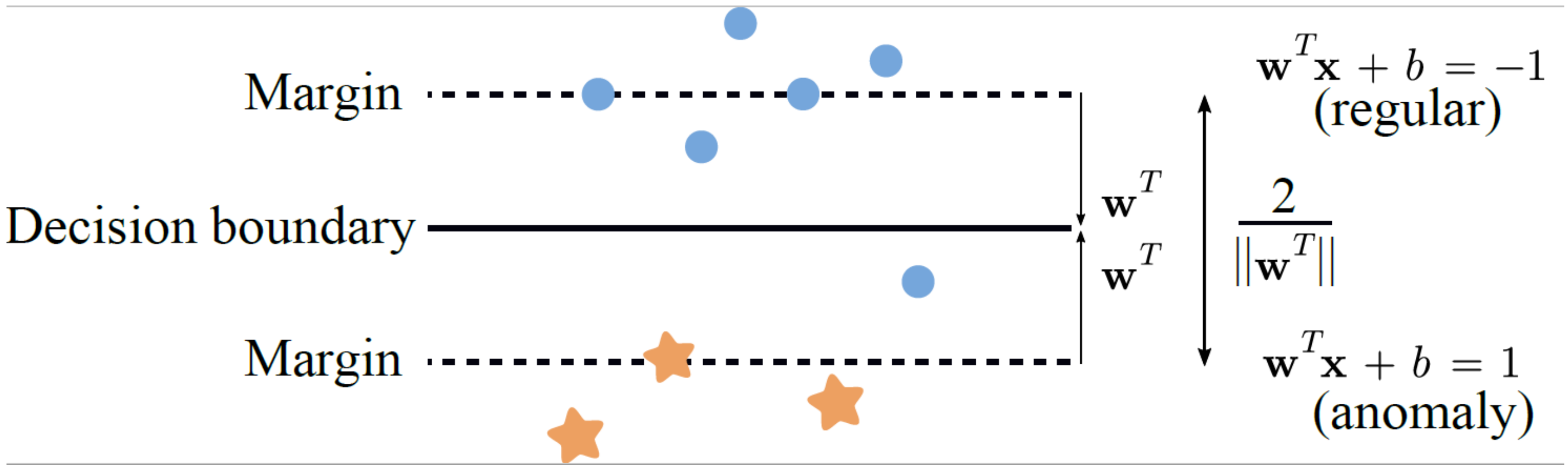
  $$\mathbf{w}^T \cdot \mathbf{x} + b = 0$$

  - $\mathbf{w}^T$: optimal weighing vector
  - *b:* bias

Support Vector Machine

- For linearly separable training data, margins are defined as:
  - $\mathbf{w}^T \cdot \mathbf{x} + b = 1$
  - $\mathbf{w}^T \cdot \mathbf{x} + b = -1$

Support Vector Machine



SVM with linear kernel: correctly classified regular (circles) and anomalous (stars) data points as well as one incorrectly classified regular (circle) data point

Support Vector Machine

- Distance between the margins: $2/\| \mathbf{w}^T \|$
- Objective function: minimize $\| \mathbf{w}^T \|$
- Let *C* be the regularization parameter that defines the separation of two classes and the error when using a training dataset. The hyperplane is acquired by minimizing the margins:

$$C \sum_{n=1}^{n} \zeta_n + \frac{1}{2} \| w \|^2,$$

with constraints $t_n y(x_n) \geq 1 - \zeta_n, n = 1, \ldots, N$

  - t_n: target value
  - ζ_n: set of slack variables

Support Vector Machine: kernel trick

- Instead of calculating each mapping, the "kernel trick" is used to directly calculate the inner product in the input space
- The mapping defines feature space and generates a decision boundary for input data points
- Using the "kernel trick" reduces the complexity of the optimization problem that now only depends on the input space instead if the feature space

Support Vector Machine

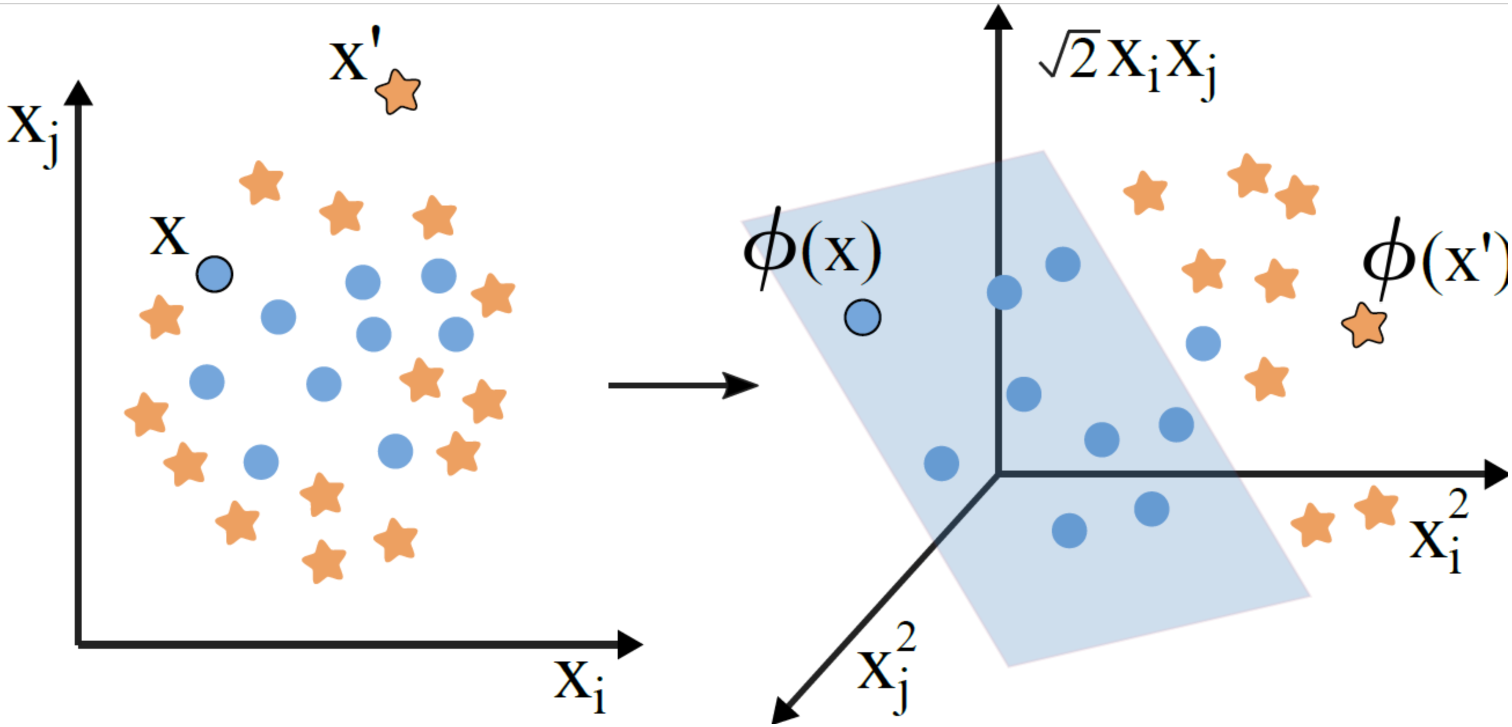
- Instead of employing a minimization model, the problem be formulated using Lagrangian dual multiplier β as:

$$\max \sum_{n=1}^{n} \beta_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \beta_n \beta_m y_n y_m \langle \mathbf{x}_n, \mathbf{x}_m \rangle,$$

subject to:

$$0 \le \beta_i \le C \; \forall \, i = 1,2,\ldots,n \text{ and } \sum_{i=1}^{n} \beta_i y_i = 0$$

Support Vector Machine



SVM with the nonlinear kernel function: the three-dimensional space shows a hyperplane dividing regular (circles) and anomalous (stars) data points

Roadmap

- Introduction
  - Border Gateway Protocol (BGP)
  - Machine learning
- Feature extraction and selection
- Support vector machine and kernels
- Research contributions
- Experimental procedure and classification results
- Conclusions and future work
- References

Research contributions

- Revised and extended our previous research findings and results by employing various SVM kernels for detecting anomalies
- Trained SVM with linear, polynomial, quadratic, cubic, Gaussian RBF, and sigmoid kernels
- Tested the models using various datasets
- Evaluated these SVM kernels based on accuracy and F-Score

Roadmap

- Introduction
  - Border Gateway Protocol (BGP)
  - Machine learning
- Feature extraction and selection
- Support vector machine and kernels
- Research contributions
- Experimental procedure and classification results
- Conclusions and future work
- References

Experimental procedure

Step 1:
  - Use 37 features or select the most relevant features using the decision tree algorithm
- Step 2:
  - Train the SVM with linear, polynomial, quadratic, cubic, Gaussian RBF, or sigmoid kernel
- Step 3:
  - Test the models using various datasets
- Step 4:
  - Evaluate the SVM kernels based on accuracy and F-Score

Training and test datasets

	Training dataset	Test dataset
Dataset 1	Slammer and Nimda	Code Red I
Dataset 2	Nimda and Code Red I	Slammer
Dataset 3	Slammer and Code Red I	Nimda
Dataset 4	Slammer	Nimda and Code Red I
Dataset 5	Nimda	Slammer and Code Red I
Dataset 6	Code Red I	Slammer and Nimda

Experimental procedure

- MATLAB 2019a Statistics and Machine Learning Toolbox
- The performance of SVM with various kernels is evaluated using combinations of datasets
- SVM performance was measured based on accuracy and F-Score
- The confusion matrix is used to evaluate performance of classification algorithms
- True positive (TP) and false negative (FN) are the number of anomalous data points that are classified as anomaly and regular, respectively

Performance measures

- Accuracy:
  - (TP+TN)/(TP+TN+FP+FN)
- F-Score signifies harmonic mean between precision and sensitivity:
  - 2 x (precision x sensitivity)/(precision + sensitivity)
  - precision: TP/(TP+FP)
  - sensitivity: TP/(TP+FN)

SVM with linear kernel

Linear kernel		Accuracy (%)			F-Score (%)
Selected features	Training dataset	Test	RIPE	BCNET	Test
1-37	Dataset 1	72.76	61.34	54.21	73.60
	Dataset 2	70.81	52.89	45.36	73.19
	Dataset 3	73.36	64.27	56.18	74.62
	Dataset 4	68.91	46.83	42.49	70.85
	Dataset 5	61.03	40.97	38.90	67.40
	Dataset 6	61.28	42.55	39.71	68.07

SVM with linear kernel

Linear kernel		Accuracy (%)			F-Score (%)
Selected features	Training dataset	Test	RIPE	BCNET	Test
1-21, 23-29, 34-37	Dataset 1	74.71	63.26	55.39	76.29
	Dataset 2	73.27	54.12	49.38	74.48
	Dataset 3	70.63	53.89	49.01	72.05
	Dataset 4	69.25	50.13	42.44	68.33
	Dataset 5	66.31	50.78	41.49	65.03
	Dataset 6	69.66	53.41	46.87	69.56

SVM with nonlinear kernels

Polynomial kernel		Accuracy (%)		F-Score (%)	
Selected features	Training dataset	Test	RIPE	BCNET	Test
1-37	Dataset 1	66.42	59.26	48.19	68.43
	Dataset 2	64.73	46.53	37.27	66.71
	Dataset 3	68.78	60.37	52.41	69.09
	Dataset 4	58.27	50.65	45.56	56.33
	Dataset 5	54.40	44.56	41.87	53.35
	Dataset 6	57.31	49.37	42.75	54.47

SVM with nonlinear kernels

Polynomial kernel		**Accuracy (%)**			**F-Score (%)**
Selected features	Training dataset	Test	RIPE	BCNET	Test
1-21, 23-29, 34-37	Dataset 1	70.26	59.43	49.86	74.39
	Dataset 2	67.51	46.69	40.73	69.84
	Dataset 3	66.80	45.38	37.41	67.05
	Dataset 4	63.02	42.95	36.03	65.73
	Dataset 5	60.29	41.24	33.92	64.24
	Dataset 6	65.37	44.50	38.10	68.59

SVM with nonlinear kernels

Quadratic kernel		Accuracy (%)			F-Score (%)
Selected features	Training dataset	Test	RIPE	BCNET	Test
1-37	Dataset 1	58.55	52.73	43.68	58.85
	Dataset 2	61.27	42.87	35.52	60.19
	Dataset 3	62.78	56.28	45.30	63.15
	Dataset 4	59.68	39.17	33.15	55.73
	Dataset 5	54.04	37.65	31.49	53.64
	Dataset 6	59.13	40.92	34.61	61.35

SVM with nonlinear kernels

Quadratic kernel		Accuracy (%)			F-Score (%)
Selected features	Training dataset	Test	RIPE	BCNET	Test
1-21, 23-29, 34-37	Dataset 1	63.84	58.51	46.39	67.24
	Dataset 2	63.36	46.55	38.73	64.68
	Dataset 3	62.53	43.30	37.12	63.09
	Dataset 4	57.40	40.59	34.78	60.33
	Dataset 5	55.58	37.13	30.53	58.29
	Dataset 6	60.21	41.85	35.17	62.48

SVM with nonlinear kernels

Cubic kernel		Accuracy (%)			F-Score (%)
Selected features	Training dataset	Test	RIPE	BCNET	Test
1-37	Dataset 1	65.33	54.31	45.53	58.85
	Dataset 2	63.15	46.23	40.17	65.49
	Dataset 3	68.83	57.57	46.47	70.03
	Dataset 4	59.50	41.44	35.88	62.15
	Dataset 5	50.37	35.47	30.17	55.28
	Dataset 6	59.28	38.04	33.20	58.33

SVM with nonlinear kernels

Cubic kernel		**Accuracy (%)**			**F-Score (%)**
Selected features	Training dataset	Test	RIPE	BCNET	Test
1-21, 23-29, 34-37	Dataset 1	69.21	58.12	49.26	70.14
	Dataset 2	67.79	49.78	42.36	69.55
	Dataset 3	65.58	48.20	40.44	66.92
	Dataset 4	58.70	41.56	35.18	56.66
	Dataset 5	55.19	37.23	32.71	51.58
	Dataset 6	61.05	45.23	38.23	61.35

SVM with nonlinear kernels

Gaussian RBF kernel		Accuracy (%)		F-Score (%)	
Selected features	Training dataset	Test	RIPE	BCNET	Test
1-37	Dataset 1	70.11	60.36	51.76	70.42
	Dataset 2	68.28	49.23	40.85	69.19
	Dataset 3	72.82	63.39	54.12	71.48
	Dataset 4	64.49	46.12	37.49	64.29
	Dataset 5	58.30	37.31	35.11	60.42
	Dataset 6	61.25	40.28	36.78	63.04

SVM with nonlinear kernels

Gaussian RBF kernel		Accuracy (%)			F-Score (%)
Selected features	Training dataset	Test	RIPE	BCNET	Test
1-21, 23-29, 34-37	Dataset 1	72.84	62.37	52.49	75.21
	Dataset 2	70.53	50.19	44.60	70.09
	Dataset 3	69.48	48.05	43.29	68.23
	Dataset 4	66.12	45.89	39.11	62.18
	Dataset 5	61.23	42.18	37.98	60.42
	Dataset 6	65.03	46.12	41.04	67.45

SVM with nonlinear kernels

Sigmoid kernel		Accuracy (%)			F-Score (%)
Selected features	Training dataset	Test	RIPE	BCNET	Test
1-37	Dataset 1	60.37	55.72	44.51	62.39
	Dataset 2	62.55	43.96	38.24	64.87
	Dataset 3	65.18	58.30	47.39	64.95
	Dataset 4	58.90	43.05	36.45	51.78
	Dataset 5	53.12	39.11	30.53	45.30
	Dataset 6	55.38	40.48	32.96	48.59

SVM with nonlinear kernels

Sigmoid kernel		Accuracy (%)			F-Score (%)
Selected features	Training dataset	Test	RIPE	BCNET	Test
1-21, 23-29, 34-37	Dataset 1	66.12	59.24	48.43	68.34
	Dataset 2	65.49	47.93	41.88	67.19
	Dataset 3	63.53	46.89	38.19	66.30
	Dataset 4	58.42	40.71	32.37	60.25
	Dataset 5	55.71	36.34	30.42	55.37
	Dataset 6	60.11	41.35	35.90	64.48

Roadmap

- Introduction
  - Border Gateway Protocol (BGP)
  - Machine learning
- Feature extraction and selection
- Support vector machine and kernels
- Research contributions
- Experimental procedure and classification results
- Conclusions and future work
- References

Conclusions

- SVM algorithm is one of the most efficient ML tools
- Kernels are used to transform the input data into a high dimensional space
- Their performance depends on both the feature selection and the type of datasets
- Analyzed BGP anomaly datasets are linearly separable
- SVM with linear and Gaussian RBF kernels outperform SVMs with polynomial, quadratic, cubic, and sigmoid kernels

Future work

- Perform concatenation, such as use 60% of data for training and 40% for testing
- Compare the present results achieved using the whole training and testing data versus 80%-20% or 60%-40% of training and testing data respectively

Roadmap

- Introduction
  - Border Gateway Protocol (BGP)
  - Machine learning
- Feature extraction and selection
- Support vector machine and kernels
- Research contributions
- Experimental procedure and classification results
- Conclusions and future work
- References

References: Data sources

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

References:

- Z. Li, P. Batta, and Lj. Trajković, "Comparison of machine learning algorithms for detection of network intrusions," in *Proc. IEEE International Conference on Systems, Man, and Cybernetics (SMC 2018)*, Miyazaki, Japan, Oct. 2018, pp. 4248-4253.
- P. Batta, M. Singh, Z. Li, Q. Ding, and Lj. Trajković, "Evaluation of support vector machine kernels for detecting network anomalies," in *Proc. IEEE Int. Symp. Circuits and Systems*, Florence, Italy, May 2018, pp. 1-4.
- Q. Ding, Z. Li, S. Haeri, and Lj. Trajković, "Application of machine learning techniques to detecting anomalies in communication networks: Datasets and Feature Selection Algorithms" in *Cyber Threat Intelligence,* M. Conti, A. Dehghantanha, and T. Dargahi, Eds., Berlin: Springer, 2018, pp. 47-70.
- Z. Li, Q. Ding, S. Haeri, and Lj. Trajković, "Application of machine learning techniques to detecting anomalies in communication networks: Classification Algorithms" in *Cyber Threat Intelligence,* M. Conti, A. Dehghantanha, and T. Dargahi, Eds., Berlin: Springer, 2018, pp. 71-92.
- Q. Ding, Z. Li, P. Batta, and Lj. Trajković, "Detecting BGP anomalies using machine learning techniques," in *Proc. IEEE International Conference on Systems, Man, and Cybernetics (SMC 2016),* Budapest, Hungary, Oct. 2016, pp. 3352-3355.
- Y. Li, H. J. Xing, Q. Hua, X.-Z. Wang, P. Batta, S. Haeri, and Lj. Trajković, "Classification of BGP anomalies using decision trees and fuzzy rough sets," in *Proc. IEEE International Conference on Systems, Man, and Cybernetics*, *SMC 2014*, San Diego, CA, October 2014, pp. 1312-1317.

References:

- N. Al-Rousan, S. Haeri, and Lj. Trajković, “Feature selection for classification of BGP anomalies using Bayesian models," in *Proc. International Conference on Machine Learning and Cybernetics, ICMLC 2012*, Xi'an, China, July 2012, pp. 140-147.
- N. Al-Rousan and Lj. Trajković, “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.

Acknowledgemnets

- Chair:
  - Prof. Ivan V. Bajić
- Senior supervisor:
  - Prof. Ljiljana Trajković
- Supervisor:
  - Prof. Parvaneh Saeedi
- SFU Examiner:
  - Mirza Faisal Beg

Thank you!

Questions?