Machine Learning Models for Feature Selection and Classification of Traffic Anomalies

Ljiljana Trajković
Simon Fraser University
Vancouver, British Columbia, Canada
Email: ljilja@cs.sfu.ca
http://www.sfu.ca/~ljilja/cnl

May 14, 2013
Roadmap

1. Introduction
2. Data Processing
   - Extraction of features
   - Selection of features
3. Performance Evaluation
4. Classification with Support Vector Machines
5. Classification with Hidden Markov Models
6. Classification with Naive Bayes
7. BGP Anomaly Detection (BGPAD tool)
8. Discussions and Conclusions
9. References
Introduction

- BGP anomalies also include: Internet Protocol (IP) prefix hijacks, miss-configurations, and electrical failures.
- BGP anomalies often occur.
- Techniques for BGP anomalies detection have recently gained visible attention and importance.
Roadmap

1. Introduction
2. Data Processing
   - Extraction of features
   - Selection of features
3. Performance Evaluation
4. Classification with Support Vector Machines
5. Classification with Hidden Markov Models
6. Classification with Naive Bayes
7. BGP Anomaly Detection (BGPAD tool)
8. Discussions and Conclusions
9. References
Datasets sources

- The RIPE and Route Views BGP update messages: multi-threaded routing toolkit (MRT) binary format
- BGP traffic traces collected from the BCNET

<table>
<thead>
<tr>
<th>Class</th>
<th>Date</th>
<th>Duration (h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slammer</td>
<td>January 25, 2003</td>
<td>16</td>
</tr>
<tr>
<td>Nimda</td>
<td>September 18, 2001</td>
<td>59</td>
</tr>
<tr>
<td>Code Red I</td>
<td>July 19, 2001</td>
<td>10</td>
</tr>
<tr>
<td>RIPE regular</td>
<td>July 14, 2001</td>
<td>24</td>
</tr>
<tr>
<td>BCNET</td>
<td>December 20, 2011</td>
<td>24</td>
</tr>
</tbody>
</table>

References

List of extracted features

- **Extracted features**: volume (number of BGP announcements) and AS-path (maximum edit distance) features:

<table>
<thead>
<tr>
<th>Feature</th>
<th>Definition</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Number of announcements</td>
<td>volume</td>
</tr>
<tr>
<td>2</td>
<td>Number of withdrawals</td>
<td>volume</td>
</tr>
<tr>
<td>3</td>
<td>Number of announced NLRI prefixes</td>
<td>volume</td>
</tr>
<tr>
<td>4</td>
<td>Number of withdrawn NLRI prefixes</td>
<td>volume</td>
</tr>
<tr>
<td>5</td>
<td>Average AS-PATH length</td>
<td>AS-path</td>
</tr>
<tr>
<td>6</td>
<td>Maximum AS-PATH length</td>
<td>AS-path</td>
</tr>
<tr>
<td>7</td>
<td>Average unique AS-PATH length</td>
<td>AS-path</td>
</tr>
<tr>
<td>8</td>
<td>Number of duplicate announcements</td>
<td>volume</td>
</tr>
<tr>
<td>9</td>
<td>Number of duplicate withdrawals</td>
<td>volume</td>
</tr>
<tr>
<td>10</td>
<td>Number of implicit withdrawals</td>
<td>volume</td>
</tr>
<tr>
<td>11</td>
<td>Average edit distance</td>
<td>AS-path</td>
</tr>
<tr>
<td>12</td>
<td>Maximum edit distance</td>
<td>AS-path</td>
</tr>
<tr>
<td>13</td>
<td>Inter-arrival time</td>
<td>AS-path</td>
</tr>
<tr>
<td>14-24</td>
<td>Maximum edit distance = n, where n = (7, ..., 17)</td>
<td>AS-path</td>
</tr>
<tr>
<td>25-33</td>
<td>Maximum AS-path length = n, where n = (7, ..., 15)</td>
<td>AS-path</td>
</tr>
<tr>
<td>34</td>
<td>Number of IGP packets</td>
<td>volume</td>
</tr>
<tr>
<td>35</td>
<td>Number of EGP packets</td>
<td>volume</td>
</tr>
<tr>
<td>36</td>
<td>Number of incomplete packets</td>
<td>volume</td>
</tr>
<tr>
<td>37</td>
<td>Packet size (B)</td>
<td>volume</td>
</tr>
</tbody>
</table>
Normalized scattering graphs

- Feature 1, feature 2, and feature 6:

- Selecting appropriate combination of features is essential for an accurate classification
Feature selection algorithms

- Features scoring algorithms:
  - Fisher
  - Minimum Redundancy Maximum Relevance (mRMR)
  - Odds Ratio

- These algorithms measure the correlation and relevancy among features

- The top ten features were selected for the Fisher feature selection

References


Fisher algorithm

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

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

$$= \frac{1}{N_a} \sum_{i \in \text{anomaly}} x_{ik}^2 - \frac{1}{N_r} \sum_{i \in \text{regular}} x_{ik}^2$$

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

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

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

- mRMR relevance between a feature set $S = \{X_1, ..., X_k, X_l, ..., X_{37}\}$ and a class vector $Y$ is based on the mutual information function $\mathcal{I}$:

$$
\mathcal{I}(X_k, X_l) = \sum_{k,l} p(X_k, X_l) \log \frac{p(X_k, X_l)}{p(X_k)p(X_l)}
$$

- Criteria for mRMR variants:

  - MID: $\max [V(\mathcal{I}) - W(\mathcal{I})]$
  - MIQ: $\max [V(\mathcal{I})/W(\mathcal{I})]$

$$
V(\mathcal{I}) = \frac{1}{|S|} \sum_{X_k \in S} \mathcal{I}(X_k, Y)
$$

$$
W(\mathcal{I}) = \frac{1}{|S|^2} \sum_{X_k, X_l \in S} \mathcal{I}(X_k, X_l)
$$
Odds Ratio algorithm

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

\[
OR(X_k) = \log \frac{\Pr(X_k|c)(1 - \Pr(X_k|\bar{c}))}{\Pr(X_k|\bar{c})(1 - \Pr(X_k|c))},
\]

where \(\Pr(X_k|c)\) and \(\Pr(X_k|\bar{c})\) are the probabilities of feature \(X_k\) being in classes \(c\) and \(\bar{c}\), respectively.
EOR, WOR, MOR, and CDM Algorithms

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

\[
EOR(X_k) = \sum_{j=1}^{J} \log \frac{\Pr(X_k|c_j)(1 - \Pr(X_k|\bar{c}_j))}{\Pr(X_k|\bar{c}_j)(1 - \Pr(X_k|c_j))}
\]

\[
WOR(X_k) = \sum_{j=1}^{J} \Pr(c_j) \times \log \frac{\Pr(X_k|c_j)(1 - \Pr(X_k|\bar{c}_j))}{\Pr(X_k|\bar{c}_j)(1 - \Pr(X_k|c_j))}
\]

\[
MOR(X_k) = \sum_{j=1}^{J} \left| \log \frac{\Pr(X_k|c_j)(1 - \Pr(X_k|\bar{c}_j))}{\Pr(X_k|\bar{c}_j)(1 - \Pr(X_k|c_j))} \right|
\]

\[
CDM(X_k) = \sum_{j=1}^{J} \left| \log \frac{\Pr(X_k|c_j)}{\Pr(X_k|\bar{c}_j)} \right|
\]

where

- \( \Pr(X_k|c_j) \) is the conditional probability of \( X_k \) given the class \( c_j \)
- \( \Pr(c_j) \) is the probability of occurrence of the \( j^{th} \) class
The top ten selected features

<table>
<thead>
<tr>
<th>Fisher</th>
<th>mRMR</th>
<th>Odds Ratio variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>MID</td>
<td>MIQ</td>
<td>MIBASE</td>
</tr>
<tr>
<td>F</td>
<td>Score</td>
<td>F</td>
</tr>
<tr>
<td>11</td>
<td>0.397758</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>0.354740</td>
<td>5</td>
</tr>
<tr>
<td>9</td>
<td>0.271961</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>0.185844</td>
<td>7</td>
</tr>
<tr>
<td>16</td>
<td>0.123742</td>
<td>4</td>
</tr>
<tr>
<td>17</td>
<td>0.121633</td>
<td>10</td>
</tr>
<tr>
<td>8</td>
<td>0.116092</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>0.086124</td>
<td>13</td>
</tr>
<tr>
<td>1</td>
<td>0.081760</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>0.081751</td>
<td>14</td>
</tr>
</tbody>
</table>
Roadmap

1. Introduction
2. Data Processing
   - Extraction of features
   - Selection of features
3. Performance Evaluation
4. Classification with Support Vector Machines
5. Classification with Hidden Markov Models
6. Classification with Naive Bayes
7. BGP Anomaly Detection (BGPAD tool)
8. Discussions and Conclusions
9. References
Definitions

- We considered: accuracy, balanced accuracy, and F-score

Definitions:

- True positive (TP): is number of anomalous training data points that are classified as anomaly
- True negative (TN): is number of regular training data points that are classified as regular
- False positive (FP): is number of regular training data points that are classified as anomaly
- False negative (FN): is number of anomalous training data points that are classified as regular

<table>
<thead>
<tr>
<th>Actual class</th>
<th>True (anomaly)</th>
<th>False (regular)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anomaly test outcome</td>
<td>Positive</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>FN</td>
</tr>
</tbody>
</table>

May 14, 2013 National Cheng Kung University, Tainan, Taiwan
Performance measures and indices

- Performance measures:
  
sensitivity = \frac{TP}{TP + FN}

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

- Performance indices:
  
  accuracy = \frac{TP + TN}{TP + TN + FP + FN}

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

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

1. Introduction
2. Data Processing
   - Extraction of features
   - Selection of features
3. Performance Evaluation
4. Classification with Support Vector Machines
5. Classification with Hidden Markov Models
6. Classification with Naive Bayes
7. BGP Anomaly Detection (BGPAD tool)
8. Discussions and Conclusions
9. References
Support Vector Machines

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

References

### SVM two-way datasets

<table>
<thead>
<tr>
<th>NB</th>
<th>Training dataset</th>
<th>Test dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM&lt;sub&gt;1&lt;/sub&gt;</td>
<td>Slammer and Nimda</td>
<td>Code Red I</td>
</tr>
<tr>
<td>SVM&lt;sub&gt;2&lt;/sub&gt;</td>
<td>Slammer and Code Red I</td>
<td>Nimda</td>
</tr>
<tr>
<td>SVM&lt;sub&gt;3&lt;/sub&gt;</td>
<td>Code Red I and Nimda</td>
<td>Slammer</td>
</tr>
</tbody>
</table>
## Two-way classification: performance

- All anomalies are treated as one class

<table>
<thead>
<tr>
<th>SVM</th>
<th>Feature</th>
<th>Accuracy (%)</th>
<th>F-score (%)</th>
<th>SVM</th>
<th>Feature</th>
<th>Accuracy (%)</th>
<th>F-score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Test dataset (anomaly)</td>
<td>RIPE (regular)</td>
<td>BCNET (regular)</td>
<td>Test dataset (anomaly)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM1</td>
<td>All features</td>
<td>64.1</td>
<td>55.0</td>
<td>62.0</td>
<td>63.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM1</td>
<td>Fisher</td>
<td>72.6</td>
<td>63.2</td>
<td>58.5</td>
<td>73.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM1</td>
<td>MID</td>
<td>63.1</td>
<td>52.2</td>
<td>59.4</td>
<td>61.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM1</td>
<td>MIQ</td>
<td>60.7</td>
<td>47.9</td>
<td>61.7</td>
<td>57.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM1</td>
<td>MIBASE</td>
<td>79.1</td>
<td>74.3</td>
<td>60.9</td>
<td>80.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM2</td>
<td>All features</td>
<td>68.6</td>
<td>97.7</td>
<td>79.2</td>
<td>22.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM2</td>
<td>Fisher</td>
<td>67.4</td>
<td>96.6</td>
<td>74.8</td>
<td>16.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM2</td>
<td>MID</td>
<td>67.9</td>
<td>97.4</td>
<td>72.5</td>
<td>19.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM2</td>
<td>MIQ</td>
<td>67.7</td>
<td>97.5</td>
<td>76.2</td>
<td>15.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM2</td>
<td>MIBASE</td>
<td>67.5</td>
<td>96.8</td>
<td>78.8</td>
<td>17.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM3</td>
<td>All features</td>
<td>81.5</td>
<td>92.0</td>
<td>69.2</td>
<td>84.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM3</td>
<td>Fisher</td>
<td>89.3</td>
<td>93.8</td>
<td>68.4</td>
<td>75.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM3</td>
<td>MID</td>
<td>75.4</td>
<td>92.8</td>
<td>71.7</td>
<td>79.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM3</td>
<td>MIQ</td>
<td>85.1</td>
<td>92.2</td>
<td>73.2</td>
<td>86.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM3</td>
<td>MIBASE</td>
<td>89.3</td>
<td>89.7</td>
<td>69.7</td>
<td>80.1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Classification results

- SVM$_3$ achieves the best F-score (86.1%) using features selected by MIQ.
- BCNET and RIPE test datasets contain no anomalies and have low F-scores:
  - Performance measure: accuracy
  - SVM$_2$: the best overall two-way classifier
- Incorrectly classified (anomaly) BCNET traffic collected on December 20, 2011 (red):
Classification results

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

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

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

<table>
<thead>
<tr>
<th>Feature</th>
<th>RIPE regular</th>
<th>BCNET</th>
</tr>
</thead>
<tbody>
<tr>
<td>All features</td>
<td>77.1</td>
<td>91.4</td>
</tr>
<tr>
<td>Fisher</td>
<td>82.8</td>
<td>85.7</td>
</tr>
<tr>
<td>MID</td>
<td>67.8</td>
<td>78.7</td>
</tr>
<tr>
<td>MIQ</td>
<td>71.3</td>
<td>89.1</td>
</tr>
<tr>
<td>MIBASE</td>
<td>72.8</td>
<td>90.2</td>
</tr>
</tbody>
</table>

References

Roadmap

1. Introduction
2. Data Processing
   - Extraction of features
   - Selection of features
3. Performance Evaluation
4. Classification with Support Vector Machines
5. Classification with Hidden Markov Models
6. Classification with Naive Bayes
7. BGP Anomaly Detection (BGPAD tool)
8. Discussions and Conclusions
9. References
Hidden Markov Models

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

- HMM model is specified by a tuple \( \lambda = (N, M, \alpha, \beta, \pi) \):
  - \( N \) = number of hidden states (cross-validated)
  - \( M \) = number of observations (11)
  - \( \alpha \) = transition probability distribution \( N \times N \) matrix
  - \( \beta \) = emission probability distribution \( N \times M \) matrix
  - \( \pi \) = initial state probability distribution matrix

- The proposed detection model consists of three stages:
  - **Sequence extractor and mapping**: all features are mapped to 1-D observation vector
  - **Training**: two HMMs for two-way classification and four HMMs for four-way classification are trained to identify the best \( \alpha \) and \( \beta \) for each class
  - **Classification**: maximum likelihood probability \( p(x|\lambda) \) is used to classify the test observation sequences
Classification

- HMMs with the same number of hidden states are compared
- Example: HMM\(_1\), HMM\(_4\), HMM\(_7\), and HMM\(_{10}\) correspond to HMMs with two hidden states for various training datasets
- HMM accuracy:

\[
\text{HMM accuracy} = \frac{\text{Number of correctly classified observation sequences}}{\text{Total number of observation sequences}}
\]
Classification

- The correctly classified observation sequence is generated by a model that has the highest probability when tested with itself.
- RIPE and BCNET were datasets to test the three anomalies.
- Two sets of features (volume) and (AS-path) are mapped to create one observation sequence for each HMM.
- volume feature set (1, 2) and AS-path feature set (6, 12) are mapped to two observation sequences.
- RIPE and BCNET test datasets have the highest F-score when tested using HMMs with two hidden states.
Two-way classification: performance

<table>
<thead>
<tr>
<th>$N$</th>
<th>Feature set</th>
<th>RIPE regular</th>
<th>BCNET</th>
<th>RIPE regular</th>
<th>BCNET</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>(1,2)</td>
<td>86.0</td>
<td>94.0</td>
<td>84.4</td>
<td>93.8</td>
</tr>
<tr>
<td>2</td>
<td>(6,12)</td>
<td>79.0</td>
<td>71.0</td>
<td>76.2</td>
<td>60.7</td>
</tr>
<tr>
<td>4</td>
<td>(1,2)</td>
<td>78.0</td>
<td>87.0</td>
<td>72.2</td>
<td>85.0</td>
</tr>
<tr>
<td>4</td>
<td>(6,12)</td>
<td>64.0</td>
<td>60.0</td>
<td>48.0</td>
<td>35.9</td>
</tr>
<tr>
<td>6</td>
<td>(1,2)</td>
<td>85.0</td>
<td>91.0</td>
<td>84.3</td>
<td>90.1</td>
</tr>
<tr>
<td>6</td>
<td>(6,12)</td>
<td>81.0</td>
<td>65.0</td>
<td>80.1</td>
<td>50.2</td>
</tr>
</tbody>
</table>

- HMMs have better F-score using set (1, 2) than set (6, 12)
Four-way classification: performance

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

<table>
<thead>
<tr>
<th>N</th>
<th>Feature set</th>
<th>RIPE regular</th>
<th>BCNET</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>(1,2)</td>
<td>72.50</td>
<td>77.50</td>
</tr>
<tr>
<td>2</td>
<td>(6,12)</td>
<td>38.75</td>
<td>41.25</td>
</tr>
<tr>
<td>4</td>
<td>(1,2)</td>
<td>66.25</td>
<td>76.25</td>
</tr>
<tr>
<td>4</td>
<td>(6,12)</td>
<td>26.25</td>
<td>33.75</td>
</tr>
<tr>
<td>6</td>
<td>(1,2)</td>
<td>70.00</td>
<td>76.25</td>
</tr>
<tr>
<td>6</td>
<td>(6,12)</td>
<td>43.75</td>
<td>42.50</td>
</tr>
</tbody>
</table>
Roadmap

1. Introduction
2. Data Processing
   - Extraction of features
   - Selection of features
3. Performance Evaluation
4. Classification with Support Vector Machines
5. Classification with Hidden Markov Models
6. Classification with Naive Bayes
7. BGP Anomaly Detection (BGPAD tool)
8. Discussions and Conclusions
9. References
Naive Bayes

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

$$
\Pr(X_k = x_k, X_l = x_l|c_j) = \Pr(X_k = x_k|c_j) \Pr(X_l = x_l|c_j)
$$

- $x_k$ is realization of feature vector $X_k$
- $x_l$ is realization of feature vector $X_l$

- 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 $x_i$ is calculated using the Bayes rule:

$$\Pr(c_j | X_i = x_i) = \frac{\Pr(X_i = x_i | c_j) \Pr(c_j)}{\Pr(X_i = x_i)}$$

$$\approx \Pr(X_i = x_i | c_j) \Pr(c_j)$$

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

$$\Pr(X_i = x_i | c_j) = \prod_{k=1}^{K} \Pr(X_{ik} = x_{ik} | c_j)$$
Likelihoods and priors

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

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

- $N_j$ 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):

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

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

- **two-way classification:**
  \[
  \max\{\Pr(c_1|X_i = x_i), \Pr(c_2|X_i = x_i)\}
  \]

- **four-way classification:**
  \[
  \max\{\Pr(c_1|X_i = x_i), \Pr(c_2|X_i = x_i), \Pr(c_3|X_i = x_i), \Pr(c_4|X_i = x_i)\}
  \]

**Example (two-way classification):**

An arbitrary training data point \(x_i\) is classified as anomalous if

\[
\Pr(c_1|X_i = x_i) > \Pr(c_2|X_i = x_i)
\]
## Two-way classification: performance

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

<table>
<thead>
<tr>
<th>No.</th>
<th>Feature set</th>
<th>RIPE regular</th>
<th>BCNET</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>All features</td>
<td>74.3</td>
<td>67.6</td>
</tr>
<tr>
<td>2</td>
<td>Fisher</td>
<td>24.7</td>
<td>34.3</td>
</tr>
<tr>
<td>3</td>
<td>MID</td>
<td>74.9</td>
<td>33.1</td>
</tr>
<tr>
<td>4</td>
<td>MIQ</td>
<td>24.6</td>
<td>34.8</td>
</tr>
<tr>
<td>5</td>
<td>MIBASE</td>
<td>75.4</td>
<td>33.1</td>
</tr>
<tr>
<td>6</td>
<td>OR</td>
<td>25.5</td>
<td>36.7</td>
</tr>
<tr>
<td>7</td>
<td>EOR</td>
<td>75.3</td>
<td>68.1</td>
</tr>
<tr>
<td>8</td>
<td>WOR</td>
<td>75.8</td>
<td>53.2</td>
</tr>
<tr>
<td>9</td>
<td>MOR</td>
<td>77.7</td>
<td>68.7</td>
</tr>
<tr>
<td>10</td>
<td>CDM</td>
<td>24.8</td>
<td>34.5</td>
</tr>
</tbody>
</table>
Classification results: **Slammer worm (January 25, 2003)**

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

1. Introduction
2. Data Processing
   - Extraction of features
   - Selection of features
3. Performance Evaluation
4. Classification with Support Vector Machines
5. Classification with Hidden Markov Models
6. Classification with Naive Bayes
7. BGP Anomaly Detection (BGPAD tool)
8. Discussions and Conclusions
9. References
BGPAD tool: Inspects BGP pcap and MRT files for anomalies
BGPAD tool:
Provides test performance indices
BGPAD tool:
Displays anomalous traffic
Roadmap

1. Introduction
2. Data Processing
   - Extraction of features
   - Selection of features
3. Performance Evaluation
4. Classification with Support Vector Machines
5. Classification with Hidden Markov Models
6. Classification with Naive Bayes
7. BGP Anomaly Detection (BGPAD tool)
8. Discussions and Conclusions
9. References
Discussion: feature extraction and selection

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

- **Effect of anomalies on AS-path features:**
  - Large length of the AS-PATH BGP attribute implies that the packet is routed via a longer path to its destination.
  - Very short lengths of AS-PATH attributes occur during BGP hijacks when the new (false) originator usually gains a preferred or shorter path to the destination.
  - Edit distance and AS-PATH length of the BGP announcements tend to have a very high or a very low value (large variance).

- The top selected AS-path features appear on the boundaries of the distributions: AS-path features 25, 32, and 24 have the highest Fisher, MID, and MIQ scores.
Discussion: classification

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

- RIPE regular and BCNET test datasets contain no anomalies and have low F-scores. For example, in two-way NB:
  - Performance measure (accuracy):
    - RIPE regular: 95.8%
    - BCNET: 95.5%

- OR algorithms often achieve better performance:
  - feature score is calculated using the probability distribution that the NB classifiers use for posterior calculations
  - features selected by the OR variants are expected to have stronger influence on the posteriors
Discussion: classification

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

References

Discussion: comparison of features categories

- The volume features accounted for 65% of selected features.
- Two-way SVM classification with only volume and with only AS-path features were applied.
- Performance of SVM using volume features was superior to AS-path.

<table>
<thead>
<tr>
<th>SVM</th>
<th>category</th>
<th>accuracy</th>
<th>precision</th>
<th>sensitivity</th>
<th>specificity</th>
<th>balanced accuracy</th>
<th>f-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM1</td>
<td>volume</td>
<td>68.5</td>
<td>53.6</td>
<td>16.6</td>
<td>73.2</td>
<td>44.9</td>
<td>27.1</td>
</tr>
<tr>
<td>SVM1</td>
<td>AS-path</td>
<td>56.4</td>
<td>6.12</td>
<td>29.5</td>
<td>58.8</td>
<td>44.1</td>
<td>3.93</td>
</tr>
<tr>
<td>SVM2</td>
<td>volume</td>
<td>87.0</td>
<td>69.6</td>
<td>12.5</td>
<td>99.1</td>
<td>55.8</td>
<td>22.3</td>
</tr>
<tr>
<td>SVM2</td>
<td>AS-path</td>
<td>86.0</td>
<td>38.7</td>
<td>1.19</td>
<td>99.6</td>
<td>50.4</td>
<td>2.36</td>
</tr>
<tr>
<td>SVM3</td>
<td>volume</td>
<td>94.8</td>
<td>79.7</td>
<td>76.4</td>
<td>97.3</td>
<td>86.8</td>
<td>85.0</td>
</tr>
<tr>
<td>SVM3</td>
<td>AS-path</td>
<td>56.9</td>
<td>19.1</td>
<td>79.4</td>
<td>53.8</td>
<td>66.6</td>
<td>64.1</td>
</tr>
</tbody>
</table>
Discussion: performance comparison

- Statistical pattern recognition approach has difficulty in estimating distributions of high dimensions
- Rule-based techniques require a priori knowledge of network conditions
- Rule based: better results in two out of three datasets
- Behavioural: worse results in all the three datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>SVM (two-way)</th>
<th>SVM (four-way)</th>
<th>HMM (two-way)</th>
<th>HMM (four-way)</th>
<th>NB (two-way)</th>
<th>NB (four-way)</th>
<th>Rule based techniques</th>
<th>Behavioural techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slammer</td>
<td>89.3</td>
<td>82.8</td>
<td>86.0</td>
<td>70.0</td>
<td>87.4</td>
<td>77.7</td>
<td>94.4</td>
<td>74.0</td>
</tr>
<tr>
<td>Nimda</td>
<td>68.6</td>
<td>82.8</td>
<td>86.0</td>
<td>70.0</td>
<td>70.1</td>
<td>77.7</td>
<td>84.1</td>
<td>74.0</td>
</tr>
<tr>
<td>Code Red I</td>
<td>79.1</td>
<td>82.8</td>
<td>86.0</td>
<td>70.0</td>
<td>74.1</td>
<td>77.7</td>
<td>74.9</td>
<td>74.0</td>
</tr>
</tbody>
</table>

References

Conclusions

- We have investigated BGP anomalies and proposed detection models based on the SVM, HMM, and NB classifiers.
- Various feature selection algorithms and models were employed to design BGP anomaly detectors.
- The OR algorithms often achieved higher F-scores in the two-way and four-way classifications with various training datasets.
- **Volume** features are more relevant to the anomaly class than the **AS-path** features.
Conclusions

- Anomalies in BGP traffic traces were successfully classified using the proposed models.
- The best achieved F-scores: SVM (86.1%), HMM (84.4%), and NB (69.7%).
- The proposed NB models may be used as online mechanisms to predict new BGP anomalies and detect the onset of worm attacks.
Roadmap

1. Introduction
2. Data Processing
   - Extraction of features
   - Selection of features
3. Performance Evaluation
4. Classification with Support Vector Machines
5. Classification with Hidden Markov Models
6. Classification with Naive Bayes
7. BGP Anomaly Detection (BGPAD tool)
8. Discussions and Conclusions
9. References
References: http://www.sfu.ca/~ljilja/cnl


References: BGP anomalies


References: BGP


References: machine learning


References: feature selection


References: naive Bayes

