Comparing results of selecting different field combinations in IDS based on decision tree algorithm

Ahmad Yoosofan[1], Masoud Moghadasian[2], and Reza Babaee[3]
[1]Department of Computer Engineering, Faculty of Engineering, University of Kashan,
yoosofan@kashanu.ac.ir
[2]Bachelor of Computer Engineering, Faculty of Engineering, University of Kashan,
moghadasian.3519@gmail.com
[3]Reza Babaee, Department of Computer Engineering, Faculty of Engineering, University of Kashan,
rbabaeecar@yahoo.com

Abstract: *One major branch of research in network security is detecting intrusions with a low false positive or negative rate. Enormous approaches have been presented to enhance the performance of intrusion detection systems, each one reflecting various viewpoints. Among them, data mining based techniques have had desirable results; however their performance are mainly affected by feature selection, and so selecting different features and feature combinations remains as an open problem. In this paper we propose a method to select the features based on their relevance to the attack group characteristics. All attacks within KDD 99 data set, which is mostly used in the literature, are categorized into four groups. Most relevant features are recognized for each group, and the binary classification algorithm is performed by a combination of key features and the category-related feature. Our experimental results on KDD 99 data set show a better outperformance in the case of selecting the feature which is frequently applied in the attacks using inherently a few numbers of connections and therefore have proportionally less number of records in the data set.*

Keywords**:** *Intrusion Detection, Data Mining, Decision Tree, Feature Selection*

1. Introduction

It Data mining was emerged as a novel approach to extract profound knowledge from large mass of data. Data mining has been practiced in many fields such as bioinformatics, knowledge discovery and network security. The huge amount of data in network security that are collected from different analysis sources, offers data mining based techniques as a wise choice in the literature. One of the most general systems used in the network security context is Intrusion Detection System (IDS).

However due to the immense amount of data that should be examined in intrusion detection systems, reducing the volume of data while keeping them as accurate as possible, is a challenging problem. Thus, although data mining approaches seem to be a practical method in IDS but there should be some enhancements in order to make them remarkably efficient.

Attacks are classified into four categories: each has their own specified features [1]: Denial of Service Attacks (DoS), User to Root Attacks (U2R), Remote to Local Attacks (R2L) and Probing Attacks. Nevertheless this classification is not still sufficient for decreasing the data volume and so we require more specification to reduce the quantity of data. Feature selection is not only an efficient method applied in many researches with the aim of redundant data elimination, but it also improves the correctness of result by removing unnecessary negative effects. Considering the attack types which the IDS must find, these features can be grouped into the associated attack categories.

The most impressive feature for each attack that is determined in the KDD 99 data set is introduced by [2]. We use this valuable information in order to combine most relative feature for each attack category (e.g. DoS) then compare results for various combinations and suggest the best one. Our performed practical experiment on KDD 99 intrusion data set represents the possible result compared to the similar feature selection techniques.

The paper is organized as follows. Next section is dedicated to a brief explanation on Intrusion Detection Systems (IDS) and related concepts. In the section 3 we describe our experimental material, i.e. KDD 99. Our approach and research results are discussed in the section 4. Finally, the conclusion and the future works are presented in the section 5.

2. Intrusion Detection System: An Overview

In current open network environment, network security plays a vital role in effective network utilization. Among all state-of-the-art systems, IDS has already attracted the careful attention of network security experts. The aim of an IDS is to discriminate between normal behaviors and any anomalies. It should then take an

appropriate action such as sounding an alarm in the case of detecting an intrusion [3].

Kruegel et al. in [4] define an intrusion as follows:

"*An intrusion is defined as a sequence of related actions performed by a malicious adversary that results in the compromise of a target system.*"

According to this definition, IDS should inspect each (sequence of) action(s) taken in the target system and raise an alarm if a suspicious activity has been recognized. However sometimes IDS might be unable to distinguish between normal and abnormal activity, which is called a false positive if it is actually a normal activity and has been detected as an abnormal activity or a false negative if it is an abnormal activity and actually it has not been detected.

2.1. IDS Taxonomy

Decreasing the false ratio in IDS is one of the most challenging topics between network security experts. Depending on the approaches IDSs take, false positive and negative ratio may differ. Two major approaches may be selected by the IDS designer based on the detection method [4]: Anomaly Detection which defines the normal behavior rather than the abnormal behavior, and Misuse-based (Signature-based) Detection that instead explicitly describes what is abnormal using particular knowledge about the attacks which is called a signature. Moreover depending on the abstraction level of audit source location the IDS has to deal with, categorizes IDSs into three groups: Network level data, Host level data and Application level inspection. Consequently, each abstraction level has some advantages and disadvantages. For instance an IDS which is scrutinizing network packets can be vulnerable to encrypted attacks while IDSs which are working on the application level are more robust against them.

2.2. IDS Analysis Methods

Several techniques have been considered for processing collected data in the IDS to find anomalies, e.g. Expert Systems, Neuron-Fuzzy Techniques [5], Linear Support Vector Machines (SVM) [6], Artificial Neural Networks [7], etc. In most of real applications the detector algorithm should handle the mass amount of data which gradually decreases the effectiveness of the IDS and it might not be applicable in real-time environments. Therefore providing a partially agile approach leads us to using algorithms designed specifically for processing huge amount of data such as data mining.

In data mining based techniques the appropriate knowledge, i.e. distinguishing between anomalies and safe activities in our case, would be extracted from large volumes of audited raw data. The data mining based algorithms are usually inserted in the anomaly detection category [8]. There are a great number of algorithms suited for different data analysis. [9] has rated top ten data mining algorithms in which they are claimed to be the most influential algorithms for classification. One of these top algorithms is C4.5 [10]. [11] show that J48 has the best performance (93.82%) among other learning machine methods they provided (Fig. 1). J48 is in fact an implementation of C4.5 release 8 that is working using decision trees [12]. Decision trees kindly provide us with

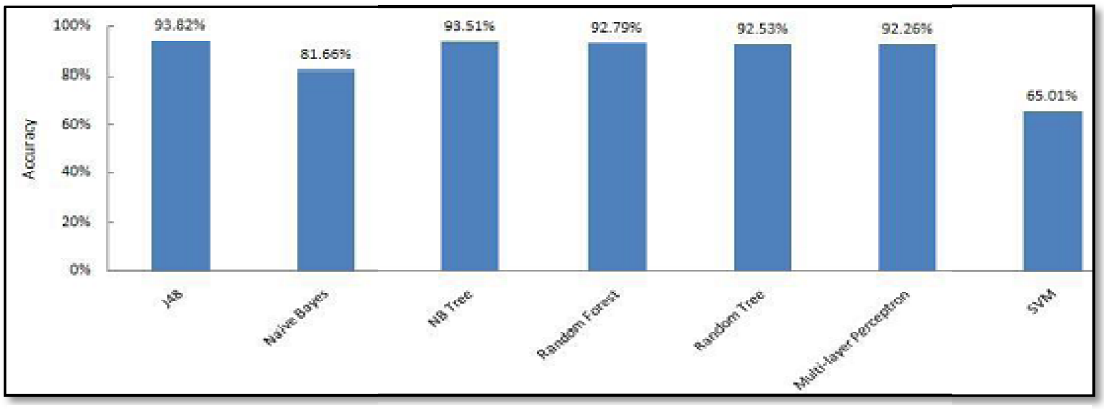

Fig. 1: The comparison of different machine learning algorithms' accuracy on KDDTest

a powerful approach to search and find anomalies in the very large data sets.

This bar chart is a useful clue to select the proper algorithm on KDDTest. Thus, we choose J48 as our data mining method which is applied on KDDTest data set. However [11] enhance the KDD 99 data set by removing redundant records which is claimed to cause statistical irregularities in KDD 99. Again J48 is almost the best algorithm among other methods within the new KDD data set.

3. Popular KDD 99

In 1998, MIT Lincoln Labs has produced a large date set named DARPA aimed at inspecting and evaluating the intrusion detection systems. The DARPA has been improved and extended by adding a label to each record indicating that whether it is a safe activity or an unsafe activity. The new data set was participated in the KDD Cup 99 and won the prize. Since then it is still the popular data set for a large number of researches and almost every new approach would be experimented with the KDD 99 [13], although extra degree of improvements has been proposed on it.

3.1. KDD Attack groups

Based on the containing records in KDD 99, every record is assigned as a normal activity or abnormal behavior. The abnormal activities are also categorized into four groups; each has its own specific characteristics so that the detection process sounds extremely convenient.

Denials of Service (DoS) attacks are those in which the intruder keeps the resources of the system quiet busy so that the system could not respond to any else requests. User to Root (U2R) attack is a traditional approach in which a local user tries to obtain an extra permission. Remote to Local (R2L) attack is occurred when a malicious agent attempts to intrude into a remote system. Probing attacks continually strive intentionally to collect data for finding any potential security holes.

Table 1: Attack Types and their category

Attack	Category
Smurf	DoS
Neptune	DoS
Back	DoS
Teardrop	DoS
Pod	DoS
Land	DoS
Perl	U2R
Loadmodule	U2R
Rootkit	U2R
Buffer_Overflow	U2R
Spy	R2L
Phf	R2L
Multihop	R2L
Imap	R2L
Guess_Passwd	R2L
Nmap	Probe
Portsweep	Probe
Ftp_Write	R2L
Warezmaster	R2L
Warezclient	R2L
Satan	Probe
Lpsweep	Probe
Normal	Normal

3.2. KDD Data Features

Each record in KDD 99, consisting of 100 bytes, is labeled with the attack type associated with the record. The attack types themselves are categorized into four specified groups (Table 1). We can focus on each attack group instead of considering the attack types themselves since attack groups represent enough knowledge to discriminate between different behaviors. Furthermore for a fast data inspection, it is more efficient to inspect fewer groups rather than inspecting all possible attacks. We will show in the section 4 that the result would not change dramatically.

4. Our direction: Selecting appropriate feature for each attack group

All data mining algorithms in the IDS should concentrate on some behavior features. In other words, the raw data shall be positioned in such a way that some meanings are extracted from them. In addition to saving time and space in processing data that feature selection provides, it may also help to make unprocessed data meaningful without almost any processing activity.

Therefore eliminating redundant and irrelevant features is one of the major open problems that has been still contemplated in many studies. We present a new approach in which we consider possible feature combinations for each type of attacks.

4.1. Related works and our contribution

Many papers propose feature selection as an approach in order that the worthless and ambiguous information would be removed. [14] employs the feature correlation analysis to find feature-feature and feature-class relationships. Their experimental outcomes have a superior performance compared to the feature selection algorithms in C4.5 and Naive Bayes, although we demonstrate that our approach has a better performance due to a more suitable feature selection mode, which is directly derived from attack groups characteristics.

Evolutionary algorithms are widely become exceedingly popular in feature selection. For instance [15] exploits Principle Component Analysis (PCA) to identify the most relevant features of the network connection. Nevertheless the weak point of their approach is the effectiveness which is principally the major problem of the most of evolutionary-based methods, although the combination of iterative algorithms (including evolutionary approaches) might achieve much acceptable performance. Using Genetic Search technique to find the best feature subsets is an approach experimented on KDD 99 by a multi-layered Neural Network (NN) in [16].

4.2. Category-feature correlation

Instead of particularly concentrating on the whole attack types to attain the most relative features and inspecting all records according to them, we notice to the category of each attack. This is because attacks in the same group have certain similarities which can be integrated in one group. The IDS then is able to thoroughly examine all records much more quickly and precisely.

On the other hand every attack group has its particular distinctions which have to be dealt separately. As an example, both categories DoS and Probe exercise repetitive connections in the relatively short periods while R2L and U2R happened usually in a single connection. All mentioned reasons stimulate a considerable demand on studying the features which are intimately related to the correspondent category.

To discover most affiliated features to each category, we have to find the most related features to each attack type, firstly, and then it becomes sufficient to calculate the frequency of each feature in every attack group. Fortunately [2] have listed the most effective feature for each attack class (Table 2). By accumulating the features of every attack type within an attack group, the most related features for every attack group are simply computed.

Table 2: Most efficient attribute in each attack class

Class Label	Feature #	Feature Name
Smurf	5	source bytes
Neptune	30	diff srv rate
Normal	5	source butes
Back	6	destination bytes
Satan	27	rerror rate
Ipsweep	37	dst host srv diff host rate
Teardrop	5	source bytes
Warezclient	5	source bytes
Portsweep	4	status flag
Pod	5	source bytes
Nmap	4	status flag
Guess_passwd	5	source bytes
Buffer_overflow	6	destination bytes
Land	7	Land
Warezmaster	6	destination bytes
Imap	3	Service
Loadmodule	6	destination bytes
Rootkit	5	source bytes
Perl	16	# root
ftp_write	5	source bytes
Phf	6	destination bytes
Multihop	6	destination bytes
Spy	39	dst host srv serror rate

Please note that each record in KDD 99 has 44 attributes representing properties of corresponding behavior whose individual number is represented as "Feature #" in the Table 2. In Table 3 we combine the features of all attacks in identical groups to specify the most relevant attribute for every category followed by their frequency.

4.3. Experimental Results

We use WEKA as a suite tool that implements many machine learning and data mining algorithms [12]. It is a free open source application available on the web and is currently applied in many research projects. The KDD 99 data set must firstly be converted into a supported format (ARFF) and then imported in the application. WEKA has the implementation of C4.5 called J48 which is used in our approach.

Table 3: Feature-Frequency in attack groups

Attack Category	Features	Frequency
DoS	source bytes	3
	diff srv rate	1
	destination bytes	1
	land	1
U2R	destination bytes	2
	source bytes	1
	# root	1
R2L	destination bytes	4
	source bytes	3
	service	1
	dst host srv serro rate	1
Probe	status flag	2
	rerror rate	1
	dst srv diff host rate	1

According to the Table 3 we select an applicable feature subset from all subsets in order to totally cover the attack groups. However attributes "protocol-type", "label" and "Service" are included in almost entire experiments since they are key features in training phase. We also split KDD 99 data set into two subsets. 75 percentage of it is used for the training phase of the algorithm and the rest is applied for testing phase.

Experimental results for four different feature combinations are shown in Fig. 2. Exp1 is the experiment that examines only features "protocol-type" and "label" which have been taken as distinctive attributes. No effective attributes from Table 3 is considered in this experiment and unsurprisingly the result is undesirable: %62.12. In Exp2 we append two other features that one is the key feature, i.e. "service" and the other is from Table 3: "destination-byte". In this case the preference is given to U2R and R2L attack groups which has amazingly the most favorable outcomes: %99.13. In the next experiment, Exp3, we select highly relevant features to the Probe attacks with frequency equivalent to 2 (Table 3). The training performance is just impressive as for Exp2; nonetheless the test result is unsatisfactory: %83.25. Finally, Exp4 exchanges floating feature with the one which is strongly related to the DoS attack group. Similarly the result for training phase is perfect but it is fairly disastrous for testing phase: %56.58.

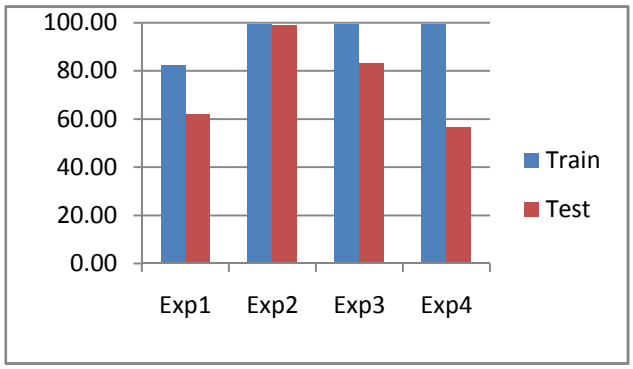

Fig. 2 : Experimantal results in various feature combinations

5. Conclusion and Future Works

Data mining based techniques play a vital role in current intrusion detection systems. The success rate of these methods is highly dependable to the types of attacks. Therefore reaching to a minimum false rate needs to find most relevant attributes to all attack classes and remove redundant data. This trend assists IDS designer in agile abnormal detection among a huge amount of data.

Our best result is regarding to the key features plus “destination bytes” which is the most effective attribute in U2R and R2L attacks. Despite our expectation the experimentation with feature combination of “source bytes” is unable to detect most of intrusions; although the “source bytes” attribute appears in large number of attacks as the most relevant feature. This means that the features existing in most of attacks (Classes in a classifying problem) are not wise choices for classification. Substitutionally, most frequent features appearing only in a limited number of classes are preferred options to achieve a remarkable accuracy.

Possible future works might be examining our approach on new KDD data sets where statistical anomalies have been removed. Combining the data mining approaches with evolutionary methods such as genetic algorithms or neural networks may direct us to a better performance on intrusion detection.

References

[1] M. A. Maloof, Machine Learning and Data Mining for Computer Security: Methods and Applications (Advanced Information and Knowledge Processing): Springer-Verlag New York, Inc., 2005.

[2] H. G. Kayacık, et al. (2005, Selecting Features for Intrusion Detection: A Feature Relevance Analysis on KDD 99 Intrusion Detection Datasets. [White Paper]. Available: Http://www.lib.unb.ca/Texts/PST/2005/pdf/kayacik.pdf

[3] S. Axelsson, "Intrusion Detection Systems: A Survey and Taxonomy," 2000.

[4] C. Kruegel, Intrusion Detection and Correlation: Challenges and Solutions: Springer-Verlag TELOS, 2004.

[5] R. Alshammari, et al., "Using Neuro-Fuzzy Approach to Reduce False Positive Alerts," presented at the Proceedings of the Fifth Annual Conference on Communication Networks and Services Research, 2007.

[6] C. Fortuna, et al., ed, 2007.

[7] I. Ahmad, et al., "Artificial neural network approaches to intrusion detection: a review," presented at the Proceedings of the 8th Wseas international conference on Telecommunications and informatics, Istanbul, Turkey, 2009.

[8] V. Chandola, et al., "Anomaly detection: A survey," ACM Comput. Surv., vol. 41, pp. 1-58, 2009.

[9] X. Wu, et al., "Top 10 algorithms in data mining," Knowl. Inf. Syst., vol. 14, pp. 1-37, 2007.

[10] R. Quinlan, C4.5: Programs for Machine Learning (Morgan Kaufmann Series in Machine Learning): Morgan Kaufmann, 1993.

[11] M. Tavallaee, et al., "A detailed analysis of the KDD CUP 99 data set," presented at the Proceedings of the Second IEEE international conference on Computational intelligence for security and defense applications, Ottawa, Ontario, Canada, 2009.

[12] I. Witten, et al., "Weka: Practical machine learning tools and techniques with java implementations," ed, 1999.

[13] T. U. K. Archive. (1999, KDD Cup 1999 Data Available: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[14] T. S. Chou, et al., "Network Intrusion Detection Design Using Feature Selection of Soft Computing Paradigms," International Journal of Computational Intelligence, pp. 196-208, 2007.

[15] Z. Bankovi, et al., "Improving network security using genetic algorithm approach," Comput. Electr. Eng., vol. 33, pp. 438-451, 2007.

[16] K. M. Faraoun and A. Rabhi, "Data dimensionality reduction based on genetic selection of feature subsets," ed, 2007