# 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
E-mail: {nalrousa, shaeri, ljilja}@sfu.ca

**Abstract:**

**Traffic anomalies in communication networks greatly degrade network performance. Early detection of such anomalies alleviates their effect on network performance. A number of approaches that involve traffic modeling, signal processing, and machine learning techniques have been employed to detect network traffic anomalies.**

**In this paper, we develop various Naive Bayes (NB) classifiers for detecting the Internet anomalies using the Routing Information Base (RIB) of the Border Gateway Protocol (BGP). The classifiers are trained on the feature sets selected by various feature selection algorithms. We compare the Fisher, minimum redundancy maximum relevance (mRMR), extended/weighted/multi-class odds ratio (EOR/WOR/MOR), and class discriminating measure (CDM) feature selection algorithms. The odds ratio algorithms are extended to include continuous features. The classifiers that are trained based on the features selected by the WOR algorithm achieve the highest F-score.**

**Keywords:**

**Feature selection, classification, Bayesian model, BGP, traffic anomaly detection, network intrusion.**

## 1. Introduction

Anomalous events in communication networks cause traffic behavior to deviate from its usual profile. Hence, network traffic anomalies may be identified by analyzing traffic patterns. Various methods have been employed for detecting traffic anomalies. Early approaches include developing traffic models using statistical signal processing techniques. A baseline profile of network regular operation is developed based on a parametric model of traffic behavior and a large collection of traffic samples to account for all possible anomaly-free cases [1]. Anomalies may then be detected as sudden changes in the mean values of variables describing the baseline model. However, it is infeasible to acquire datasets that include all possible cases. In a network with quasi-stationary traffic, statistical signal processing methods may be employed to detect anomalies as correlated abrupt changes in network traffic [2].

In recent years, machine learning techniques have been employed for traffic classification. Unsupervised machine learning models are used to detect anomalies in networks with non-stationary traffic [3]. The one-class neighbor machine [4] and recursive kernel based online anomaly detection [5] algorithms are effective methods for detecting anomalous network traffic [6]. The Naive Bayes (NB) estimators perform well for categorizing the Internet traffic based on various applications [7].

The Border Gateway Protocol (BGP) [8] is used for routing packets between the Internet Autonomous Systems (ASs). BGP anomalies may be categorized as misconfigurations, worms, or blackouts. Rule based methods have been already employed for detecting anomalous BGP events [9]. However, they are non-adaptive, have high computational complexity, and require a priori knowledge of network conditions. A BGP anomaly detector has been proposed and implemented using statistical pattern recognition techniques [10]. For example, a Bayesian detection algorithm was designed to show that unexpected route misconfigurations may be identified as statistical anomalies [11]. An instance-learning framework may also employ wavelets to systematically identify anomalous BGP route advertisements [12]. We have recently evaluated Support Vector Machine (SVM) and Hidden Markov Model (HMM) classifiers for detecting BGP anomalies [13].

In the past, the main focus of proposed approaches was to develop models for traffic classification. However, the accuracy of a classifier depends on the underlying model, the extracted features, and the combination of features used for developing the model. In this paper, we address feature selection process to detect BGP anomalies. The employed algorithms belong to the category of filters, where feature selection is independent of the underlying learning algorithm [14].

This paper is organized as follows. In Section 2, we describe feature extraction from raw BGP data. A brief review of the employed feature selection algorithms is presented in Section 3. Design and implementation of the proposed NB classifiers are described in Section 4 while their performance is evaluated in Section 5. We conclude with Section 6.

## 2. Feature Extraction

BGP update messages are available to the research community through the Route Views project [15] and the Routing Information Service (RIS) project within the Réseaux IP Européens (RIPE) community [16]. The BGP messages are collected in multi-threaded routing toolkit (MRT) binary format [17]. The anomalous traffic traces are collected by RIPE during Slammer, Nimda, and Code Red I attacks. The list of collected anomalies along with regular (anomaly-free) datasets is given in Table 1.

We used the Zebra tool [18] to convert MRT data to ASCII format. We also developed a tool that employs the regular

Table 1: BGP datasets.

| | 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 |

expression library of C# to extract features from the ASCII files.

The BGP protocol generates four types of messages: open, update, keepalive, and notification. We only consider the BGP update messages because they contain all necessary features for anomaly classification. The extracted features are categorized into *volume* and *AS-path* features. The AS-PATH is a BGP update message attribute that enables the protocol to select the best path for routing packets. The update messages carry information about paths that BGP packets traverse. A feature is categorized as *AS-path* if it is derived from the AS-PATH attribute. Otherwise, it is categorized as a *volume* feature. Extracted features $\mathcal{F}$ and their categories are listed in Table 2.

Table 2: List of features extracted from BGP update messages.

| 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* |

BGP traffic features are sampled every minute within a five-day window. Hence, 7,200 samples are generated for each anomalous event. Samples from two days before and after each anomaly are used as regular test datasets. Each sample is a point in a 17-dimensional space, where $k^{th}$ dimension is a column vector $\mathbf{X}_k$ representing one feature. For example, $\mathbf{X}_1$ is a $7,200$ column vector representing the number of announcements in each sampling window of one minute. The scatterings of anomalous and regular classes for Feature 6 (*AS-path*) vs. Feature 1 (*volume*) and Feature 2 (*volume*) in two-way classification are shown in Figure 1 (top) and (bottom), respectively. The graphs indicate spatial separation of features. While selecting Features 1 and 6 may lead to a feasible classification based on visible clusters ($\mathsf{o}$and $*$), using only Features 2 and 6 would lead to poor classification. Hence, selecting appropriate combination of features is essential for an accurate classification.

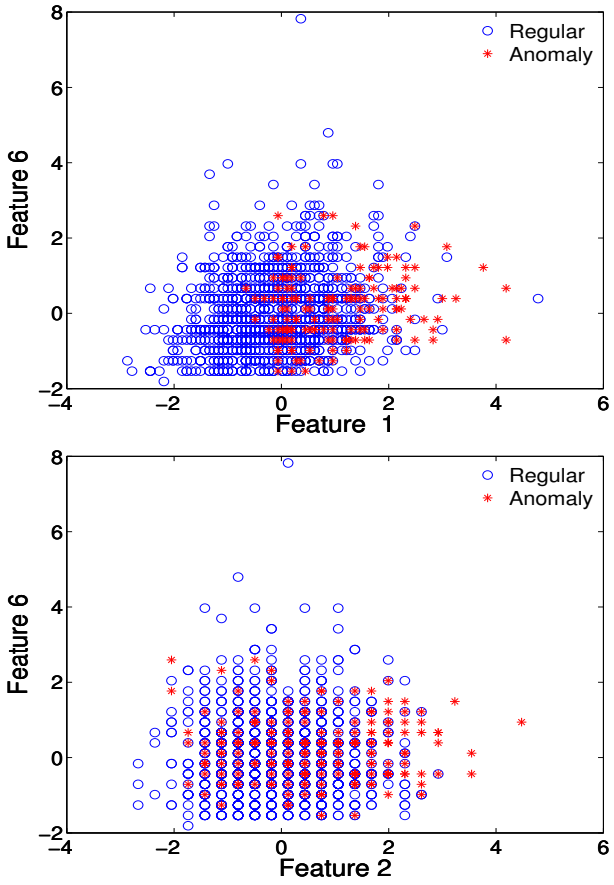


Figure 1: Scattered graph of Feature 6 vs. Feature 1 (top) and Feature 2 (bottom) extracted from the BCNET traffic. Feature values are normalized to have zero mean and unit variance. Shown are two traffic classes: regular ($\mathsf{o}$) and anomaly ($*$).

## 3. Feature Selection

Feature selection algorithms improve classification accuracy by selecting features that are most relevant to the classification task. We employ the Fisher [19], [20], three variants of the minimum Redundancy Maximum Relevance (mRMR) [21], extended/weighted/multi-class odds ratio (EOR/WOR/MOR), and class discriminating measure (CDM) [22] selection algorithms. We selected the top ten features while neglecting weaker and distorted features.

The Fisher feature selection algorithm computes the score $\Phi_k$ for the $k^{th}$ feature as a ratio of inter-class separation and intra-class variance. Features with higher inter-class separation and lower intra-class variance have higher Fisher scores. If there are $N_a^k$ anomalous samples and $N_r^k$ regular samples of the $k^{th}$ feature, the mean values $m_a^k$ of anomalous samples and $m_r^k$ of regular samples are calculated as

$$m_a^k = \frac{1}{N_a^k} \sum_{i \in \mathbf{a}_k} x_{ik}$$

$$m_r^k = \frac{1}{N_r^k} \sum_{i \in \mathbf{r}_k} x_{ik}, \qquad (1)$$

where $\mathbf{a}_k$ and $\mathbf{r}_k$ are the sets of anomalous and regular samples for feature $k$, respectively. The Fisher score for the $k^{th}$ feature is calculated as

$$\Phi_k = \frac{|\left(m_a^k\right)^2 - \left(m_r^k\right)^2|}{\frac{1}{N_a^k}\sum_{i\in\mathbf{a}_k}\left(x_{ik} - m_a^k\right)^2 + \frac{1}{N_r^k}\sum_{i\in\mathbf{r}_k}\left(x_{ik} - m_r^k\right)^2}. \quad (2)$$

The mRMR algorithm relies on an information theory approach for feature selection. It selects a subset of features that contains more information about the target class while having less pairwise mutual information. A subset of features $S = \{\mathbf{X_1}, \ldots, \mathbf{X_k}, \ldots\}$ with $|S|$ elements has the minimum redundancy if it minimizes

$$\mathcal{W} = \frac{1}{|S|^2}\sum_{\mathbf{X}_k,\mathbf{X}_l\in S}\mathcal{I}(\mathbf{X}_k, \mathbf{X}_l) \quad (3)$$

and maximum relevance to the classification task if it maximizes

$$\mathcal{V} = \frac{1}{|S|}\sum_{\mathbf{X}_k\in S}\mathcal{I}(\mathbf{X}_k, \mathbf{C}), \quad (4)$$

where $\mathbf{C}$ is a class vector and $\mathcal{I}$ denotes the mutual information function calculated as

$$\mathcal{I}(\mathbf{X}_k, \mathbf{X}_l) = \sum_{k,l} p(\mathbf{X}_k, \mathbf{X}_l) log \frac{p(\mathbf{X}_k, \mathbf{X}_l)}{p(\mathbf{X}_k)p(\mathbf{X}_l)}. \quad (5)$$

We use three variants of the mRMR algorithm for feature selection: Mutual Information Difference (MID), Mutual Information Quotient (MIQ), and Mutual Information Base (MIBASE). If $\Omega$ is the set of all features, MID and MIQ select the features based on $\max_{S\subset\Omega}[\mathcal{V} - \mathcal{W}]$ and $\max_{S\subset\Omega}[\mathcal{V}/\mathcal{W}]$, respectively.

The odds ratio (OR) algorithm and its variants perform well for selecting features to be used in binary classification with NB models. In case of a binary classification with two target classes $c$ and $\bar{c}$, the odds ratio for a feature $\mathbf{X}_k$ is

$$OR(\mathbf{X}_k) = \log \frac{\Pr(\mathbf{X}_k|c)\big(1 - \Pr(\mathbf{X}_k|\bar{c})\big)}{\Pr(\mathbf{X}_k|\bar{c})\big(1 - \Pr(\mathbf{X}_k|c)\big)}, \quad (6)$$

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.

The extended odds ratio (EOR), weighted odds ratio (WOR), multi-class odds ratio (MOR), and class discriminating measure (CDM) are variants that enable multi-class feature selection. In case of a classification problem with $\gamma = \{c_1, c_2, \ldots, c_J\}$ classes,

$$\begin{aligned}
EOR(\mathbf{X}_k) &= \sum_{j=1}^{J} \log \frac{\Pr(\mathbf{X}_k|c_j)\big(1 - \Pr(\mathbf{X}_k|\bar{c_j})\big)}{\Pr(\mathbf{X}_k|\bar{c_j})\big(1 - \Pr(\mathbf{X}_k|c_j)\big)} \\
WOR(\mathbf{X}_k) &= \sum_{j=1}^{J} \Pr(c_j) \times \log \frac{\Pr(\mathbf{X}_k|c_j)\big(1 - \Pr(\mathbf{X}_k|\bar{c_j})\big)}{\Pr(\mathbf{X}_k|\bar{c_j})\big(1 - \Pr(\mathbf{X}_k|c_j)\big)} \\
MOR(\mathbf{X}_k) &= \sum_{j=1}^{J} \left| \log \frac{\Pr(\mathbf{X}_k|c_j)\big(1 - \Pr(\mathbf{X}_k|\bar{c_j})\big)}{\Pr(\mathbf{X}_k|\bar{c_j})\big(1 - \Pr(\mathbf{X}_k|c_j)\big)} \right| \\
CDM(\mathbf{X}_k) &= \sum_{j=1}^{J} \left| \log \frac{\Pr(\mathbf{X}_k|c_j)}{\Pr(\mathbf{X}_k|\bar{c_j})} \right|, \quad (7)
\end{aligned}$$

where $\Pr(\mathbf{X}_k|c_j)$ is the conditional probability of $\mathbf{X}_k$ given the class $c_j$ and $\Pr(c_j)$ is the probability of occurrence of the $j^{th}$ class. The OR algorithm may be extended by computing $\Pr(\mathbf{X}_k|c_j)$ for continuous features. If the sample points are independent and identically distributed, (6) may be written as

$$OR(\mathbf{X}_k) = \sum_{i=1}^{|\mathbf{X}_k|} \log \frac{\Pr(X_{ik} = x_{ik}|c)\big(1 - \Pr(X_{ik} = x_{ik}|\bar{c})\big)}{\Pr(X_{ik} = x_{ik}|\bar{c})\big(1 - \Pr(X_{ik} = x_{ik}|c)\big)},$$

where $|\mathbf{X}_k|$ and $X_{ik}$ denote the size and the $i^{th}$ element of the $k^{th}$ feature vector, respectively. A realization of the random variable $X_{ik}$ is denoted by $x_{ik}$. Other variants of the OR algorithm may be extended to continuous cases in a similar manner. The top ten selected features are listed in Table 3.

## 4. Classification with Naive Bayes (NB)

The Bayesian classifiers are among the most efficient machine learning classification tools. They assume conditional independence among features. Hence,

$$\begin{aligned}
\Pr(\mathbf{X}_k = \boldsymbol{x}_k, \mathbf{X}_l = \boldsymbol{x}_l|c_j) = \\
\Pr(\mathbf{X}_k = \boldsymbol{x}_k|c_j)\Pr(\mathbf{X}_l = \boldsymbol{x}_l|c_j),
\end{aligned} \quad (8)$$

where $\boldsymbol{x}_k$ and $\boldsymbol{x}_l$ are realizations of feature vectors $\mathbf{X}_k$ and $\mathbf{X}_l$, respectively. In a two-way classification, classes $c_1$ and $c_2$ denote anomalous and regular data points, respectively. We arbitrarily assign labels $c_1 = 1$ and $c_2 = -1$. For a four-way classification, we define four classes $c_1 = 1$, $c_2 = 2$, $c_3 = 3$, and $c_4 = 4$ that denote Slammer, Nimda, Code Red I, and Regular data points, respectively. Even though it is naive to assume that features are independent conditioned on a given class (8), in certain applications NB classifiers perform better compared to other classifiers. They also have low complexity and may be trained effectively with smaller datasets.

We train generative Bayesian models that may be used as classifiers using labeled datasets. In such models, the probability distributions of the priors $\Pr(c_j)$ and the likelihoods $\Pr(\mathbf{X}_i = \boldsymbol{x}_i|c_j)$ are estimated using the training datasets. Posterior of a data point represented as a row vector $\boldsymbol{x}_i$ is calculated using the Bayes rule

$$\begin{aligned}
\Pr(c_j|\mathbf{X}_i = \boldsymbol{x}_i) &= \frac{\Pr(\mathbf{X}_i = \boldsymbol{x}_i|c_j)\Pr(c_j)}{\Pr(\mathbf{X}_i = \boldsymbol{x}_i)} \\
&\approx \Pr(\mathbf{X}_i = \boldsymbol{x}_i|c_j)\Pr(c_j). \quad (9)
\end{aligned}$$

The naive assumption of independence among features helps calculate the likelihood of a data point as

$$\Pr(\mathbf{X}_i = \boldsymbol{x}_i|c_j) = \prod_{k=1}^{K} \Pr(X_{ik} = x_{ik}|c_j), \quad (10)$$

where $K$ denotes the number of features. The probabilities on the right-hand side (10) are calculated using the Gaussian distribution

$$\Pr(X_{ik} = x_{ik}|c_j, \mu_k, \sigma_k) = \mathcal{N}(X_{ik} = x_{ik}|c_j, \mu_k, \sigma_k), \quad (11)$$

Table 3: The top ten selected features $\mathcal{F}$ based on the scores calculated by various feature selection algorithms.

| 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 |

where $\mu_k$ and $\sigma_k$ are the mean and standard deviation of the $k^{th}$ feature, respectively. We assume that priors are equal to the relative frequencies of the training data points for each class $c_j$. Hence,

$$\Pr(c_j) = \frac{N_j}{N}, \tag{12}$$

where $N_j$ is the number of training data points that belong to the $j^{th}$ class and $N$ is the total number of training points.

The parameters of two-way and four-way classifiers are estimated and validated by a tenfold cross-validation. In a two-way classification, an arbitrary training data point $\boldsymbol{x}_i$ is classified as anomalous if the posterior $\Pr(c_1|\mathbf{X}_i = \boldsymbol{x}_i)$ is larger than $\Pr(c_2|\mathbf{X}_i = \boldsymbol{x}_i)$.

## 5. Performance Evaluation

We use the MATLAB statistical toolbox to develop NB classifiers. The feature matrix consists of 7,200 rows for each dataset corresponding to the number of training data points and 17 columns representing features for each data point. Two classes are targeted: anomalous (true) and regular (false). In a two-way classification, all anomalies are treated to be of one type while in a four-way classification, each training data point is classified as Slammer, Nimida, Code Red I, or Regular. We use three datasets listed in Table 4 to train the two-way classifiers. Performance of two-way and four-way classifiers is evaluated using various datasets. The results are verified by using regular RIPE and regular BCNET [23] datasets. The regular BCNET dataset is collected at the BCNET location in Vancouver, British Columbia, Canada [24], [25].

Table 4: 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 |

The proposed classifiers are trained using the top selected features listed in Table. 3. We use the accuracy and F-score to compare the proposed models. The performance measures are calculated as

$$\begin{aligned} \text{accuracy} &= \frac{TP + TN}{TP + TN + FP + FN} \\ \text{F-score} &= 2 \times \frac{\text{precision} \times \text{sensitivity}}{\text{precision} + \text{sensitivity}}, \end{aligned} \tag{13}$$

where

$$\begin{aligned} \text{sensitivity} &= \frac{TP}{TP + FN} \\ \text{precision} &= \frac{TP}{TP + FP}. \end{aligned} \tag{14}$$

Furthermore,

- true positive (TP) is the number of anomalous training data points that are classified as anomaly;
- true negative (TN) is the number of regular training data points that are classified as regular;
- false positive (FP) is the number of regular training data points that are classified as anomaly;
- false negative (FN) is the number of anomalous training data points that are classified as regular.

The sensitivity measures the ability of the model to identify the anomalies (true positives) among all labeled anomalies (true). The precision is the ability of the model to identify the anomalies (true positives) among all data points that are classified as anomalies (positives). The accuracy treats the regular data points to be as important as the anomalous training points. Hence, it is not an adequate measure when comparing performance of classifiers. For example, if a dataset contains 900 regular and 100 anomalous data points and the NB classifies these 1,000 data points as regular, its accuracy is 90%, which seems high at the first glance. However, no anomalous data point is correctly classified and, hence, the F-score is zero. Therefore, the F-score is often used to compare performance of classification models. It is the harmonic mean of the precision and the sensitivity and reflects the success of detecting anomalies rather than detecting both anomalies and regular data points.

### 5.1 Two-Way Classification

The results of the two-way classification are shown in Table 5. The combination of Code Red I and Nimda training data points (NB3) achieves the best classification results. The NB models classify the training data points of regular RIPE and regular BCNET datasets with 95.8% and 95.5% accuracies, respectively. There are no anomalous data points in these datasets and, thus, both TP and FN values are zero. Hence, the sensitivity is not defined and precision is equal to zero. Consequently, the F-score is not defined for these cases and the accuracy reduces to

$$\text{accuracy} = \frac{TN}{TN + FP}. \tag{15}$$

Table 5: Performance of the two-way naive Bayes classification.

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

Classifiers trained based on features selected by the OR algorithms often achieve higher accuracies and F-scores for training and test datasets listed in Table 4. The OR selection algorithms perform well when used with the NB classifiers because the feature score (6) is calculated using the probability distribution that the NB classifiers use for posterior calculations (9). Hence, the features selected by the OR variants are expected to have stronger influence on the posteriors calculated by the NB classifiers [26]. The WOR feature selection algorithm achieves the best F-score for all NB classifiers.

The Slammer worm test data points that are incorrectly classified (false positives and false negatives) using the NB3 classifier trained based on the features selected by WOR in the two-way classification are shown in Figure 2 (top). Correctly classified anomalies (true positives) during the 16 hours time interval are shown in Figure 2 (bottom). Most anomalous data points with large number of IGP packets (*volume* feature) are correctly classified.

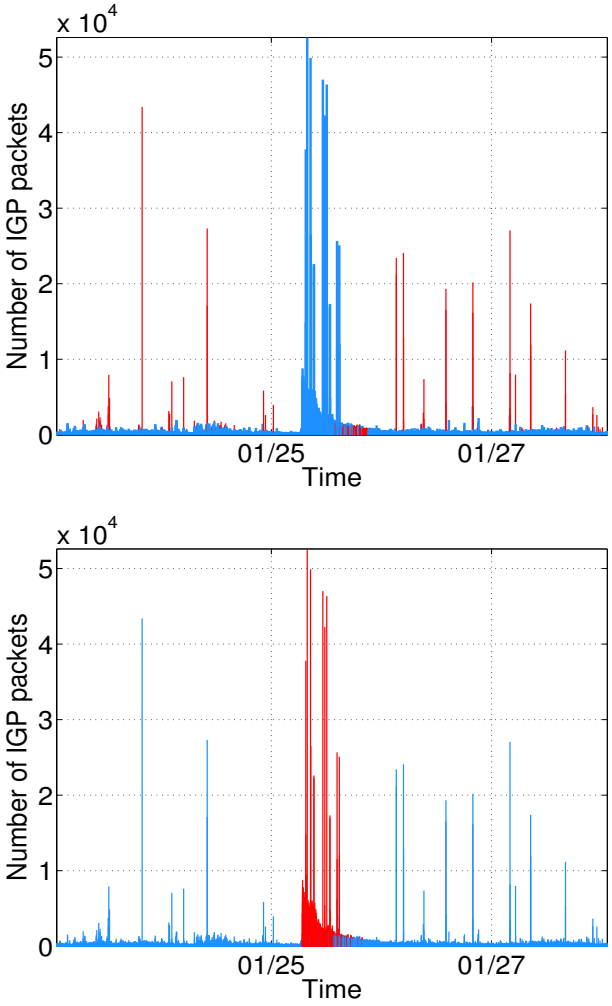


Figure 2: Shown in red are incorrectly classified regular (false positives) and anomaly (false negatives) data points (top) and correctly classified anomaly (true positives) data points (bottom) on January 25, 2003. Correctly classified regular (true negatives) traffic is not shown.

### 5.2 Four-Way Classification

The four-way classification results are shown in Table 6. The four-way NB model classifies data points as Slammer, Nimda, Code Red I, or Regular. Both regular RIPE and regular BCNET datasets are tested. Regular BCNET dataset classification results are also listed in order to verify the performance of the proposed classifiers. Although it is more difficult to classify four distinct

classes, the classifier trained based on the features selected by the MOR algorithm achieves 68.7% accuracy.

Table 6: Accuracy of the four-way naive Bayes classification.

| No. | Feature | Average accuracy (%) | |
|---|---|---|---|
| | | 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 |

Performance of the NB classifiers is often inferior to the SVM and HMM classifiers [13]. However, the NB2 classifier trained on Slammer and Code Red I datasets performs better than the SVM classifier.

## 6. Conclusions

In this paper, we successfully classified anomalies in BGP traffic traces using NB models. We employed various feature selection algorithms and generative NB models to design anomaly detectors. We extended the usage of the OR algorithms from categorical to continuous features. The OR algorithms often achieved higher F-scores in the two-way and four-way classifications with various training datasets. The NB classifiers may be used for online detection of anomalies because they have low complexity and may be trained effectively on smaller datasets.

## References


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

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

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] Y. Rekhter and T. Li, "A Border Gateway Protocol 4 (BGP-4)," RFC 1771, *IETF*, Mar. 1995.

[9] 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.

[10] S. Deshpande, M. Thottan, T. K. Ho, and B. Sikdar, "An online mechanism for BGP instability detection and analysis," *IEEE Trans. Comput.*, vol. 58, no. 11, pp. 1470–1484, Nov. 2009.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] University of Oregon Route Views project. [Online]. Available: http://www.routeviews.org/.

[16] RIPE RIS raw data. [Online]. Available: http://www.ripe.net/data-tools/stats/ris/ris-raw-data.

[17] T. Manderson, "Multi-threaded routing toolkit (MRT) Border Gateway Protocol (BGP) routing information export format with geo-location extensions," RFC 6397, *IETF*, Oct. 2011.

[18] Zebra BGP parser. [Online]. Available: http://www.linux.it/~md/software/zebra-dump-parser.tgz.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] BCNET. [Online]. Available: http://www.bc.net.

[24] 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.

[25] 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.

[26] 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.