Open Source Software Tools for Anomaly Detection Analysis

by Robert F. Erbacher and Robinson Pino

ARL-MR-0869

April 2014

Approved for public release; distribution unlimited.
NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer’s or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.
Open Source Software Tools for Anomaly Detection Analysis

Robert F. Erbacher and Robinson Pino
Computational and Information Sciences Directorate, ARL
Open Source Software Tools for Anomaly Detection Analysis

Robert F. Erbacher and Robinson Pino

The goal of this report is to perform an analysis of software tools that could be employed to perform basic research and development of Anomaly-Based Intrusion Detection Systems. The software tools reviewed include: Environment for Developing KDD-Applications Supported by Index-Structures (ELKI), RapidMiner, SHOGUN (toolbox) Waikato Environment for Knowledge Analysis (Weka) (machine learning), and Scikit-learn. From the analysis, it is recommended to employ the SHOGUN (toolbox) or Scikit-learn as both tools are written in C++ and offers an interface for Python. The python language software is currently employed as a research tool within our in-house team of researchers.
### Contents

List of Figures iv
List of Tables iv

1. Introduction 1

2. Environment for Developing KDD-Applications Supported by Index Structures (ELKI) 1

3. RapidMiner 2

4. SHOGUN (toolbox) 3

5. Waikato Environment for Knowledge Analysis (Weka) (Machine Learning) 4

6. Scikit-Learn 5

7. Results and Discussion 6

8. Conclusions 7

9. References 9

Appendix – List of Algorithm Per Tool 11

List of Symbols, Abbreviations, and Acronyms 15

Distribution List 16
List of Figures

Figure 1. ELKI (a) user interface and (b) output results (4) .......................................................... 2
Figure 2. RapidMiner output results (7) .......................................................................................... 3
Figure 3. Screenshot of SHOGUN software tool results (8) ............................................................ 4
Figure 4. Screenshot of Weka software tool results (10) ................................................................. 5
Figure 5. Scikit-learn results performing binary classification using nonlinear Support Vector
Classification (SVC) with Radial Basis Function (RBF) kernel. The target to predict is an Exclusive Or (XOR) of the inputs. The color map illustrates the decision function learned by the SVC (14). .................................................................................................................. 6

List of Tables

Table 1. Side-by-side comparison of algorithms offered by SHOGUN and Scikit-learn .............. 7
1. Introduction

Anomaly-based intrusion detection is the concept for detecting computer intrusions and misuse by monitoring network and computer activity and classifying it as either normal or anomalous \(^1\). Classification is commonly based on heuristics or rules, rather than patterns or signatures, and will detect any type of misuse that falls out of normal system operation \(^2\). In the case of signature based detection, the system can only detect attacks for which a signature has previously been created. In order to determine what attack traffic is, the system must be taught to recognize normal system activity. This can be accomplished using artificial intelligence techniques or neural networks \(^1\). Another method is to define what normal usage of the system comprises using a strict mathematical model, and flag any deviation from this as an attack, known as strict anomaly detection \(^1\). The goal of this report is to determine the suitability of current open source software packages in their usage and ability to enable our in-house team of researchers to perform basic research on anomaly-based intrusion detection algorithms.

2. Environment for Developing KDD-Applications Supported by Index Structures (ELKI)

The software tool ELKI stands for Environment for Developing KDD-Applications Supported by Index Structures and is a knowledge discovery in databases (KDD), data mining, and software framework developed for use in research and teaching by the database systems research unit of Professor Hans-Peter Kriegel at the Ludwig Maximilian University of Munich, Germany \(^3\). The ELKI software package is written in Java\(^*\) and intended to allow development and a platform for independent evaluation of data mining algorithms \(^4\). The software framework is open source for scientific usage; see figure 1 for an overview.

\(^*\) Java is a registered trademark of Oracle.
ELKI provides a suite of algorithms that include: K-means clustering, anomaly detection, spatial index structures, apriori algorithm, dynamic time warping, and principal component analysis. However, an internet search for publications using this particular software application platform yields results authored by the software developers. In 2011, a book chapter published by Achert et al. (5) talks about spatial outlier detection: data, algorithms, and visualization. The manuscript focuses on showcasing ELKI’s ability to integrate a geographic/geospatial information system (GIS) and a data mining system (DMS) within a single framework supported by the tool. In the demonstration, the authors demonstrated an integrated GIS-DMS system for performing advanced data mining tasks such as outlier detection on geospatial data, but which also allows the interaction with an existing GIS (5).

3. RapidMiner

The RapidMiner software, formerly Yet Another Learning Environment (YALE), is an environment for machine learning, data mining, text mining, predictive analytics, and business analytics. It is used for research, education, training, rapid prototyping, application development, and industrial applications (6). The software is distributed under the Affero General Public License (AGPL) open source license and has been hosted by SourceForge since 2004 (7).
RapidMiner provides data mining and machine learning procedures including: data loading and transformation (extract, transform, load [ETL]), data preprocessing and visualization, modeling, evaluation, and deployment. Two examples of graphical results are shown in figure 2. The data mining processes can be made up of arbitrarily nestable operators, described in eXtensible Markup Language (XML) files and created in RapidMiner's graphical user interface (GUI). RapidMiner is written in the Java programming language, and can be used for text mining, multimedia mining, feature engineering, data stream mining and tracking drifting concepts, development of ensemble methods, and distributed data mining (7). In addition, advanced features can be purchased as a commercial version of the base software and are available at the rapid-i.com Web site. In particular, beyond the free community edition, three enterprise software packages can be purchased from small, standard, and developer editions, which offer an increasing number of options and capabilities, respectively. The algorithms included in the RapidMiner software include: machine learning, data mining, text mining, predictive analytics, and business analytics.

Figure 2. RapidMiner output results (7).

4. **SHOGUN (toolbox)**

The focus of SHOGUN is on kernel machines such as support vector machines for regression and classification problems. SHOGUN also offers a full implementation of Hidden Markov models. The core of SHOGUN is written in C++ and offers interfaces for MATLAB,* Octave, Python,† R, Java, Lua, Ruby, and C#. SHOGUN has been under active development since 1999. Today there is a user community all over the world using SHOGUN as a base for research and

---

*MATLAB is a registered trademark of The MathWorks, Inc.
†Python is a registered trademark of Python Software Foundation.
education, and contributing to the core package. SHOGUN is a free software, open source toolbox written in C++. It offers numerous algorithms and data structures for machine learning problems. SHOGUN is licensed under the terms of the GNU General Public License version 3 or later (8). Figure 3 shows a screenshot of SHOGUN’s code and output results. The software can be obtained from the official Web site (9).

Among the software tool packages reviewed, SHOGUN offers the most features for research and development. Some of SHOGUN’s algorithms and features included in the software tool are: support vector machines, dimensionality reduction, online learning, clustering, and implemented kernels for numeric data analysis algorithms. Over 20 publications about SHOGUN are featured on its Wikipedia page (8).

5. Waikato Environment for Knowledge Analysis (Weka) (Machine Learning)

Weka is a suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. The Weka workbench contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with GUIs for easy access to this functionality. Weka is free software available under the GNU General Public License (10). The Weka software is available for download at the official Web site (11), see figure 4.
Some of the features of this software tool include: data preprocessing, clustering, expectation maximization, classification, regression, visualization, and feature selection. The primary learning methods in Weka are “classifiers,” and they induce a rule set or decision tree that models the data. Weka also includes algorithms for learning association rules and clustering data. All implementations have a uniform command-line interface. A common evaluation module measures the relative performance of several learning algorithms over a given data set (12).

6. Scikit-Learn

Scikit-learn (formerly scikits.learn) is an open source machine learning library for the Python programming language (13). It features various classifier engines, regression, and clustering algorithms including support vector machines, logistic regression, naive Bayes, k-means, and DBSCAN; and is designed to interoperate with NumPy and SciPy (13). The scikit license is open source, commercially usable, and the software code can be downloaded on the official Web site (14).
Figure 5. Scikit-learn results performing binary classification using nonlinear Support Vector Classification (SVC) with Radial Basis Function (RBF) kernel. The target to predict is an Exclusive Or (XOR) of the inputs. The color map illustrates the decision function learned by the SVC (14).

Scikit-learn is a Python module integrating a wide range of machine learning algorithms for supervised and unsupervised problems. The software package focuses on bringing machine learning to nonspecialists using a general-purpose high-level language; it has minimal dependencies, and is distributed under the simplified Berkeley Software Distribution (BSD) license, for use in both academic and commercial settings (15).

7. Results and Discussion

We have reviewed several open source software tools for performing research and development on anomaly detection for network security. After reviewing the flexibility and popularity of usage, we believe that in order to proceed with our evaluation we should select two packages for additional in-house testing and evaluation. Table 1 describes the two anomaly detection tools that we feel offer the most flexibility and the most number of anomaly detection algorithms for our in-house research purposes. From the table, we can see that the two packages share most of the basic algorithms that we can use in-house for performing basic research in anomaly detection. Therefore, we have submitted a formal request within our branch to install SHOGUN and Scikit-learn within the U.S. Army Research Laboratory’s (ARL) computers and we are awaiting feedback.
Table 1. Side-by-side comparison of algorithms offered by SHOGUN and Scikit-learn.

<table>
<thead>
<tr>
<th>Shogun</th>
<th>Scikit-Learn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support vector machines</td>
<td>Supervised learning:</td>
</tr>
<tr>
<td>Dimensionality reduction algorithms:</td>
<td>Generalized Linear Models</td>
</tr>
<tr>
<td>PCA, Kernel PCA, Locally Linear</td>
<td>Support Vector Machines</td>
</tr>
<tr>
<td>Embedding, Hessian Locally Linear Embedding, Local Tangent Space</td>
<td>Stochastic Gradient Descent</td>
</tr>
<tr>
<td>Alignment, Linear Embedding, Kernel Locally</td>
<td>Nearest Neighbors</td>
</tr>
<tr>
<td>Tangent Space Alignment, Kernel Local Tangent Space</td>
<td>Gaussian Processes</td>
</tr>
<tr>
<td>Alignment, Multidimensional Scaling, Isomap,</td>
<td>Partial Least Squares</td>
</tr>
<tr>
<td>Diffusion Maps, Laplacian Eigenmaps</td>
<td>Naive Bayes</td>
</tr>
<tr>
<td>Online learning algorithms:</td>
<td>Decision Trees</td>
</tr>
<tr>
<td>such as SGD-QN, Vowpal Wabbit</td>
<td>Ensemble methods</td>
</tr>
<tr>
<td>Clustering algorithms:</td>
<td>Multiclass and multilabel algorithms</td>
</tr>
<tr>
<td>k-means and GMM Kernel</td>
<td>Feature selection</td>
</tr>
<tr>
<td>Ridge Regression Support</td>
<td>Semi-Supervised</td>
</tr>
<tr>
<td>Vector Regression Hidden</td>
<td>Linear and Quadratic Discriminant Analysis</td>
</tr>
<tr>
<td>Markov Models</td>
<td></td>
</tr>
<tr>
<td>K-Nearest Neighbors</td>
<td></td>
</tr>
<tr>
<td>Linear discriminant analysis</td>
<td></td>
</tr>
<tr>
<td>Kernel Perceptrons</td>
<td></td>
</tr>
<tr>
<td>Kernels for numeric data:</td>
<td></td>
</tr>
<tr>
<td>linear gaussian</td>
<td></td>
</tr>
<tr>
<td>polynomial</td>
<td></td>
</tr>
<tr>
<td>sigmoid kernels</td>
<td></td>
</tr>
<tr>
<td>The supported kernels for special data include:</td>
<td></td>
</tr>
<tr>
<td>Spectrum</td>
<td></td>
</tr>
<tr>
<td>Weighted Degree</td>
<td></td>
</tr>
<tr>
<td>Weighted Degree with Shifts</td>
<td></td>
</tr>
</tbody>
</table>

8. Conclusions

In this report, we have performed a review of various software tools that can be leveraged in-house to perform basic research and development of anomaly-based intrusion detection algorithms. Out of the five software tools described, it is recommended to employ the Scikit or
SHOGUN (toolbox) as both tools are written in C++ and offer an interface for Python. The python language software is commonly employed as a research tool within our in-house team of researchers.
9. References


Appendix – List of Algorithm Per Tool

ELKI
Cluster analysis:
  K-means clustering
  Expectation-maximization algorithm
  Single-linkage clustering
  DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
  OPTICS (Ordering Points To Identify the Clustering Structure), including the extensions
  OPTICS-OF, DeLi-Cmu, HiSC, HiCO and DiSH
SUBCLU (Density-Connected Subspace Clustering for High-Dimensional Data)
Anomaly detection:
  LOF (Local outlier factor) OPTICS-OF
  DB- Outlier (Distance-Based Outliers) LOCI (Local Correlation Integral)
  LDOF (Local Distance-Based Outlier Factor) EM-Outlier
Spatial index structures:
  R-tree R*-tree M-tree
Evaluation:
  Receiver operating characteristic (ROC curve) Scatter plot
  Histogram
  Parallel coordinates
Other:
  Apriori algorithm Dynamic time warping Principal component analysis
RapidMiner
  Machine learning

This appendix appears in its original form, without editorial change.
Data mining Text mining Predictive analytics Business analytics.

SHOGUN

Support vector machines

Dimensionality reduction algorithms:

- PCA, Kernel PCA, Locally Linear Embedding, Hessian Locally Linear Embedding, Local Tangent Space Alignment, Linear Local Tangent Space Alignment, Kernel Locally Linear Embedding, Kernel Local Tangent Space Alignment, Multidimensional Scaling, Isomap, Diffusion Maps, Laplacian Eigenmaps

Online learning algorithms:

- such as SGD-QN, Vowpal Wabbit

Clustering algorithms:

- k-means and GMM Kernel Ridge Regression Support Vector Regression Hidden Markov Models
- K-Nearest Neighbors
- Linear discriminant analysis
- Kernel Perceptrons

Implemented kernels for numeric data include:

- linear gaussian polynomial sigmoid kernels

The supported kernels for special data include:

- Spectrum
- Weighted Degree
- Weighted Degree with Shifts

Weka

Data mining:

- Data preprocessing
- Clustering
- Expectation maximization
- Classification Regression Visualization
Feature selection

Scikit

Supervised learning: Generalized Linear Models Support Vector Machines Stochastic Gradient Descent Nearest Neighbors
  Gaussian Processes Partial Least Squares Naive Bayes
  Decision Trees
  Ensemble methods
  Multiclass and multilabel algorithms
  Feature selection

Semi-Supervised
  Linear and Quadratic Discriminant Analysis

Unsupervised learning: Gaussian mixture models Manifold learning Clustering
  Decomposing signals in components (matrix factorization problems) Covariance estimation
  Novelty and Outlier Detection
  Hidden Markov Models
INTENTIONALLY LEFT BLANK.
### List of Symbols, Abbreviations, and Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGPL</td>
<td>Affer General Public License</td>
</tr>
<tr>
<td>ARL</td>
<td>U.S. Army Research Laboratory</td>
</tr>
<tr>
<td>BSD</td>
<td>Berkeley Software Distribution</td>
</tr>
<tr>
<td>DMS</td>
<td>Data mining system</td>
</tr>
<tr>
<td>ELKI</td>
<td>Environment for Developing KDD-Applications Supported by Index-Structures</td>
</tr>
<tr>
<td>ETL</td>
<td>Extract, transform, load</td>
</tr>
<tr>
<td>GIS</td>
<td>Geographic/geospatial information system</td>
</tr>
<tr>
<td>GUI</td>
<td>Graphical user interface</td>
</tr>
<tr>
<td>KDD</td>
<td>Knowledge discovery in databases</td>
</tr>
<tr>
<td>RBF</td>
<td>Radial Basis Function</td>
</tr>
<tr>
<td>SVC</td>
<td>Support Vector Classification</td>
</tr>
<tr>
<td>Weka</td>
<td>Waikato Environment for Knowledge Analysis</td>
</tr>
<tr>
<td>XML</td>
<td>eXtensible Markup Language</td>
</tr>
<tr>
<td>XOR</td>
<td>Exclusive Or</td>
</tr>
<tr>
<td>YALE</td>
<td>Yet Another Learning Environment</td>
</tr>
<tr>
<td>No. of Copies</td>
<td>Organization</td>
</tr>
<tr>
<td>---------------</td>
<td>---------------------------------------------------</td>
</tr>
<tr>
<td>1 (PDF)</td>
<td>DEFENSE TECHNICAL INFORMATION CTR DTIC OCA</td>
</tr>
<tr>
<td>2 (PDF)</td>
<td>DIRECTOR US ARMY RSRCH LAB RDRL CIO LL RDRL IMAL HRA RECORDS MGMT</td>
</tr>
<tr>
<td>1 (PDF)</td>
<td>GOVT PRINTG OFC A MALHOTRA</td>
</tr>
<tr>
<td>2 (PDF)</td>
<td>DIR USARL RDRL CIN D R ERBACHER R PINO</td>
</tr>
</tbody>
</table>