The Effect of Feature Selection on Detection Accuracy of Machine Learning Algorithms

Noureldien A. Noureldien
University of Science and Technology
Department of Computer Science
Omdurrman, Sudan

Raghda A. Hussain
University of Science and Technology
Department of Computer Science
Omdurman, Sudan

Ahmed Khalid
Najran University
Community College
KSA

Abstract

Machine learning algorithms are commonly used to detect anomalies in network traffic. Recently, many research studies are focus on the detection performance of classification algorithms.

Determining the optimistic performance of an algorithm is dependent on various factors and determining the optimistic detection performance for a given algorithm is a challenging research problem.

In this paper an experiment was conducted to see the effect of feature selection on the detection performance of machine learning algorithms. The algorithms Trees.J48, Bayes.BayesNet, Functions.Logistic, Meta.Bagging and Rules.ZeroR are used to test their detection performance of DoS attacks in KDDCup99 data set using different sets of features.

The experimental results show that an algorithm detection performance is dependent on the selected features and the general detection behavior is independent of the number of selected features.

1. Introduction

Network intrusion detection aims to protect networks and systems from malicious attacks. Intrusion detection techniques can be divided into two complementary approaches: misuse detection, and anomaly detection. Misuse detection systems stores patterns of known attacks and scan the system data for occurrences of these patterns, on the other hand anomaly detection systems works by monitoring significant deviations from a normal or expected behavior of the system or users.

The anomaly based detection system first learns normal system or user activities and then alerts the system or user behaviors that deviate from the already learned activities. The main negative aspect of anomaly based detection systems is that it erroneously classifies the normal system or user behaviors as attacks, which would result in false positive alarms.

In anomaly detection systems classifiers or machine learning algorithms are used to differentiate normal behavior from malicious one. Typically machine learning algorithms are trained to learn normal behavior so that they can detect abnormal or malicious behavior in new data. The learning process is either supervised or unsupervised. In supervised learning, the class labels of training data are already known. The task of a supervised learner is to find a function to approximate the mapping between training data and their classes so that it can predict the classes of new data.

There are many algorithms proposed for supervised learning, such as artificial neural networks [1], naïve Bayes classifiers [2], decision trees [3], K-nearest neighbor [4], support vector machines (SVMs) [5] and random forests [6].

In order to improve the learning process, before the algorithm starts training and learning, the training data set go through many operations, known as data preprocessing. One of the major techniques that are used frequently in data preprocessing is feature selection.

Feature selection is about how to select informative features from the data set features to remove irrelevant, redundant or noisy ones from data. By reducing the dimensionality of data, feature selection reduces the overall computational cost, improves the performance of learning algorithms and enhances the comprehensibility of the data models.

With the help of feature selection, machine learning algorithms become more scalable, reliable and accurate. Many feature selection algorithms have been proposed in the literature [7, 8, 9, 10, 11, 12, 13]. These algorithms are categorized into two groups, wrapper employs learning algorithms and the filter algorithms.

From this enormous and increasing number of classification and feature selection algorithms, it becomes important to answer questions such as "Which classification algorithm have a high detection performance for a given attack type?", "What is the optimistic feature set for a given classification algorithm that achieves best performance?", "How features selection affect the detection performance of an algorithm?", "Under what criteria can we compare machine learning algorithms performance?"

In this paper, experiments are conducted to test the effect of feature selection on the detection performance of machine learning algorithms, and to observe the performance behavior when attributes are changed.

The remainder of this paper is organized as follows: Section 2 is dedicated for related work, in section 3 we describe our empirical method and the used tools. The experiments and results are discussed in section 4 and finally the conclusion and future work are drawn in section 5.

2. Related Work

There are various research studies that compare the efficiency of machine learning algorithms**.**

N. S. Chandolikar and V. D. Nandavadekar [14] evaluate the detection performance of two well known classification algorithms, Bayesnet and J48 algorithms on KDDcup99 dataset. To test and evaluate the algorithms they use 10-fold cross validation, in which the data set is divided into 10 subsets. Each time, one of the 10 subsets is used as the test set and the other 9 subsets form the training set. Performance statistics are calculated across all 10 trials. They evaluate the algorithms on the bases of true positive (TP) and false positive (FP) rates.

L. Portnoy, E. Eskin, and S. Stolfo [15] partition the KDDcup99 data set into ten subsets, each contain approximately 490,000 instances or 10% of the data. However, they observe that the distribution of the attacks in the KDD data set is very uneven which made cross-validation very difficult. They conclude that many of these subsets contain instances of only a single type. For example, the 4th, 5th, 6th, and 7th, 10% portions of the full data set contained only smurf attacks, and the data instances in the 8th subset were almost entirely neptune intrusions [16].

G. Kalyani and A. Jaya Lakshmi [17] compares the performance and accuracy of the algorithms; Naive Bayes, j48, OneR, PART and RBF Network Algorithm. The experiments and assessments were performed using WEKA with NSL-KDD dataset. 75% data is used for training and the remaining is for testing purposes, they conclude from the simulation results that, the best algorithm based on the intrusion detection data is PART classifier.

Adetunmbi A. Olusola. et.al [18] introduces an analytical study to find the relevance of each feature in KDD ’99 intrusion detection dataset to the detection of each attack class. Their empirical results show that seven features were not relevant in the detection of any class.

Saeed Abu-Nimeh, et.al [19] presents a study that compares the predictive accuracy of several machine learning methods for predicting phishing emails. They use a data set of 2889 phishing and legitimate emails in the comparative study. In addition, 43 features are used to train and test the classifiers.

Upendra and Yogendra Kumar [22] are aim to find out which classifier is better among five machine learning algorithm, namely, J48, BayesNet, OneR, NB and ZeroR. They use many performance criteria including; accuracy, precision, recall, F-Measure, incorrectly classified instances, kappa statistic, and mean absolute error. They carry out their experiments on KDDCup99 and they use two different sets of features, 41 attributes and 7 attributes respectively.

3. Material and Methods

Since the objective is to see the impact of feature selection on the performance of machine learning algorithms, a number of algorithms; Trees.J48, Bayes.BayesNet, Functions.Logistic, Meta.Bagging and Rules.ZeroR which are belongs to different machine learning categories are experimented to measure their detection performance using different sets of attributes.

Experiments are carried using KDDCup99data set which is a subset of the DARPA benchmark data set [20]. Each KDDCup’99 training connection record contains 41 features and is labeled as either normal or an attack.

KDD dataset covers four major categories of attacks: Probing attacks (information gathering attacks), denialof-Service (DoS) attacks (deny legitimate requests to a system), user-to-root (U2R) attacks (unauthorized access to local super-user or root), and remote-to-local (R2L) attacks (unauthorized local access from a remote machine).

The detection performance of each of the selected algorithms will be checked against four sets of selected attributes, namely the full dataset features (41 attributes), the DoS attack relevant features (21 attributes), the Neptune attack relevant features (14 attributes) and the Smurf attack relevant features (12 attributes). Table (1) shows the attributes numbers of the selected features subsets [21].

Table (1): DoS, Neptune and Smurf Features

DoS Relevant Features		Neptune Relevant Features	Smurf Relevant Features
2	32	3	2
3	34	4	3
4	36	5	5
5	37	23	6
6	38	26	12
7	39	29	25
8		30	29
12		31	30
23		32	32
25		34	36
26		36	37
29		37	39
30		38	
31		39	

The WEKA tool version 3.6.2 is used to analyses the detection accuracy of the selected algorithm, and accuracy is used as the performance measure. Accuracy is one of the most basic measures of performance of machine learning algorithms; it determines the percentage of correctly classified instances, i.e. the ratio of true positives and true negatives to the total number of instances.

Therefore the WEKA's parameter Correctly Classified Instances and Incorrectly Classified Instances, which determine respectively the percentage of instances in the data set that are correctly and incorrectly classified by the algorithm, are used as a measure for performance detection.

4. Experiments Results and Discussion

The experiments are done using the standard KDD Cup99 data set and WEKA version 3.6.2, the simulationplatform is an Intel® Core (TM) i5-2430 processor system running at 2.40 GHz with installed memory (RAM) 3.00 GB under Microsoft Windows7 Professional operating system.Tables (2-6) show the experimental results for each algorithm using 66% split of the data set.

Table (2): J48

Performance	Correctly Classified Instances	Incorrectly Classified Instances
Features		
Full Features (41 Attributes)	99.9412%	0.0588%
DoS Features (21 Attributes)	99.9412%	0.0588%
DoS Features (21 Attributes)	99.8824%	0.1167%
Smurf Features (12 Attributes)	99.9412 %	0.0588%

Table (3): BayesNet

Performance	Correctly Classified Instances	Incorrectly Classified Instances
Features		
Full Features (41 Attributes)	98.0588%	1.9412%
DoS Features (21 Attributes)	98.7794%	1.2206%
DoS Features (21 Attributes)	99.5735%	0.4265%
Smurf Features (12 Attributes)	98.4265%	1.5735%

Table (4): Logistic

Performance	Correctly Classified Instances	Incorrectly Classified Instances
Features		
Full Features (41 Attributes)	99.9118%	0.0882%
DoS Features (21 Attributes)	99.9265%	0.0735%
DoS Features (21 Attributes)	99.8824%	0.1176%
Smurf Features (12 Attributes)	99.8676 %	0.1324%

Table (5): Bagging

Performance	Correctly Classified Instances	Incorrectly Classified Instances
Features		
Full Features (41 Attributes)	99.9265%	0.0735%
DoS Features (21 Attributes)	99.9265%	0.0735%
DoS Features (21 Attributes)	99.9265%	0.0735%
Smurf Features (12 Attributes)	99.9412 %	0.0588%

Table (6): ZeroR

Performance	Correctly Classified Instances	Incorrectly Classified Instances
Features		
Full Features (41 Attributes)	80.1912%	19.8088%
DoS Features (21 Attributes)	80.1912%	19.8088%
DoS Features (21 Attributes)	80.1912%	19.8088%
Smurf Features (12 Attributes)	80.1912%	19.8088%

The results generally show that the detection accuracy of algorithms behaves differently with the change of the feature sets, which is chosen to be in a decreasing manner in order of the number of features. The J48 detection accuracy fluctuates with the change of features sets while other algorithms such as Bayesnet and Logistic their detection accuracy increase and then decrease steadily. The Bagging algorithm detection accuracy keeps constant and then increase, while ZeroR detection accuracy is not affected absolutely by the change in feature sets.

The results reveal the fact that the detection accuracy and performance of most algorithms is directly affected by the selected attribute set and this affection is not predictable. Therefore the comparing of performance of different classification algorithms should either be justified only to the selected attributes set, or to the optimistic performance of each algorithm.

5. Conclusion and Future Work

Feature selection is a major preprocessing stage for machine learning algorithms. Selection of good features will reduce data dimensionality and improve algorithm performance. In this paper we show through experiments that the detection performance of an algorithm is independent of the number of selected attributes, and therefore features, comparing machine learning algorithms can only be affirmative under the optimistic performance of each algorithm.

Our future work will focus on determining among available machine learning algorithms, the algorithm with optimum performance for each of the four attack types found in the KDDCup99 data set.

References

[1] Lawrence, J. "Introduction to Neural Networks", California Scientific Software Press. ISBN 1-883157-00-5, 1994.

[2] Langley, P., Iba, W. and Thompson, K. "An analysis of Bayesian classifier", In Proceeding of the Tenth National Conference on Artificial Intelligence, 223-228, 1992.

[3] Quinlan, J.R.," C4.5: Programs for Machine Learning". Morgan Kaufmann, 1993.

[4] Cover, T. M. and Hart, P. E.," Nearest neighbor pattern classification", IEEE, Transactions on Information Theory, 13:21-27, 1967.

[5] Burges, C. J. C.,"A tutorial on support vector machines for pattern recognition". Data mining and Knowledge Discovery, 2(2), 121-167, 1998.

[6] Breiman, L. Random Forest. Technical Report, Stat.Dept. UCB. 2001

[7] Liu, Y.," A comparative study on feature selection methods for drug discovery", Journal of Chemical Information and Computer Sciences 44(5): 1823-1828, 2004.

[8] Liu, H. Li, J. and Wong, L.,"A comparative study on feature selection and classification methods using gene expression profiles and proteomic pattern", Genomic Informatics, 13, 5160, 2002.

[9] Weston, J. et al. "Feature selection for SVMs". Advances in Neural Information Processing Systems 13, 2000.

[10] Liu, H. and Setiono, R. Chi,"Feature selection and discretization of numeric Attributes". Proc. of IEEE 7th International Conference on Tools with artificial Intelligence, 338-391, 1995

[11] Guyon, I. and Elisseeff, A. "An Introduction to Variable and Feature Selection (Kernel Machines Section)". JMLR, 3: 1157-1182, 2003

[12] Yang, J. and Honavar, V. "Feature subset selection using a genetic algorithm", ACM Computing Classification System Categories, 1997.

[13] Liu, H. et al. "Evolving Feature Selection", Journal of Intelligent Systems. IEEE Volume 20, Issue 6, Nov.-Dec. Page(s):64-76, 2005.

[14] N. S. Chandolikar, Dr. V. D. Nandavadekar, "Efficient Algorithm for Intrusion Attack Classification by Analyzing KDD Cup 99", 978-1-4673-1989-8/12 ©2012 IEEE.

[15] L. Portnoy, E. Eskin, and S. Stolfo, "Intrusion detection with unlabeled data using clustering," Proceedings of ACM CSS Workshop on Data Mining Applied to Security, Philadelphia, PA, November, 2001.

[16] Mahbod Tavallaee, et al., "A Detailed Analysis of the KDD CUP 99 Data Set", In proceedings of the 2009 IEEE Symposium on Computational intelligence in security and defense Applications CISDA 2009.

[17] G. Kalyani and A. Jaya Lakshmi, "Performance Assessment of Different Classification Techniques for Intrusion Detection", IOSR Journal of Computer Engineering (IOSRJCE)
ISSN: 2278-0661, ISBN: 2278-8727 Volume 7, Issue 5 (Nov-Dec. 2012), PP 25-29, 2012.

[18] Adetunmbi .A. Olusola. et.al; "Analysis of KDD '99 Intrusion Detection Dataset for
Selection of Relevance Features", Proceedings of the World Congress on Engineering and Computer Science 2010 Vol I, WCECS 2010, October 20-22, 2010, San Francisco, USA

[19] Saeed Abu-Nimeh, et.al; "A Comparison of Machine Learning Techniques for
Phishing Detection", APWG eCrime Researchers Summit, October 4-5, 2007, Pittsburgh,
PA, USA

[20] Tavallaee, M., E. Bagheri, W. Lu and A.A. Ghorbani. "A detailed analysis of the KDD CUP 99 data set". Proceedings of the 2nd IEEE Symposium on Computational Intelligence for Security and Defense Applications, Jul. 2009, NRC, Canada, pp: 1-7.

[21] H. GünesKayacık, A. NurZincir-Heywood, Malcolm I. Heywood, "Selecting Features for Intrusion Detection: A Feature Relevance Analysis on KDD 99 Intrusion Detection Datasets",

[22] Upendra, Yogendra Kumar, "An Empirical Comparison and Feature Reduction
Performance Analysis of Intrusion Detection", International Journal of Control Theory and Computer Modelling (IJCTCM) Vol.2, No.1, January 2012.