JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 06, ISSUE 07, 2019

Estimating Vehicle Fuel Consumption and Emissions Based on Instantaneous Speed and Acceleration Levels

J.VijaySree,N. Mownika,Dr.M. Murugesan
Assistant Professor1,2,Professor3
Dept. of CSE,
mail-id:jvijaya.pavani@gmail.com , mounika.nekkanti@gmail.com,murugesanvim@gmail.com
Anurag Engineering College,Anatagiri(V&M),Suryapet(Dt),Telangana-508206

Abstract - The rapid development of business and othertransaction systems over the Internet makes computer securitya critical issue. In recent times, data mining and machine learning have been subjected to extensive research in intrusion detection with emphasis on improving the accuracy of detection classifier. But selecting important features from input data lead to a simplification of the problem, faster and more accurate detection rates. In this paper, we presented therelevance of each feature in KDD '99 intrusion detection dataset to the detection of each class. Rough set degree of dependency and dependency ratio of each class were employed to determine the most discriminating features for each class. Empirical results show that seven features were not relevant in the detection of any class.

Keywords: Intrusion detection, machine learning, relevance feature, rough set, degree of dependency.

INTRODUCTION

As Internet keeps growing with an exponential pace, so also iscyber attacks by crackers exploiting flaws in Internet protocols, operating system and application software. Several protective measures such as firewall have been put in place to check the activities of intruders which could not guarantee the full protection of the system. Hence, the need for a more dynamic mechanism like intrusion detection system (IDS) as a second line of defense. Intrusion detection is the process of monitoring events occurring in a computer system or network and analyzing them for signs of intrusions [1]. IDSs are simply classified as host-based or

This manuscript was submitted May, 2007. The work was self sponsored.

Adeola, S. Oladele is an Oracle Certified Professional with a core competency is Oracle Database, Microsoft Basic, VB.Net, PL/SQL and Computer Networking. He is a member of professional bodies such as Computer Professional of Nigeria (CPN), Nigeria Computer Society (NCS), IEEE Computer society as well as Association of Computer Machinery (ACM). He has worked in different companies as a Network Engineer and Programmer. He was a consultant to a number of establishments in Nigeria including ALCATEL Nigeria and Nigeria Police Force. He is currently with the Federal University of Technology, P.M.B 704, Akure, Nigeria (phone: +234-8033749944; e-mail: deleadeola@yahoo.com)

Adetunmbi A. Olusola holds a PhD in Computer Science from the Federal University of Tech., Akure, Nigeria. He worked in different organization in Nigeria including Associated Business Information and Computer Services, Lagos, Nigeria. He was also a lecturer at Adeyemi College of Education, Ondo, Nigeria and University of Ado-Ekiti, Nigeria. He is a member of professional bodies such as Nigeria Computer Society, IEEE Computer Society and International Studies on Advanced Intelligence. He is currently a researcher with the Federal University of Tech., P.M.B 704, Akure, Nigeria (e-mail: bayo_adetunmbi@yahoo.com)

Daramola, O. Abosede holds a M.Tech degree in Computer science. He is a member of different professional bodies such as Nigeria Computer Society, Third World Organization of Women Scientists and Science Association of Nigeria. She is currently pursuing her PhD at the Department of Computer Science of the Federal University of Technology, Akure, P.M.B 704, Akure, Nigeria

network-based. The former operates on information collectedfrom within an individual computer system and the latter collect raw networks packets as the data source from the network and analyze for signs of intrusions. The two different detection techniques employed in IDS to search forattack patterns are Misuse and Anomaly. Misuse detection systems find known attack signatures in the monitored resources. Anomaly detection systems find attacks by detecting changes in the pattern of utilization or bahaviour ofthe system.

Majority of the IDS currently in use are either rule-based or expert-system based. Their strengths depend largely on theability of the security personnel that develops them. The former can only detect known attack types and the latter is prone to generation of false positive alarms. This leads to theuse of an intelligence technique known as data mining/machine learning technique as an alternative to expensive and strenuous human input. These techniques automatically learn from data or extract useful pattern from data as a reference for normal/attack traffic behaviour profile from existing data for subsequent classification of network traffic.

Intelligent approach was first implemented in mining auditdata for automated models for intrusion detection(MADAMID) using association rule [2]. Several others machine-learning paradigms investigated for the design of IDS include: neural networks learn relationship between given input and output vectors to generalize them to extract new relationship between input and output [3,4,5], fuzzy generalize relationship between input and output vector basedon degree of membership [5,6], decision tree learns knowledge from a fixed collection of properties or attributes in a top down strategy from root node to leave node [5,7,8], support vector machine simply creates Maximum-marginhyper planes during training with samples from two classes [3,9,10] .

Rough sets produce a set of compact rules made up of relevant features only suitable for misuse and anomalous detection (9,11,12,13,14]. Bayesian approaches are powerful tools for decision and reasoning under uncertain conditions employing probabilistic concept representations [15,16].

Prior to the use of machine learning algorithms raw network traffic must first be summarized into connection recordscontaining a number of within-connection features such as service, duration, and so on. Identification of important features is one of major

factors determining the success of any learning algorithm on a given task. Feature selection in learning process leads to reduction in computational cost, over fitting, model size and leads to increase in accuracy.

Previous works in feature selection for intrusion detection include the work of [17, 18]. In this paper, attempt was made to investigate the relevance of each feature in KDD 99 intrusion detection dataset to substantiate the performance of

certainty to belong to the subject of interest, while upper approximation is a description of objects which possibly belong to the subset [19].

Definition 1:

The following is the outline for this paper: Section 2: A Description of refers to a non-empty, limited collection of characteristics that may be utilized to objects; V characterizes the values of all attributes;

After introducing the evaluation dataset for intrusion detection, we briefly discuss the rough set and discretization methods.

if the characteristics in S are split into two independent sets, condition attributes (C) and decision attributes (D), with A = C D and C D =, then the resulting table is a decision table (DT).

Our experiment use the intrusion detection benchmarking dataset KDD Cup 1999 [21]. Dataset consisted of simulated raw TCP dump information gathered U,C,D,V, f DT

The characteristics that fall within the B A categorz specifies a synonym nine weeks on a LAN, or Local Area Network. The Instruction

Seven weeks of network traffic and two weeks of testing data were processed to a total of around five million connections records.

Almost two million relationship records were gleaned from the data. The

relationship (also known as the "Indiscernibility Relationship") on the set U (represented by the symbol IND) (B).

Out of the total of 39 possible assaults, only 22 are included in the training data.

For any two numbers x and y, the expression "IND(B)" is written as "xUxU | f (x, b) f (y, b)b B.(2) included in the experimental data When we talk about attacks, we may divide them into two categories: those we're already familiar with, which are those found in the training dataset, and those we're just learning about, which are the new assaults found in the test datasets. There are four broad types of assault:

Synonymous with "Denial of Service," or "DOS," are attacks such syn flooding.

Surveillance and other forms of probing, such port scanning, are examples of (2)probing.

3)U2R: buffer overflow attacks, or any other means of gaining local root access without permission.

Classes with the same B-indiscernibility equivalents are B represents a variable with unknown value, indicated by symbol [x].

This expression may also be written as [x]B y U | (x, y) IND(B).

To further define, if we have B A and X U, then X approximated by building the B lower and B-upper approximations of set X defined as:

Password guessing or other forms of unwanted remote access ((4).R2L)

There were a total of 494,021 entries in the training dataset; 97,277 (19.69%) were considered typical, while 391,458 (79.24%) were classified as having some kind of DOS.

There were 4,107 (0.83%) Probe, 1,126 (0.23%) R2L, and 52 (0.01%).

Interactions between the upper and lower extremities, or U2R.

There are 41 properties per connection, each defining a distinct aspect of the relationship, and a label.

BX = [x]B X 0 | x X

Defined Third: If A = C D, then C D =. In the relation IND (D), POSC (D), the positive region for a given set of condition attributes C is defined as designated as a kind of assault or as the norm for each individual. Class labels and sample counts for the "10% KDD" training dataset are shown in Table 1. Specifically, Appendix II contains:

Feature Extraction from the KDD 99 Intrusion Detection Dataset.

II. PRINCIPLES OF ROUGH SET DESIGN

When dealing with ambiguity or a lack of information, Rough Set may help you minimize the size of your data sets, uncover previously unseen patterns, and arrive at sound decisions. In a significant way, the idea of a reduct owes its existence to rough set

theory. , with D* standing for the set of equivalence classes specified by the IND connection (D). For each set U, POSC(D) includes all items that fall neatly into

The IND-defined categories (D).
The positive area also includes all objects of type U that can be partitioned into blocks of type Q using attribute B, given attribute subsets B, Q A. Definition of "B"

Determines which attribute subsets are necessary for a predictive model to work. If you have a collection of discrete data and want to get rid of any duplicate characteristics, rough sets are the way to go

Entropy is a supervised splitting approach that uses the class label to assess the informativeness of a certain input attribute about the output attribute for a subset. One defining feature [20] is the search for the partition that maximizes information gain. A quick calculation yields this result:

Consider the training set D, which consists of a list of qualities and their labels.
one method is to count how many times each class is used in the dataset and use that information to calculate the degree of dependence across classes. As a result, it represents the feature's ability to distinguish the target class from others. Second, there is a mapping between class labels and other attributes. That is, in order to detect all the relevant features distinguishing one class from another, one must first generate a frequency table of a particular class label against others based on variations in each attribute, and then compare these tables to generate the dependency ratio of predominant classes (see Appendix I for details). This research also makes use of graphical analysis to identify important traits for each class.

Calculating the dependence ratio is as easy as An expression for the entropy of the variable D is as follows
If possible, when choosing a spilt-point for attribute A, choose a value for A that yields the bare minimum of information, as achieved when E(D,T) is minimal. To do this, we use a recursive procedure on an attribute where the information need is low (0).

dependability and categorization into two distinct groups. In other words, for each class, an instance in the dataset is in-class if it has the same label as the class, and out-class if it does not. For each class label in a dataset, a dependency degree is calculated using the class's instance count. Class labels in the training set are most dependent on other class labels as seen in Table 2. The dependence ratio of the most important characteristics chosen for each class is shown in Table 3. Of the 23 groups, the DOS category accounts for half of the six that use the quantity of data sent as a defining characteristic. Whether the assault is a denial of service or a probe, then its very short or very lengthy duration is par for the course.

This may be written as Ent(S) = E(T, S) =.

(10)

connections. It was determined that criterion 7 (which is connected to land attacks) was the most telling.
LABORATORY TESTS AND RESULTS

This study makes use of the "10% KDD" (kdncup data gz file) dataset as its training set. Following the discussion in Section 3.1, continuous features are discretized in order to compute the degree of dependence for discrete features. As rough set does not need duplicate instances to classify and find discriminating features, the discretization process begins with the elimination of redundant records from the dataset.

Feature 8 (wrong fragment) was shown to be the most useful in differentiating pod and teardrop attacks. The study also found a high degree of reliance on the third element, "Service," which indicates that various services are abused to carry out a variety of attacks. Imap4, ftp data, and telnet are all vulnerable to attacks like imap, warezclient, and buffer overflow. The most discriminatory name for each trait is shown in Table 4. The most common appearances are the Normal, Neptune, and Smurf types. Table 1: Class labels and the number of samples that appears in "10% KDD" dataset

Attack	Original Number of Samples	Number of samples after removing duplicated instances	Class
back	2,203	994	DOS
land	21	19	DOS
neptune	107,201	51,820	DOS
pod	264	206	DOS
smurf	280,790	641	DOS
teardrop	979	918	DOS
satan	1,589	908	PROBE
ipsweep	1,247	651	PROBE
nmap	231	158	PROBE
portsweep	1,040	416	PROBE

normal	97,277	87,831	NORMAL
Guess_passwd	53	53	R2L
ftp_write	8	8	R2L
imap	12	12	R2L
phf	4	4	R2L
multihop	7	7	R2L
warezmaster	20	20	R2L
warezclient	1,020	1020	R2L
spy	2	2	R2L
Buffer_overflow	30	30	U2R
loadmodule	9	9	U2R
perl	3	3	U2R
rootkit	10	10	U2R

Table 2: Attribute with the highest degree of dependency that distinctly distinguish some class labels from the training data set.

Attack	Degree of dependency	Selected features	Feature Name	Other distinct features
back	0.9708	5	source bytes	6
neptune	0.0179	3	service	39
teardrop	0.9913	8	wrong fragment	25
satan	0.0319	30	diff srv rate	27,3
portsweep	0.0264	4	flag	30,22,5
normal	0.0121	6	destination bytes	5,3,10,11,1
guess_passwd	0.0189	11	failed logins	-
imap	0.3333	26	srv error rate	-
warezmaster	0.7500	6	destination bytes	-
warezclient	0.2686	10	hot	5,1

discriminating classes for most of the features which consequently make their classification easier. Moreover, these three classes dominating the testing dataset and this account to high detection rate of machine learning algorithm on them. The research also shows how importanta particular feature is to detection of an attack and normal. For some class label a feature sufficient to detect an attack type while some requires combination of two or more features. For features with few representatives in the dataset such as spy and rootkit, it is very difficult detecting a feature or features that can clearly differentiate thembecause of the dominance of some class labels like normal and Neptune. These difficult to classify attacks belong totwo major groups, user to root and remote to local. The

involvement of each feature has been analyzed forclassification. Features 20 and 21 (see appendix I) make no contribution to the classification of either an attack or normal. Hence these two features (outbound commandcount for FTP session and hot login) have no relevance in intrusion detection. There are other features that makeslittle significant in the intrusion detection data set. Fromthe dependency ratio table in Appendix I, these features include 13, 15, 17, 22 and 40 (number of compromised Table 3: The most relevant feature for each attack type and normal conditions, su attempted, number of file creation operations, is guest login, dst host rerror rate conditions, su attempted, number of file creation operations, is guestlogin, dst host rerror rate respectively). Table 3: The most relevant feature for each attack type and normal

Attack	Most relevant features	Feature Name	Variations	Dependency ratio	Class
Back	5	source bytes	66,64,60	0.9708	DOS
Land	7	land	2	0.9999	DOS
neptune	5	source bytes	0	0.9328	DOS
Pod	8	wrong fragment	1	0.9853	DOS
Smurf	5	source bytes	39	0.7731	DOS
teardrop	8	wrong fragment	2	0.9913	DOS
Satan	30	diff srv rate	30	0.7648	PROBE
ipsweep	36	dst host name src port rate	13,14,15,17	0.8282	PROBE
Nmap	5	source bytes	4	0.6448	PROBE
portsweep	28	srv error rate	9	0.8057	PROBE
normal	29	same srv rate	28	0.8871	NORMAL
guess_passwd	11	failed login	1	0.9622	R2L
ftp_write	23	count	1	0.7897	R2L
Imap	3	service	60	0.9980	R2L
Phf	6	destination bytes	28	0.9976	R2L
multihop	23	count	1	0.7898	R2L
warezmaster	6	destination bytes	33	0.7500	R2L
warezclient	3	service	13	0.6658	R2L
Spy	39	dst host srv serror rate	8	0.9997	R2L
buffer_overflow	3	service	6	0.6965	U2R

loadmodule	36	dst host name srcport rate	29	0.6279	U2R
Perl	14	root shell	1	0.9994	U2R
rootkit	24	srv count	1	0.7269	U2R

CONCLUSION

In this paper, selection of relevance features is carried out on KDD '99 intrusion detection evaluation dataset. Empirical results revealed that some features have no relevance in intrusion detection. These features include 20 and 21 (outbound command count for FTP session and hot login) while features 13, 15, 17, 22 and 40 (number of compromised conditions, su attempted, number of file creation operations, is guest login, dst host rerror rate respectively) are of little significant in the intrusion detection.

In our future work, additional measures including sophisticated statistical tools will be employed

. **REFERENCES**

Reference: Bace, R., and Mell, P. (2001). NIST SP 800: Intrusion Detection System, November.
Lee, W., S.J. Stolfo, and K. Mok [2]. (1999). Intrusion detection testing using data mining in a workflow setting. Proceedings of the Ninth Annual Conference on Knowledge Discovery and Data Mining (1999).
According to [3] (Mukkamala, S., Janoski, G., & Sung, A. (2002). Neural networks and support vector machines for intrusion detection. IEEE International Joint Conference on Neural Networks Proceedings, Volume 2, Issue 3, Pages 1702-1707.

According to [4]'s Byunghae, C., kyung, W.P., and Jaittyun, S., "Neural Networks Approaches for Host Anomaly Intrusion Detection Using Fixed Pattern Transformation at the International Conference on Computer Science and its Applications," LNCS 3481, pp. 254-263.
Ajith, A., Ravi, J., Johnson, T., and Sang, Y.H. (2005). Journal of Network and Computer Applications, Elsevier, pp. 1–19, "D-SCIDS: Distributed Soft Computing Intrusion Detection System."
Susan M.B. and Rayford B.V. (2000). Proceedings of the 12th Annual Canadian Information Technology Security Conference, Ottawa, Canada, June 19-23, 2000, Pages 109-122; Intrusion detection using fuzzy data mining.
Quinlan, J.L. (1993) C4.5 Software for Machine Learning, To paraphrase, Morgan Kaufmann Publishers, Inc.
Pavel L., Patrick D., Christia S., and Konrad R. (2005). Supervised or Unsupervised Learning for Intrusion Detection, International Conference on Image Analysis and Processing, 2005 (3617), Italy, pages 50–57.
Zhang L, Zhang G, YU L, Zhang J, and Bai Y. (2004) Intrusion detection using Rough Set Classification, Journal of Zhejiang University SCIENCE, 5(9), 1076-1086 [9].
[10] Machine Learning Approach to Real-time Intrusion Detection System, by K. Byung-Joo and K. Il-Kon, was published in 2005 in Lecture Notes in Artificial Intelligence, issue 3809. Printed on acid-free paper at Springer-verlag Berline, Heidelberg, between pages 153 and 163, and edited by S.Zhang and R.Jarvis.
[11].
Alese, B.K.; Adetunmbi, A.O.; Falaki, S.O.; Adewale, O.S. (2007a) The Basic Plan

for Identifying Familiar and Unfamiliar Network Intrusion, AICTTRA 2007: The Second International Conference on the Integration of Information and Communication Technologies Into Academic and Research Institutions

OAU, Ife, pages 190–200, Kehinde, L.O., E.R. Adagunodo, and G.A. Aderounmu.
[12].
The authors of this paper are Adetunmbi, A.O., Alese, B.K., Ogundele, O.S., and Falaki, S.O. (2007b). Data Mining for Network Intrusion Detection, Journal of Computer Science and Its Applications, 14(2), 24–37. [13]. Intrusion Detection Using Rough Sets and k-Nearest Neighbors. Adetunmbi, A.O., Falaki, S.O., Adewale, O.S., and Alese, B.K. (2008). International Journal of Computing and ICT Research. Volume 2. Pages 61–66.
[14]. Transactions on Rough Sets IV, LNCS 3700, 2005, pp. 144–161; Sanjay, R., V.P. Gulati, and K.P. Arun. 2005. A Rapid Host-Based Intrusion Detection System Utilizing Rough Set Theory.
[15]. Authors: Axelsson, S. (1999). ACM, "The Base-Rate Fallacy and Its Implication for the Difficulty of Attack Detection," in Proceedings of the Sixth ACM Symposium on Computer and Communication Security, pp. 127–141, 2010.
[16]. Naive Bayes vs Decision Trees for Intrusion Detection Systems, ACM Conference on Applied Computing, 2004, pages 420–424. Amor, N.B., S. Beferhat, and Z. Elouedi.
Reference: Sung, A.H., and Mukkamala, S., "Identifying Key Characteristics for Intrusion Detection using Support Vector Machines and Neural Networks," IEEE Proceedings of the 2003 Conference on Applications and the Internet, p.
[18]. H.G. Kayacik; A.N. Zincir-Heywood; M.L. Heywood (2006). Choose the Right Characteristics for Intrusion Detection: An Examination of the KDD 99 Intrusion Detection Datasets.
[19]. Rough Sets: A Tutorial. Komorowski, J., L. Pokowski, and A. Skowron. 1998. citeseer.ist.psu.edu/komorowski98rough.htm
Data Mining Concepts and Techniques, 2nd Edition, Jiawei, H., and Micheline, K. (2006), China Machine Press, pp. 296–303.
[21]. These are the datasets from the 1999 KDD Cup: http://kdd.ics.uci.edu/kddcup99/