Chinese Traditional Medical Journal

An Advanced Ensemble Enhanced Feature Selection Approach for Intrusion Detection System

Pegg Bosh, Heik Staud, Sujun Ye

Donders Centre for Cognition, Radboud University Nijmegen, Nijmegen
Psychiatric Research Institute, LVR-Klinik Bedburg-Hau, Bedburg-Hau
Research Group of Pain and Neuroscience, Kyung Hee University, Seoul

Abstract : In recent years Intrusion Detection Systems (IDS's) play a key part in data security. The goal of IDS is to assist computer systems in dealing with attacks. the IDS for identifying the attacks effectively has been suggested and actualized. For this reason, another component determination method called

Optimal Feature Selection approach. The idea of an Information Gain Ratio has been put up in light of this strategy. The primary purpose of algorithm determines appropriate amount of features from KDD Cup dataset. SVM and Rule Based Classification have both been used to characterise the data, making it possible to categorise it even more effectively. In comparison to previously published findings, our OFSA model shows promise in detection.

.Keywords: Intrusion Detection, Information Gain, Support Vector Machine Feature Selection Technique and Classification

Introduction

2. As network based PC frameworks assume increasingly important parts in modern civilization, they have turned into the targets of our enemies and hoodlums. As a result, we must seek out the most optimal methods possible in order to safeguard our foundations. The security of a PC framework is traded off when an interruption arises. An interruption may be classified [HLMS90] as "any arrangement of operations that aim to trade off the honesty, categorization or accessibility of an asset". Client validation (e.g. using passwords or biometrics), avoiding programming errors, and data assurance (e.g. encryption) are all interruption counteraction measures that have been used to assure PC frameworks as a first line of defence. Interruption counteractive action alone is not adequate in view of the way that since frameworks grow out to be persistently unexpected, there are constantly exploitable inadequacies in the

frameworks because of outline and programming faults, or various "socially intended" infiltration strategies. The existence of exploitable "cushion flood" due to programming errors in certain late framework programming was first documented many years ago and has persisted ever since.

3. Because of the tactics that balance ease of use with stringent control over a framework and its data, it is now impossible for a business activity to be completely safe. As a secondary line of defence against unauthorised access, intrusion detection systems are essential. To guard these frameworks from being assailed by interlopers, another Intrusion Detection System has been designed and built in this project effort, which combines a basic feature selection method and SVM approach to discover

.

4. attacks. Using KDD cup data set and Data Mining retrieve the hidden predictive information from big Databases. Using this powerful new technology, corporations can zero in on the most critical information in their data warehouses, which has enormous potential.

.

5 Any kind of data repository may be linked to information mining. Algorithms and techniques, on the other hand, might differ depending on the kind of data they are working with. In recent years, the internet has become an integral part of our daily lives. The current virtual worlds taking into consideration data preparation systems are predisposed to various type of risks which lead to diverse sorts of damages bringing about major calamities. As a result, data security is becoming more important. Protecting administrative systems against unauthorised access via disclosure, interruption, alteration, or pulverisation is the most important goal of system security. System security also reduces the risks associated with the core security goals, such as confidentiality, integrity, and openness.

6 Related Work

Lately, network security has been the focus of various study efforts with arrival of web. There are several books in the literary that discuss about Intrusion Detection System. IDSs are applied to distinguish the attacks made by gatecrashers. [1]Sindhu et al presented a heredity based component determination calculation for decreasing the computational complex nature of the classifier.

Jianping Li et al [2] proposed another technique taking into account Continuous Random Function for picking suitable capabilities to execute system interruption discovery. Many SVM-related order computations may be found in the IDS writings. For instance, a computation called Tree Structured Multiclass SVM has been presented by Snehal A. Mulay et al [3] for grouping information viably. There are several works in the literature that evaluate about pre- preparation. The greater part of the actual troubles undoubtedly demand an ideal and deserving arrangement as opposed to figuring them utterly at the price of debased execution, time and space. The element choice quest started with invalid set where elements were incorporated one by one or it was commenced with a complete arrangement of elements where components were dispensed with one by one. Li et al suggested a wrapper based element choice computation with a defined end aim to build up an IDS. Using Geetha Raman's[4] component determination algorithm, we can better parse the massive KDD Cup dataset. There are various works in the literature that explore about grouping methods and tools.

In the beginning, classifiers like Bolster Vector Machines (SVM) were designed with paired characterisation in mind. IDS's

Neural Network model was developed by Debar et al.. Dewan Md. Farid proposed another learning approach for system interruption recognition utilising innocent Bayesian classifier and ID3 algorithm is introduced, which recognises compelling traits from the preparation dataset, ascertains the restrictive probabilities for the best property estimations, and after that accurately

7. each and every instance of setting up and running tests on a dataset is grouped together. The SVM- based interruption recognition framework consolidates a varied levels bunching technique, a fundamental component choosing methodology, and the SVM process.

8 Proposed Approach

A. Information Preparation Subsystem

1) Information Collector

The records from the KDD'99 cup data set are gathered by the data accumulation operator. The pre-processing module receives this data and uses it to pre-process the data. The records acquired from the KDD cup dataset could be a normal information or an assaulted information.

2) Pre-processing Module

Pre-processing strategies are critical for information decreasing because it is extremely hard to manage huge amount of system movement information with all components to detect gatecrashers progressively and to supply anticipation procedures.

B. Classification Subsystem

1) Rule Based Classifier

The application of rules fired by the rule system called by intelligent agents improves judgments on anomalous intrusion detection and prevention in this system. Rule-based decision-making on incursions is made easier with the use of a knowledge base.

2) Support Vector Machine

SVM is the learning machine that can execute double order and relapse estimating jobs. They are coming out to be progressively well known as another worldview of order and learning on account of two key aspects. To start with, different to the following arrangement systems, SVM reduces the usual blunder as opposed to minimising the characterisation error. To achieve a twofold problem, SVM uses the duality hypothesis of numerical programming, which admits powerful computing techniques.

C. Proposed Algorithm for Optimal Feature Selection Approach

The Information Gain Ratio for property selection was used to build this method. Keeping in mind the final objective to do this, the information set D is partitioned into n number of classes Ci. The qualities Fi having highest quantity of non-zero qualities are selected by the professional and the Information Gain Ratio (IGR) is figured utilising conditions:

$$\text{Info (D)} = -\sum_{j=1}^{m}\left[\frac{\text{freq(Cj,D)}}{|D|}\right]\log_2\left[\frac{freq(C_j,D)}{|D|}\right]\text{.....(}$$

$$\text{Info (F)} = \sum_{i=1}^{n}\left[\frac{|F_i|}{|F|}\right] * Info(|F_i|)\text{.....(2)}$$

$$\text{IGR (Ai)} = \left[\frac{Info(D)-Info(F)}{Info(D)+Info(F)}\right] * 100\text{.....(3)}$$

Ten key components have been identified by the OFS algorithm in order to more quickly detect possible attacks.

12. IMPLEMENTATION

A. Enhanced Feature Selection

Algorithm: Intelligent Agent based Property Selection Algorithm

Input: Set of 41 features from KDD'99 Cup data set

Output: Reduced set of features R

Step 1: Select the traits which have variety in their qualities.

Step 2: Calculate the Info (D) values for the chose traits utilizing the condition 1.

Step 3: Select the traits which have most extreme number of non- zero values.

Step 4: Calculate the Info(F) esteem for the properties chose in step 3 utilizing the condition 2. Step 5: Calculate the IGR esteem utilizing the condition 3.

Step 6: Depending on the IGR esteem, select the characteristics.

The conventional component choice techniques set aside substantial computation time for determining IGR values. A new component determination technique, dubbed the Enhanced Feature Selection algorithm, is now presented and implemented in this study in order to reduce computation time. This procedure calculates the Information Gain Ratio (IGR) esteem for the altering characteristics in the information collection. It performs segment diminution in light of the IGR esteem. OFS enhances the exactness in identification and reduces the false caution rates. All of the re-enacted attacks fall under one of four categories: Denial of Service (DoS), User to Root (U2R), Remote to Local (R2L), or Probe assault.

TABLE 1 THE 41 FEATURES IN KDD'99 DATASET

S.NO	FEATURE NAME	S.NO	FEATURE NAME
1	Duration	22	Is_guest_login
2	Protocol type	23	Count
3	Service	24	Serror_rate
4	Src_byte	25	Rerror_rate
5	Dst_byte	26	Same_srv_rate
6	Flag	27	Diff_srv_rate
7	Land	28	Srv_count
8	Wrong_fragment	29	Srv_serror_rate
9	Urgent	30	Srv_rerror_rate
10	Hot	31	Srv_diff_host_rate
11	Num_failed_logins	32	Dst_host_count
12	Logged_in	33	Dst_host_srv_count
13	Num_compromised	34	Dst_host_same_srv_count
14	Root_shell	35	Dst_host_diff_srv_count
15	Su_attempted	36	Dst_host_same_src_port_rate
16	Num_root	37	Dst_host_srv_diff_host_rate
17	Num_file_creations	38	Dst_host_serror_rate
18	Num_shells	39	Dst_host_srv_serror_rate
19	Num_access_shells	40	Dst_host_rerror_rate
20	Num_outbound_cmds	41	Dst_host_srv_rerror_rate
21	Is_hot_login		

B. Calculation of info (D)

The data pick up base is gained from data hypothesis. The important notion of data hypothesis is that the data carried on by a message relies on upon the probability and may be quantified in bits as less the logarithm of base 2 of that likelihood. Think of a dataset D that has q classifications C1..Cn. Assume additionally that we have a hypothetical test x with m outcomes that allotments D into m subgroups D1….Dm. Since parallel split is all we're doing, m=2 for a numeric quality. The chance that is picked one record from the set D of information records and report that if has a location with some class Cj is provided by ,

m j=1

freq(Cj,D)

Where freq (Cj, D) refers to the amount of information records(points) of the class Cj in D, whereas |D| is the aggregate number of information records in D. So the info that is handed on is

−log2

[freq(Cj,D)] bits
|D|

To discover the normal data anticipated to differentiate the class of an information record in D before apportioning occurs, summing is conducted across the classes in extent to their frequencies in D, providing

Info(D)=− ∑m

[freq(Cj,D)] log
|D|

[freq(Cj,D)] (1)
|D|

The dataset D has been partitioned into m equal parts based on the test x findings. The normal measure of data anticipated to differentiate the class of an information record in D after the parcelling has occurred, may be determined as the weighted whole over the subsets, as:

Info (F) = ∑n

[|Fi|] ∗ Info(|F |)(2)

i=1

|F| i

where |Fi| denotes the number of data records in the subset Di after the partitioning has happened.

The information received owing to the partition is:

Gain(Ai) = Info(D) – Info(F) (6)

Plainly, it is vital to raise the addition. The addition foundation is to pick the test or slice the widens the increase to parcel the existing information

Info(D)−Info(F)

IGR (Ai) = [( )

] ∗ 100 (3)

Info D +Info(F)

6. RESULT

Underneath, we evaluate and plan the execution and time investigation for the various types of attacks. Table's exhibits the discovery exactness and calculation time got using the parts of the KDD'99 Cup information set by applying the element choosing procedures of current and prospective work.

A. Rule Based Classification

The Rule based characterisation is the beginning step in the organisation of several types of attacks. The execution examination as far as accuracy and time spent for characterising the attacks using Rule Based Classifier is established in the TABLE 2. The accuracy and processing time for recognising 5000 records are shown in the table below.

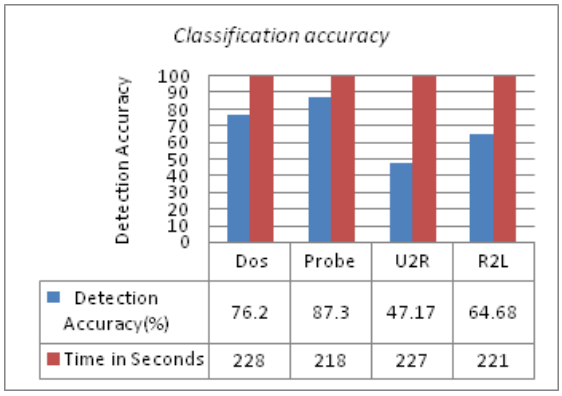
Attacks	No. of Records	Detection Accuracy(%) for selected features	Detection Accuracy(% for tot features
Dos	1581	99.11	99.11
Probe	1902	92.03	96.31
U2R	1745	91.51	96.15
R2L	1745	91.51	96.15

Table 1 Performance Analysis For Rule Based Classifica

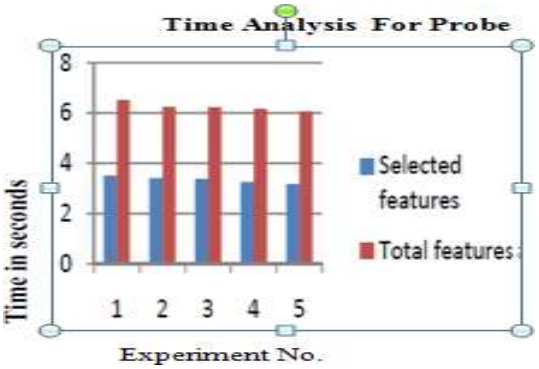
Attacks	Detection Accuracy(%)	Time in Seconds
Dos	76.2	228
Probe	87.3	218
U2R	47.17	227
R2L	64.68	221

Table 2 Performance Analysis

The Classification and detection accuracy for rule based classification Table 3time Analysis For U2R Attack In Svm



Experiment No	Accuracy(%)	
	selected features using OFS(10)	Total features (41)
1	3.51	6.52
2	3.41	6.25
3	3.37	6.23
4	3.26	6.17
5	3.17	6.07
Avg	**3.34**	**6.24**



7. Conclusion

By combining the EFS algorithm with two order techniques and a novel IDS, this research proposes and implements a more secure framework. The computation time necessary for identifying and arranging the records employing all the forty one aspects of the KDD'99 cup information set is observed to be large. The suggested highlight choice computation choices only the key parts that aid in lowering the time spent for identifying and sorting the records.

As an added bonus, SVM outperforms the standard-based classifier in terms of

accuracy. The fundamental point of interest of the suggested IDS is that it minimises the false positive rates additionally lessens the computation time.

References

1. 1. “Analysis of KDD’99 Intrusion Detection Dataset for Selection of Relevance Features” byDaramola O. Abosede, Adetunmbi A. Olusola, AdeolaS.Oladele,.,from Proceedings of the World Congress on Engineering and Computer Science, Vol. I, October 20-22, 2010.
2. 2. “Intrusion Detection System utilising Support Vector Machine and Decision Tree “, by Devale.P.R,Garje.G.V.,SnehalA.Mulay, 2012. International Journal of Computer (0975– 8887), Vol. \s3, June 2010.
3. Debar, Becker, and Siboni, 1992, "A Neural Network Component for an Intrusion Detection System," IEEESymposium on Research in Computer Security and Privacy, p. 240250;
4. (4) Du Hongle, TengShaohua, and Zhu Qingfang, "Intrusion detection based on Fuzzy support vector machines", International Conference on Networks Security, Wireless Communications, and Trusted Computing, pp. 639–642, 2009.
5. 5. Wei Lu, MahbodTavallaee, EbrahimBagheri, Alia A.Ghorbani,. Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Security and Defense Applications, Vol. 97, pp. 4244–37641, 2009. A detailed analysis of the KDD CUP 99 dataset.
6. 6. Weka software, Machine Learning. “Weka 3–Data Mining using Open Source Machine Learning Software in Java” Machine Learning Group at University of Waikato Website,

.