Machine Learning Techniques for Intrusion Detection on Public Dataset

Udaya Sampath K. Perera Miriya Thanthrige, Jagath Samarabandu, Xianbin Wang
Electrical and Computer Engineering, University Of Western Ontario, London, Ontario, Canada
Email:{mperer4, jagath, xianbin.wang}@uwo.ca

***Abstract*—The development of computer based systems expands the usage of computer based application in human life. It can be observed that illegal activities such as unauthorized data access, data theft, data modification and various other intrusion activities are rapidly growing during last decade. Hence, deployment and continuous improvement of Intrusion Detection Systems (IDS) are of paramount importance. Training, testing and evaluation of IDS with real network traffic is significant challenge, so most of IDS evaluation is based on intrusion datasets. Therefore, analysis of intrusion datasets are of paramount importance. In this paper, we evaluated Aegean Wi-Fi Intrusion Dataset (AWID) with different machine learning techniques. Feature reduction techniques such as Information Gain (IG) and Chi-Squared statistics (CH) were applied to evaluate dataset performance with feature reduction. Results of experiments show that feature reduction can lead to better analysis in terms of accuracy, processing time and complexity. It was observed that, the maximum increment of classification accuracy with feature reduction from 110 to 41 is 2.4%.**

I. Introduction

The interaction between human and computer networks are rapidly growing day by day. This increase in interaction with computer based systems result in exchanging valuable and sensitive information via computer networks. Hence, deployment and continuous improvement of network security systems are of paramount importance. Within the broader area of network security research, there are many research activities that aim to improve intrusion detection techniques. Although real time intrusion detection is an important feature of an Intrusion Detection System (IDS), most IDSs operate in offline mode due to the need to analyze a huge amount of network activity data. For research, offline mode provides opportunity for in depth analysis of patterns and behaviors of intrusions. In addition to that, it provides opportunity for testing of intrusion detection algorithms. In operational point of view offline mode provides in depth analysis of past data and generate prevention methods for future occurrences. As monitoring a large number of features can increase computational time of IDS, it is important to select the best features that efficiently contribute to detection process.

Feature selection methods can be mainly classified as filter-based methods and wrapper-based methods [1]. Although filter-based methods have better efficiency when compared with wrapper-based methods, wrapper-based methods are shown to have better accuracy than filter-based methods [2]. There are a number of studies that have used older datasets such as KDDCUP 99 [3], NSLKDD [4] and many researchers indicate that these datasets are outdated now [5], [6]. Hence, it is important to evaluate new datasets that can replace these old datasets.

In this paper we evaluate the publicly available Aegean Wi-Fi Intrusion Dataset (AWID)[7] with respect to different machine learning techniques with feature selection approach. The dataset contains real traces of both normal and intrusion activities of 802.11 Wi-Fi network. Each record of the dataset is classified as either normal or a specific intrusion type (ie., class attribute of a record is refers to a type of intrusion or normal network activity). The AWID datasets can be mainly classified into two types based on class labeling, the high-level labelled dataset contain 4 major classes while other dataset has a more finer grained class labelling [7]. OneR, Ada Boost, J48 Decision tree, Random Forest and Random Tree machine learning techniques are used to evaluate the AWID dataset with information gain and chi squared statistics based feature selection.

Kolias et al. [8] were first to analyze the AWID dataset with machine learning techniques. They analyzed high-level labelled AWID reduced dataset with different machine learning techniques. Our contribution differs from their work that, we perform feature selection using both information gain and chi squared statistics based feature selection methods and we evaluated both high-level and finer grained labelled datasets with different machine learning techniques. Also we have performed experiments with different tree sizes to observe the behavior of random forest algorithm.

It was observed that, processing times were significantly decreased (in the range of 15.29% to 51.85% and 16.54% to 71.37%) with a maximum accuracy increment of 2.4% and 1.8% for high-level labelled data set and finer grained labelled dataset respectively with respect to feature reduction from 111 to 41. For feature reduction from 111 to 10, processing times were significantly decreased (in the range of 27.46% to 79.70% and 43.87% to 83.88%) with decrement of accuracy of the classification (in the range of 0 to 2.5% and 0 to 3.8%) for high-level labelled dataset and finer grained labelled dataset respectively. Because of the lack of studies in 802.11 based dataset such as AWID dataset, we believe that it is beneficial to community of researchers to evaluate different dataset with different intrusion detection methods.

The rest of this paper is structured as follows. Section 2 describes the classification of intrusion detection techniques. Section 3 describes dataset classification and Aegean Wi-

978-1-4673-8721-7/16/$31.00 ©2016 IEEE

Fi Intrusion Dataset. Section 4 describes methodology and experimental results and findings.

II. Classification of Intrusion Detection Methods

Typical network security intrusion detection methods can be classified as misuse-based detection, anomaly-based detection (i.e. behavior-based detection) or combination of both [9], [10]. The misuse-based detection mechanism compares network packet flow with known malicious threat patterns [11], [12]. The advantages of misuse-based intrusion detection are higher accuracy and easy implementation. But it has several disadvantages such as inability to detect unknown intrusions and the need to perform database updates regularly.

Anomaly-based detection is based on behavior of the system/user. Behavior of the system is classified as normal operation or abnormal operation mode based on measurements of system parameters and characteristics [13], [14]. The main advantage of behavior-based detection is that it is capable of identifying unknown attacks. The main disadvantage is lower accuracy. Main approaches of intrusion detection can be categorized as below.

1) Knowledge based systems
   - Pattern matching
   - State transition based methods
   - Expert systems
2) Statistical methods
3) Protocol analysis based methods
4) Soft computing and Classification based methods
5) Hybrid methods (Combination of two or more intrusion detection methods or techniques)

III. Datasets for training, testing and evaluation of Intrusion Detection

Dataset can be mainly categorized into three main types as

1) Real network dataset :- Real network datasets include data captured from real networks over few days. The captured data include normal and abnormal behaviors of the network.
2) Benchmark dataset :- Benchmark dataset was generated by simulated environments in large network. Simulated environment generates different intrusion scenarios.
3) Synthetic dataset :- Synthetic dataset is generated to meet a specific requirements. Synthetic dataset is usually used to evaluate the system prototype theoretically.

A. Aegean Wi-Fi Intrusion Dataset (AWID)

The Aegean Wi-Fi Intrusion Dataset (AWID) is a labeled dataset which was developed based on real traces of both normal and intrusion activities of a 802.11 Wi-Fi network [7]. The dataset consists of a large dataset and a reduced dataset. It include separate dataset for training (denoted as Trn) and testing (denoted as Tst). Each record of the dataset is classified as either normal or a specific intrusion type (ie., class attribute of a record is refers to a type of intrusion or normal network activity). The AWID datasets can be mainly classified into two types as high-level labelled dataset (AWID-CLS) and finer grained labelled dataset (AWID-ATK) based on class distribution. The higher level class distribution dataset (AWID-CLS-F and AWID-CLS-R) contain four major classes namely Flooding, Impersonation, Injection and Normal. Other dataset (AWID-ATK-F and AWID-ATK-R) include more detailed class labeling. The training datasets (AWID-ATK-F-Trn and AWID-ATK-R-Trn) contain 10 classes namely Amok, Arp, Authentication request, Beacon, Cafe latte, Deauthentication, Evil twin, Fragmentation, Probe response and Normal. The large test dataset (AWID-ATK-F-Tst) contain 17 classes with 7 additional classes than training dataset (AWID-ATK-F-Trn) namely Chop chop, Cts, Disassociation, Hirte, Power, Probe response and Rts. The reduced test dataset (AWID-ATK-R-Tst) contain 15 classes namely Amok, Arp, Beacon, Cafe latte, Chop chop, Cts, Deauthentication, Disassociation, Evil twin, Rts, Fragmentation, Hirte, Power saving, Probe request and Normal [7].

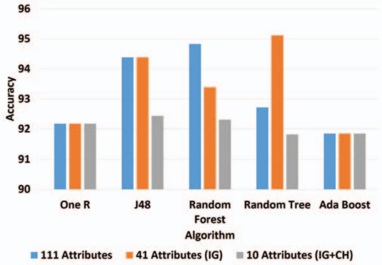

Fig. 1. Correctly Classified % of high-level class distribution dataset.

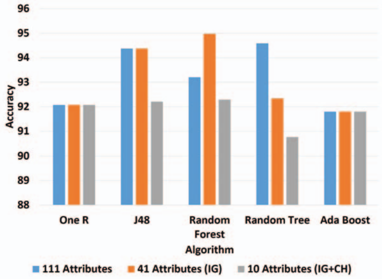

Fig. 2. Correctly Classified % of Finer grained class distribution dataset.

The large datasets include 162,375,247 records in training datasets (AWID-ATK-F-Trn and AWID-CLS-F-Trn) and 48,524,866 record in testing datasets (AWID-ATK-F-Tst and AWID-CLS-F-Tst), which has 97% of normal class records. Reduced datasets include 1,795,575 records in the training (AWID-ATK-R-Trn and AWID-CLS-R-Trn) datasets. Which include 90% of normal class records. Reduced testing datasets include 575,643 record (AWID-ATK-R-Tst and AWID-CLS-R-Tst) with 92% of normal class records [7].

IV. Methodology and Experimental Results

AWID reduced datasets were chosen for machine learning evaluation with feature selection approach. Intel Core i5 3.30GHz and 8GB RAM System running Ubuntu 14.04 LTS 64bit and Waikato Environment for Knowledge Analysis (Weka) [15] were used for this experiments. The chosen

TABLE I
MACHINE LEARNING EVALUATION OF AWID WITH 111 ATTRIBUTES

AWID-CLS-R-Trn and AWID-CLS-R-Tst (High-level class distribution)					
	OneR	J48	Random Forest	Random Tree	Ada Boost
Correctly Classified%	92.17	94.39	94.83	92.72	91.85
Incorrectly Classified%	7.83	5.61	5.17	7.28	8.15
TP Rate	0.922	0.944	0.948	0.927	0.918
FP Rate	0.898	0.141	0.445	0.616	0.255
Precision	0.861	0.969	0.942	0.876	0.909
F-Measure	0.888	0.938	0.933	0.901	0.91
ROC Area	0.512	0.884	0.974	0.954	0.946
Time	25.3	222.9	734.0	116.6	486.9
AWID-ATK-R-Trn and AWID-ATK-R-Tst (Finer grained class distribution)					
Correctly Classified%	92.07	94.37	93.21	94.58	91.80
Incorrectly Classified%	7.93	5.63	6.79	5.42	8.20
TP Rate	0.921	0.944	0.932	0.946	0.918
FP Rate	0.898	0.117	0.735	0.418	0.254
Precision	0.853	0.944	0.894	0.92	0.907
F-Measure	0.885	0.944	0.906	0.93	0.908
ROC Area	0.512	0.915	0.962	0.964	0.932
Time	25.2	198.2	849.4	93.3	278.3

TABLE II
MACHINE LEARNING EVALUATION OF AWID WITH 41 ATTRIBUTES

AWID-CLS-R-Trn and AWID-CLS-R-Tst (High-level class distribution) Attribute Selection Method :- Information Gain Attribute Evaluation					
	OneR	J48	Random Forest	Random Tree	Ada Boost
Correctly Classified%	92.17	94.39	93.39	95.12	91.85
Incorrectly Classified%	7.83	5.61	6.61	4.88	8.15
TP Rate	0.922	0.944	0.934	0.951	0.918
FP Rate	0.898	0.141	0.646	0.538	0.255
Precision	0.861	0.969	0.928	0.91	0.909
F-Measure	0.888	0.938	0.912	0.93	0.91
ROC Area	0.512	0.884	0.957	0.704	0.946
Time	14.0	188.8	353.4	58.3	305.3
AWID-ATK-R-Trn and AWID-ATK-R-Tst (Finer grained class distribution) Attribute Selection Method :- Information Gain Attribute Evaluation					
Correctly Classified%	92.07	94.37	94.97	92.34	91.80
Incorrectly Classified%	7.93	5.63	5.03	7.66	8.20
TP Rate	0.921	0.944	0.95	0.923	0.918
FP Rate	0.898	0.117	0.573	0.826	0.254
Precision	0.853	0.944	0.907	0.864	0.907
F-Measure	0.885	0.944	0.927	0.891	0.908
ROC Area	0.512	0.915	0.967	0.694	0.932
Time	15.4	160.9	280.3	26.7	232.3

TABLE III
MACHINE LEARNING EVALUATION OF AWID WITH 10 ATTRIBUTES

AWID-CLS-R-Trn and AWID-CLS-R-Tst (High-level class distribution) Information Gain Attribute Evaluation and Chi-Square Attribute Evaluation					
	OneR	J48	Random Forest	Random Tree	Ada Boost
Correctly Classified%	92.17	92.44	92.31	91.82	91.85
Incorrectly Classified%	7.83	7.56	7.69	8.18	8.15
TP Rate	0.922	0.924	0.923	0.918	0.918
FP Rate	0.898	0.888	0.893	0.791	0.255
Precision	0.861	0.909	0.895	0.893	0.909
F-Measure	0.888	0.89	0.889	0.896	0.91
ROC Area	0.512	0.84	0.936	0.578	0.913
Time	5.14	78.71	151.86	19.03	48.23
AWID-ATK-R-Trn and AWID-ATK-R-Tst (Finer grained class distribution) Information Gain Attribute Evaluation and Chi-Square Attribute Evaluation					
Correctly Classified%	92.07	92.21	92.29	90.76	91.80
Incorrectly Classified%	7.93	7.79	7.71	9.24	8.20
TP Rate	0.921	0.922	0.923	0.908	0.918
FP Rate	0.898	0.922	0.899	0.915	0.254
Precision	0.853	0.85	0.872	0.85	0.907
F-Measure	0.885	0.885	0.887	0.878	0.908
ROC Area	0.512	0.814	0.926	0.535	0.932
Time	4.7	111.25	182.71	24.67	44.85

datasets were AWID-CLS-R-Trn, AWID-CLS-R-Tst, AWID-ATK-R-Trn and AWID-ATK-R-Tst. AdaBoost, J48, OneR, Random Forest and Random Tree machine learning techniques were used to evaluate the chosen dataset. In order to reduce preprocessing complexity, string attributes were removed from the dataset and 111 attributes were selected for the experiment. In the first step machine learning evaluation was done without feature selection. In the second step information gain was applied for feature selection, based on that, 40 attributes were selected. In the third step, chi squared statistics based feature selection method was applied on features which were selected in the second step. Based on chi squared statistics, 9 features were selected. The performance of machine learning techniques with 111, 41 and 10 attributes were listed in Table I, II and III respectively. These results indicate that all 5 algorithms achieved over 90 % classification accuracy. Random Tree with 41 attributes achieved the highest accuracy of 95.12% for high-level labelled dataset and Random Forest with 41 attributes achieved the highest accuracy of 94.97% for finer grained labelled dataset.

It was observed that, accuracy of the classification was increased by maximum of 2.4% (random tree algorithm) and 1.8% (random forest algorithm) for high-level labelled dataset and finer grained labelled dataset respectively with respect to feature reduction from 111 to 41. The only exception to this was random forest algorithm in high-level labelled dataset and random tree algorithm in finer grained labelled dataset, in which accuracy of the classification was decreased by 1.44% and 2.24% respectively. When number of features was reduced to 10, accuracy was decreased by maximum of 2.5% for random forest algorithm and maximum of 3.8% for random tree algorithm for high-level labelled dataset and finer grained labelled dataset respectively. For feature reduction from 111 to 40 and 111 to 10, processing times were significantly decreased by maximum of 51.85% (random forest algorithm) and 79.70% (oner algorithm) for high-level labelled dataset. For finer grained labelled dataset it was decreased by maximum of 71.37% (random tree algorithm) and 83.88% (ada boost algorithm). Accuracy of the classification and processing times were shown in figure 1, 2, 3 and 4.

The AWID dataset can be mainly divided in to two types based on it's class distribution (high-level labelling and finer grained labelling). It can be observed that with increase of classes, accuracy of the classification is slightly reduced (maximum difference of 2.78% for random tree algorithm). The area under ROC curve (AUC) is a better measurement to compare performance of classifiers (Higher AUC is better). It can be observed that the ROC area is very low in OneR compare to all other methods in every cases. Random Forest and Ada boost achieved over 0.9 values for AUC in all experiments.

TABLE IV
Top 10 Attribute Selection

Information Gain Attribute Evaluation	Chi-Square Attribute Evaluation
1.frame.len	1.frame.cap.len
2.frame.cap.len	2.frame.len
3.radiotap.mactime 4.frame.time.relative 5.frame.time.epoch 6.wlan.seq 7.frame.time.delta 8.frame.time.delta.displayed 9.data.len	
10.wlan.duration	10.wlan.frag

TABLE V
Performance Evaluation of Random Forest Algorithm with Number of Trees for AWID dataset with 111 Attributes

Number of Trees	high-level class distribution (AWID-CLS-R)	Finer grained class distribution (AWID-ATK-R)
	Correctly Classified %	Correctly Classified %
5	93.8212	94.6698
10	93.3902	94.9739
15	93.3294	93.7001
20	93.64	93.1951
25	94.1179	93.7857

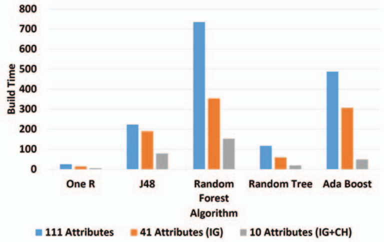

Fig. 3. Build Time of AWID Dataset with high-level class distribution.

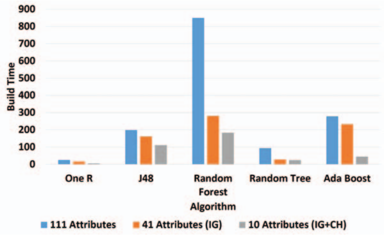

Fig. 4. Build Time of AWID Dataset with finer grained class distribution.

J48 achieved lowest FP rate for both 111 and 41 attributes while Adaboost achieved for 10 attributes. It can be seen that Information Gain attribute evaluation (IG) and Chi-Square attribute evaluation (CH) produce nine common attribute when they applied separately to the original dataset, these attributes are presented in table IV. Table V represents the performance evaluation of random forest algorithm with respect to number of trees. Results indicate that slight accuracy improvement can be achieved with respect to increase of the number of trees for the high-level labelled dataset. But for the finer grained labelled dataset, a slight decrease of accuracy is observed.

V. Conclusion

Intrusion datasets are used to train, test and evaluate intrusion detection systems. It is challenging to identify relevant features that has a significant effect on the accuracy of intrusion detection. With better feature identification, it is possible to develop an efficient IDS. The experimental results indicate that feature reduction can improve the detection accuracy and classification speed, but if we further reduce the number of features, it can decrease the detection accuracy. Frequent evaluation of datasets are important to identify weakness of datasets which helps in developing new datasets with better utility. Network intrusions are growing rapidly and new attack vectors are continuously being developed which makes it is very challenging to generate a dataset that includes sufficient amount of relevant intrusion types. Hence continuous evaluation and development of datasets are very important.

VI. Acknowledgment

The support of the NSERC (Natural Sciences and Engineering Research Council of Canada) and University of Western Ontario is gratefully acknowledged by the authors.

References

[1] R. Kohavi and G. H. John, “Wrappers for feature subset selection,” *Artificial intelligence*, vol. 97, no. 1, pp. 273–324, 1997.

[2] H. Min and W. Fangfang, “Filter-wrapper hybrid method on feature selection,” in *Intelligent Systems (GCIS), 2010 Second WRI Global Congress on*, vol. 3, Dec 2010, pp. 98–101.

[3] KDDCUP99, “Kdd cup99 data set,” 1999. [Online]. Available: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[4] NSL-KDD, “Nsl-kdd data set for network-based intrusion detection systems.” 2009. [Online]. Available: http://www.unb.ca/research/iscx/dataset/iscx-NSL-KDD-dataset.html

[5] M. Tavallaee, E. Bagheri, W. Lu, and A. Ghorbani, “A detailed analysis of the kdd cup 99 data set,” in *Computational Intelligence for Security and Defense Applications, 2009. CISDA 2009. IEEE Symposium on*, July 2009, pp. 1–6.

[6] M. Sabhnani and G. Serpen, “Why machine learning algorithms fail in misuse detection on kdd intrusion detection data set,” *Intelligent Data Analysis*, vol. 8, no. 4, pp. 403–415, 2004.

[7] AWID, “Awid-wireless security datasets project data set,” 2014. [Online]. Available: http://icsdweb.aegean.gr/awid/features.html

[8] C. Kolias, G. Kambourakis, A. Stavrou, and S. Gritzalis, “Intrusion detection in 802.11 networks: Empirical evaluation of threats and a public dataset,” *Communications Surveys Tutorials, IEEE*, vol. PP, no. 99, pp. 1–1, 2015.

[9] J. Cannady and J. Harrell, “A comparative analysis of current intrusion detection technologies,” in *Proceedings of the Fourth Technology for Information Security Conference*, vol. 96. Citeseer, 1996.

[10] S. Axelsson, “Intrusion detection systems: A survey and taxonomy,” Technical report Chalmers University of Technology, Goteborg, Sweden, Tech. Rep., 2000.

[11] M. Salour and X. Su, “Dynamic two-layer signature-based ids with unequal databases,” in *Information Technology, 2007. ITNG ’07. Fourth International Conference on*, April 2007, pp. 77–82.

[12] P. Gupta, C. Raissi, G. Dray, P. Poncelet, and J. Brissaud, “Ss-ids: Statistical signature based ids,” in *Internet and Web Applications and Services, 2009. ICIW ’09. Fourth International Conference on*, May 2009, pp. 407–412.

[13] O. Linda, T. Vollmer, and M. Manic, “Neural network based intrusion detection system for critical infrastructures,” in *Neural Networks, 2009. IJCNN 2009. International Joint Conference on*, June 2009, pp. 1827–1834.

[14] N. Aissa and M. Guerroumi, “A genetic clustering technique for anomaly-based intrusion detection systems,” in *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2015 16th IEEE/ACIS International Conference on*, June 2015, pp. 1–6.

[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA data mining software: An update,” *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009. [Online]. Available: http://www.sigkdd.org/explorations/issues/11-1-2009-07/p2V11n1.pdf