Evaluation of Support Vector Machine Kernels for Detecting Network Anomalies

Prerna Batta, Maninder Singh, Zhida Li, Qingye Ding, and Ljiljana Trajković
Simon Fraser University
Vancouver, British Columbia, Canada
Email: {pbatta, msa198, zhidal, qingyed, ljilja}@sfu.ca

***Abstract*—Border Gateway Protocol (BGP) is used to exchange routing information across the Internet. BGP anomalies severely affect network performance and, hence, algorithms for anomaly detection are important for improving BGP convergence. Efficient and effective anomaly detection mechanisms rely on employing machine learning techniques. Support Vector Machine (SVM) is a widely used machine learning algorithm. In this paper, we evaluate performance of SVM with linear, quadratic, and cubic kernels. The SVM kernels are compared based on accuracy and the F-Score when detecting BGP anomalies in Internet traffic traces. The performance heavily depends on the selected features and their combinations.**

I. Introduction

Border Gateway Protocol (BGP) is a path vector protocol whose main function is to optimally route data between Autonomous Systems (ASes). An AS is a collection of BGP routers (peers) within a single administrative domain. BGP relies on Transmission Control Protocol (TCP), using port 179 for reliable router–to–router communication with only four message types (open, keepalive, update, and notification). Due to this simple design, BGP is prone to malicious attacks. In most cases, an attacked router advertises fraudulent information via BGP thus causing a large scale redirection of the Internet traffic [1], [2]. BGP anomalies include the worm attacks such as Slammer, Nimda, and Code Red I as well as routing misconfigurations, Internet Protocol (IP) prefix hijacks, and electrical failures. Statistical approaches have been extensively used for detection of BGP anomalies. However, they are not suitable to detect anomalies having high number of features. Rule-based detection techniques are based on prior knowledge of network conditions. They are not adaptable learning mechanisms, are slow, have high degree of computational complexity, and require a prior knowledge.

Machine learning models classify data using a feature matrix. The rows and columns of the matrix correspond to data points and feature values, respectively. By providing a sufficient number of relevant features, machine learning approaches help build a generalized model to classify data and lead to smaller number of classification errors. Machine learning models have been recently used to optimally solve a variety of engineering and scientific problems. Among various machine learning approaches such as Logistic Regression, Naive Bayes, Support Vector Machine (SVM), SVM is often used due to its robust nature [3]. It provides better generalization of features by defining decision boundary to geometrically lie midway between the support vectors. Support vectors are the data points that lie closest to the decision surface. SVM deals with both linearly and nonlinearly separable features of the input dataset by using kernels. Kernel functions map a nonlinearly separable into a higher-dimensional linearly separable data [4].

In this paper, we have revised and extended our previous research findings and results [5]–[10] by employing various SVM kernels for detecting BGP anomalies. The performance of anomaly classifiers is closely related to the selected features. Past approaches [5] to select BGP feature employed minimum Redundancy Maximum Relevance (mRMR) algorithms such as Mutual Information Deference (MID), Mutual Information Quotient (MIQ), and Mutual Information Base (MIBASE). In this paper, we employ the decision tree algorithm for feature selection and evaluate performance of SVM using linear, quadratic, and cubic kernels.

The paper is organized as follows. In Section II, we describe BGP features. The SVM algorithm and kernels are discussed in Section III. In Section IV, experimental setup is described. Classification models and their performance measures are introduced in Section V. We conclude with Section VI.

II. BGP Features and Anomalies

BGP messages are categorized as AS-*path* or *volume* features as shown in Table I [10]. They are BGP update message attributes that enable the protocol to select the best path for routing packets.

We consider three well-known BGP anomalies: Slammer, Nimda, and Code Red I, which occurred in January 2003, September 2001, and July 2001, respectively. The Route Views [11] and Réseaux IP Européens (RIPE) [12] BGP update messages are publicly available. Data from the RIPE Network Coordination Centre (NCC) were collected in December 2011 and reflect higher traffic volume due to the historical growth of the Internet. Their traffic traces are shown in Fig. 1. Slammer infected Microsoft SQL servers through a small piece of code that generated IP addresses at random. The number of infected machines doubled approximately every 9 seconds. Nimda exploited vulnerabilities in the Microsoft Internet Information Services (IIS) web servers for Internet Explorer 5. The worm propagated by sending an infected attachment that was automatically downloaded once the email was viewed. The Code Red I worm attacked Microsoft IIS web servers by replicating itself through IIS server weaknesses.

TABLE I
EXTRACTED BGP FEATURES

Feature	Definition	Type	Category
1	Number of announcements	continuous	volume
2	Number of withdrawals	continuous	volume
3	Number of announced NLRI prefixes	continuous	volume
4	Number of withdrawn NLRI prefixes	continuous	volume
5	Average AS-PATH length	categorical	AS-PATH
6	Maximum AS-PATH length	categorical	AS-PATH
7	Average unique AS-PATH length	continuous	volume
8	Number of duplicate announcements	continuous	volume
9	Number of duplicate withdrawals	continuous	volume
10	Number of implicit withdrawals	continuous	volume
11	Average edit distance	categorical	AS-PATH
12	Maximum edit distance	categorical	AS-PATH
13	Inter-arrival time	continuous	volume
14-24	Maximum edit distance = n, where n = (7;...; 17)	binary	AS-PATH
25-33	Maximum AS-PATH length = n, where n = (7;...;15)	binary	AS-PATH
34	Number of Interior Gateway Protocol packets	continuous	volume
35	Number of Exterior Gateway Protocol packets	continuous	volume
36	Number of incomplete packets	continuous	volume
37	Packet size (B)	continuous	volume

Unlike the Slammer worm, Code Red I searched for vulnerable servers to infect. The rate of infection was doubling every 37 minutes. Datasets containing BGP anomalies are collected from RIPE while regular datasets are collected from both RIPE and BCNET [13]. We use 37 features extracted from BGP update messages that originated from AS 513 (route collector rrc 04). The data were collected during periods of Internet anomalies. Five-day periods are considered: the day of the attack as well as two days prior and two days after the attack [5]–[10]. The duration of anomalies are listed in Table II.

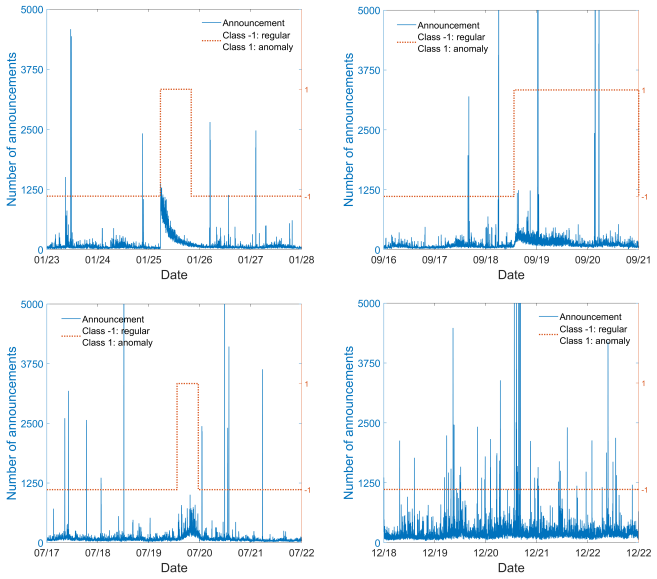

Fig. 1. Number of BGP announcements in Slammer (top left), Nimda (top right), Code Red I (bottom left), and regular RIPE (bottom right) traffic.

Feature selection reduces redundancy among features and improves the classification accuracy. Various combinations of extracted features also affect classification results. We use decision tree algorithm for feature selection implemented

TABLE II
DURATION OF ANALYZED BGP EVENTS

	Anomaly (min)	Regular (min)
Slammer	869	6,331
Nimda	3,521	3,679
Code Red I	600	6,600

in the publicly available software tool C5.0 [14]. Decision tree is one of the most successful techniques for supervised classification learning because it may handle both numerical and categorical features. Some features are removed from the constructed tree because they are repeatedly used.

III. SUPPORT VECTOR MACHINE

SVM is a discriminative classifier used for classification and regression tasks. SVM defines a separating hyperplane in order to assign the target variables into distinct categories. As a supervised learning algorithm, SVM is a non-probabilistic classifier that is used for classification problems and in pattern recognition applications [15]. It is a modified version of logistic regression [16], [17].

For a given dataset $\mathbf{x}$ with n number of training data, SVM finds the maximum margin hyperplane separating different classes of data:

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

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

$$\mathbf{w}^T \cdot \mathbf{x} + b = 0, \tag{2}$$

where $\mathbf{w}^T$ is the optimal weighing vector and b is the bias. For linearly separable training data, margins are defined as:

$$\mathbf{w}^T \cdot \mathbf{x} + b = 1 \text{ and } \quad \mathbf{w}^T \cdot \mathbf{x} + b = -1. \tag{3}$$

The distance between the margins is given by $2/||\mathbf{w}^T||$, as shown in Fig. 2. Hence, the objective function is to minimize $||\mathbf{w}^T||$. In practice, it is difficult to linearly separate the training dataset. 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 [3]:

$$C \sum_{n=1}^{n} \zeta_n + \frac{1}{2}||\mathbf{w}||^2, \tag{4}$$

with constraints $t_n y(x_n) \geqslant 1 - \zeta_n, \ n = 1, ..., N$, where t_n is the target value and ζ_n is the set of slack variables.

Instead of employing a minimization model (4), the problem may be formulated using Lagrangian dual multiplier β as:

$$\max \sum_{n=1}^{N} \beta_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \beta_n \beta_m y_n y_m \langle \mathbf{x}_n, \mathbf{x}_m \rangle, \tag{5}$$

subject to:

$$0 \leq \beta_i \leq C \ \ \forall \ \ i = 1, 2, ..., n, \ \text{ and } \ \sum_{i=1}^{n} \beta_i y_i = 0. \tag{6}$$

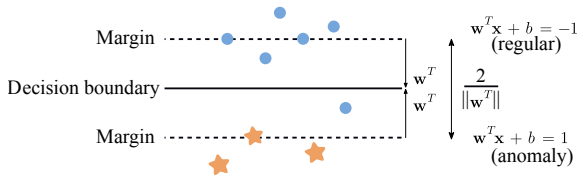

Fig. 2. Illustration of SVM with linear kernel: Shown are correctly classified regular (circles) and anomalous (stars) data points as well as one incorrectly classified regular (circle) data point.

Various SVM kernels may be employed for classification. If dataset is not linearly separable, SVM with linear kernel fails to perform. In such a scenario, polynomial nonlinear SVM kernels are used. Nonlinear kernels enlarge the feature space of input datasets to generate nonlinear boundary between the classes. This is achieved by mapping the data to a higher dimensional space. Such a boundary created by nonlinear kernels appears as a linear hyperplane and is used to extend SVM to nonlinear surfaces.

The kernel function defines inner products $\langle \mathbf{x}_n,\ \mathbf{x}_m \rangle$ or the similarity in the transformed space. Let $\phi(\mathbf{x})$ be a mapping of feature vectors from an input space to a feature space:

$$\langle \mathbf{x}_n, \mathbf{x}_m \rangle \rightarrow \langle \phi(\mathbf{x}_n),\ \phi(\mathbf{x}_m) \rangle. \tag{7}$$

The input space represents the selected features from the original dataset while the feature space is generated by the mapping $\phi(\mathbf{x})$. An example of mapping $\phi(\mathbf{x}) : \mathbf{R}^2 \rightarrow \mathbf{R}^3$ is shown in Fig. 3, where:

$$(x_i, x_j) \rightarrow (x_i^2, \sqrt{2}x_i x_j, x_j^2). \tag{8}$$

If $\mathbf{x}(x_1, x_2)$ and $\mathbf{x}'(x_1', x_2')$ are two feature vectors (data points) in the input space, then

$$\begin{aligned} &\langle \phi(x_1, x_2), \phi(x_1', x_2') \rangle = \\ &\langle ({x_1}^2, \sqrt{2}x_1 x_2, x_2^2), ({x_1'}^2, \sqrt{2}x_1' x_2', {x'}_2^2) \rangle = \\ &(x_1 x_1' + x_2 x_2')^2 = (\langle x, x' \rangle)^2 = k(x, x'). \end{aligned} \tag{9}$$

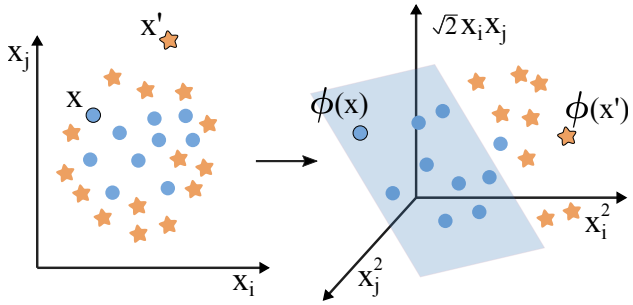

Fig. 3. Illustration of SVM using the nonlinear kernel function $k(\mathbf{x}_n, \mathbf{x}_m) = \langle \mathbf{x}_n, \mathbf{x}_m \rangle^2$. View in the three-dimensional space shows a hyperplane dividing regular (circles) and anomalous (stars) data points.

Instead of calculating each $\phi(\mathbf{x})$, the "kernel trick" is introduced to directly calculate the inner product in the input space:

$$k(\mathbf{x}_n, \mathbf{x}_m) = \langle \phi(\mathbf{x}_n),\ \phi(\mathbf{x}_m) \rangle, \tag{10}$$

where k is a kernel function. The mapping (10) 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 of the feature space.

The objective function for SVM with nonlinear kernel has the form [3]:

$$\max \sum_{n=1}^{N} \beta_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \beta_n \beta_m y_n y_m k(\mathbf{x}_n, \mathbf{x}_m). \tag{11}$$

In this study we use polynomial kernel functions:

$$k(\mathbf{x}_n,\ \mathbf{x}_m) = (\langle \mathbf{x}_n,\ \mathbf{x}_m \rangle + 1)^p, \tag{12}$$

where p is a parameter defining degree of the polynomial. Using polynomial kernel ($p > 1$) instead of the linear kernel ($p = 1$) results in flexible SVM decision boundary [17]. SVM classifies features of the input dataset mapped into higher dimensional space by using polynomials kernels of degree p.

IV. Experimental Procedure

In classification problems, unbalanced datasets are very frequently used. They have disproportionately high number of examples from one class. Since any classifier evaluated with unbalanced dataset tends to be biased towards one class, we use only balanced training datasets to evaluate SVM models using linear, quadratic, and cubic kernel. The classification procedure consists of four steps:

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

The performance of SVM with various kernels is evaluated using combinations of datasets shown in Table III. Experiments were performed using MATLAB 2017b with the Statistics and Machine Learning Toolbox.

TABLE III
Training and Test Datasets

	Training dataset	**Test dataset**
Dataset 1	Slammer and Nimda	Code Red I
Dataset 2	Nimda and Code Red I	Slammer

V. Performance Evaluation

We measure the SVM performance based on accuracy and F-Score. The confusion matrix [5] 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. Accuracy reflects the true prediction over the entire dataset. However, it assumes equal cost for misclassifications and a relatively uniform distribution of classes. Hence, we also use the F-Score that considers false predictions. F-Score signifies harmonic mean between precision and sensitivity that further measure

the discriminating ability of the classifier to identify classified and misclassified anomalies. The performance measures are:

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{13}$$

$$\text{F-Score} = 2 \times \frac{\text{precision} \times \text{sensitivity}}{\text{precision} + \text{sensitivity}}. \tag{14}$$

The precision and sensitivity are defined as:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}\ , \ \text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{15}$$

A. SVM with Linear Kernel

Performance of SVM with linear kernel for Dataset 1 and Dataset 2 is shown in Table IV. For both datasets, SVM with linear kernel gives the best performance when features are selected using the decision tree feature selection algorithm. The best accuracy (74.71%) and F-Score (76.29%) are obtained when the model is trained using Dataset 1.

TABLE IV
ACCURACY AND F-SCORE USING 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
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

B. SVM with Nonlinear Kernels

Results obtained using SVM with quadratic and cubic kernels are shown in Table V. SVM with quadratic and cubic kernels show the best accuracy (63.84% and 69.21%, respectively) and F-Score (67.24% and 70.14%, respectively) using the decision tree feature selection algorithm when tested using Dataset 1.

TABLE V
ACCURACY AND F-SCORE USING 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	63.15
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
Cubic kernel		Accuracy (%)			F-Score (%)
Selected features	Training dataset	Test	RIPE	BCNET	Test
1-37	Dataset 1	65.33	54.31	45.53	68.55
	Dataset 2	63.15	46.23	40.17	65.49
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

C. Performance Comparison

The best accuracy (74.71%) and F-Score (76.29%) are achieved using the linear SVM kernel when trained with Dataset 1. The results are obtained using dataset with the most relevant features rather than considering all 37 features. For both datasets, the SVM with linear and cubic kernels perform better than the SVM with quadratic kernel. The linear SVM kernel results in better accuracy and F-Score for both Dataset 1 and Dataset 2.

VI. CONCLUSION

The SVM algorithm is one of the most efficient machine learning tools. They use a set of mathematical functions called kernels that transform the input data into a high dimensional space before classifying the data points into distinct clusters. Examples of kernel functions are polynomial, Gaussian, Radian basis function, and sigmoid. The performance of an SVM kernel depends on both the feature selection and the type of dataset. If the input dataset is linearly separable, then linear SVM kernel performs better than other kernels. Analyzed BGP anomaly datasets are linearly separable and, hence, the linear SVM kernel outperforms quadratic and cubic SVM kernels. Therefore, using SVM classifier with linear kernels may serve as a feasible approach for detecting BGP anomalies in communication networks.

REFERENCES

[1] T. Ahmed, B. Oreshkin, and M. Coates, "Machine learning approaches to network anomaly detection," in *Proc. USENIX Workshop Tackling Comput. Syst. Problems with Mach. Learn. Techn.*, Cambridge, MA, USA, Apr. 2007, pp. 1–6.

[2] M. Bhuyan, D. Bhattacharyya, and J. Kalita, "Network anomaly detection: methods, systems and tools, *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 303–336, Mar. 2014.

[3] C. M. Bishop, *Pattern Recognition and Machine Learning*. Secaucus, NJ, USA: Springer-Verlag, 2006, pp. 325–358.

[4] T. Hofmann, B. Schölkopf, and A. J. Smola, "Kernel methods in machine learning," *The annals of statistics*, vol. 36, no. 3, pp. 1171–1220, 2008.

[5] Q. Ding, Z. Li, P. Batta, and Lj. Trajković, "Detecting BGP anomalies using machine learning techniques," in *Proc. IEEE Int. Conf. Syst., Man, and Cybern.*, Budapest, Hungary, Oct. 2016, pp. 3352–3355.

[6] 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 Trans. Syst., Man, Cybern.*, San Diego, CA, USA, Oct. 2014, pp. 1331–1336.

[7] Q. Ding, Z. Li, S. Haeri, and Lj. Trajkovic, "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, to appear.

[8] Z. Li, Q. Ding, S. Haeri, and Lj. Trajkovic, "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, to appear.

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

[10] N. Al-Rousan, S. Haeri, and Lj. Trajković, "Feature selection for classification of BGP anomalies using Bayesian models," in *Proc. Int. Conf. Mach. Learn. Cybern., ICMLC 2012*, Xi'an, China, July 2012, pp. 140–147.

[11] University of Oregon Route Views projects [Online]. Available: http://www.routeviews.org/. Accessed: Feb. 28, 2018.

[12] RIPE NCC [Online]. Available: http://www.ripe.net/data-tools/. Accessed: Feb. 28, 2018.

[13] BCNET [Online]. Available: http://www.bc.net/. Accessed: Feb. 28, 2018.

[14] C5.0 [Online]. Available: http://www.rulequest.com/see5-info.html/. Accessed: Feb. 28, 2018.

[15] M. M. Hossain and M. S. Miah, "Evaluation of different SVM kernels for predicting customer churn," in *18th Int. Conf. Comput. and Inform. Technology,* Dhaka, Bangladesh, Dec. 2015, pp. 1–4.

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

[17] C. Cortes and V. Vapnik, "Support-vector networks," *J. of Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sept. 1995.