International Journal For Recent Developments in Science & Technology

2581-4575

Crossref

A Peer Reviewed Research Journal

A FILTER BASED FEATURE SELECTION ALGORITHM FOR NIDS USING MULTI CLASS LS-SVM

[1]K.VAMSIKRISHNA, [2]SD.SIRAJUNNISA, [3]M.N.V. LOKESWARAO, [4]N.MEENA

[1]Associate Professor, CSE Department, NRI INSTITUTE OF TECHNOLOGY, POTHAVARAPPADU
[2,3,4]UG Student, CSE department, NRI INSTITUTE OF TECHNOLOGY, POTHAVARAPPADU

ABSTRACT: Redundant and inapplicable features in information have caused an extended haul issue in network traffic classification. These highlights backtrack the procedure of arrangement still as keep a classifier from selecting precise selections, significantly once adapting to large information. Mutual information based mostly algorithmic program has used that analytically selects the simplest feature for classification. We propose a Hybrid Feature Selection algorithm that analytically selects the optimal feature for classification. This mutual information based principally feature selection algorithmic program will handle dependent information options linearly and nonlinearly. Its effectiveness is evaluated among the cases of network intrusion detection. Associate in Network Intrusion Detection System (IDS), named least sq. Support Vector Machine primarily based IDS (LSSVM-IDS), is made exploitation the features elite by the planned feature selection algorithmic program. The performance of LSSVM-IDS is evaluated by intrusion detection analysis datasets, significantly KDD Cup ninety nine dataset. The results of LS-SVM +HFSA show that our feature selection algorithm contributes more critical features for LSSVM-IDS to grasp better accuracy and lower computational cost compared with the state-of-the-art methods.

KEYWORDS: Feature Selection, Intrusion Detection, Least sq. Support Vector Machine, support vector machine

1. INTRODUCTION

Over the past decades, web and pc systems have raised various security problems thanks to the explosive use of networks. Any malicious intrusion or attack on the network could create to serious disasters. Intrusion could be a malicious, harmful entity that is chargeable for network attack. They violate integrity, confidentiality and availableness of a system resource. During this case, system is did not respond for information taken or being lost. So, Intrusion Detection Systems (IDSs) area unit should to decrease the intense influence of those attacks.

Intrusion Detection System is outlined because the system or code tool to notice unauthorized access to a network or ADPS. IDS is capable of detection every type attack like malicious, harmful attack, vulnerability, information driven attacks, host based attacks as an example privilege violation, sensitive file access, unauthorized logins and malwares. Then want IDS once have firewall as a result of the networks having firewall weren't designed to notice attack at network layer and application layer like worms, viruses, Denial of services (DoS), distributed denial of services (DDoS) and Trojans. The work of firewall is to prevent external traffic from coming into within the internal network. The intrusions area unit like viruses, worms, Trojans, or network attacks like unauthorized login, access of sensitive files, or information driven attacks on application. The intrusion violates the integrity, confidentiality and availableness. Because of this technique is unable to retort or access is denied. Therefore intrusion detection suggests that detection of unauthorized use of system or AN attack on a system or network. The Intrusion detection system (IDS) could be a hardware or computer code tool to detect these activities. Feature choice could be a technique for eliminating inapplicable and redundant options and choosing the most optimum set of options that manufacture a far better characterization of patterns happiness to completely different categories. Methods for feature choice square measure usually classified into filter and wrapper ways.

This paper specializes in filter methods for IDS. Motivation:

1. The Detection accuracy of anomaly system ought to maximize.
2. The detector generation time ought to less.

2. BACKGROUND

Early development started with understanding the attacks and their identities. System to trace occurrences of these identities was the order of the day. This gave rise to the classic signature based intrusion detection systems [3]. Though robust these systems lack the potential to handle novel attacks [4]. Data mining techniques are often used to make signature matching more efficient. With increasing complexity of the networks to state the least the increasing protocol layers, the attackers have modeled the attacks to exploit the loop holes in this complex system. Hence there is a need to generalize the Intrusion Detection process. Newer models to detect intrusion have become prevalent replacing mechanisms which fish for specific character set to find attacks. This lead to the rise of another class of IDSs which uses anomaly in the traffic characteristics for intrusion detection. Of the several models available one model describes the use of protocols, time based analysis for building the model. This is coupled with algorithm for learning from the conditional rules [5]. Another model involves the combination of signature based detection for known attacks and anomaly based detection for new attacks forming a hybrid intrusion detection system. They illustrate the use of fuzzy data mining techniques for anomaly based detection [6]. Another approach is the application of Genetic algorithm, a machine learning tool, in intrusion detection systems. The use of

Genetic Algorithm and Decision Trees to automatically generate rules for classifying the network connections [7].

3. DATASET FOR EVALUATION, TRAINING AND TESTING

Application of machine learning techniques involves use of a good dataset and extraction of relevant features from the dataset. Many of such research based dataset often come in handy but are never up to the mark due various reasons ranging from the out datedness of the dataset (hence the attacks which are present in it) to the inadequate spread of the incident attacks in the dataset. Keeping these in mind researchers tend to generate specific sets involving highly specific attacks and thus use them in evaluation of IDSs. In the forthcoming method used for Intrusion Detection, we have used the standard KDD CUP99 Intrusion detection and evaluation dataset (98–99). In the KDD CUP99 IDS evaluation dataset, all the network traffic including the entire payload of each packet was recorded in TCP dump format and provided for evaluation. The test network consisted of a mix of real and simulated machines; background traffic was artificially generated by the real and simulated machines while the attacks were carried out against the real machines. Classification of the attacks into four main classes namely, Denial of Service (DoS), Remote to Local (R2L), User to Remote (U2R) and the Data attacks/surveillance and probing [11].

4. EXISTING SYSTEM

The current network traffic data, which are often huge in size, present a major challenge to IDSs These "big data" slow down the entire detection process and may lead to unsatisfactory classification accuracy due to the computational difficulties in handling such data. Classifying a huge amount of data usually causes many mathematical difficulties which then lead to higher computational complexity. As a well-known intrusion evaluation dataset, KDD Cup 99 dataset is a typical example of large-scale datasets. This dataset consists of more than five million of training samples and two million of testing samples respectively. Such a large scale dataset retards the building and testing processes of aclassifier, or makes the classifier unable to perform due to system failures caused by insufficient memory. Furthermore,large-scale datasets usually contain noisy, redundant, or uninformative features which present critical challenges to knowledge discovery and data modeling.

Disadvantages:

1. Computer systems and internet have become a major part of the critical system. The current network traffic data, which are often huge in size, present a major challenge to IDSs.
2. These "big data" slow down the entire detection process and may lead to unsatisfactory classification accuracy due to the computational difficulties in handling such data.
3. Classifying a huge amount of data usually causes many mathematical difficulties which then lead to higher computational complexity.

5. PROPOSED SYSTEM

We Proposed an Intrusion detection system (IDS) is computer code or hardware that

monitors for intrusions and anomalies from the environment it's set to protect. Normally the IDS could be a security observance tool sort of a firewall that tries to find and possibly stop malicious activity. The Proposed LS-SVM based IDS a Knowledge primarily based IDS is applied on outcome of Filter Based Feature selection Algorithm i.e..This Feature Selection Algorithm Selects the most relevant feature of data, on this features we apply LS-SVM IDS for Attack Recognition. Knowledge primarily based IDS monitors a system victimization pattern of famed intrusions. The framework of the planned intrusion detection system is comprised of 4 main phases:

(1) Knowledge assortment, wherever sequences of network packets square measure collected,

(2) knowledge preprocessing, wherever

coaching and check knowledge square measure preprocessed and necessary options that may distinguish one class from the others square measure elite,

(3) Classifier coaching, wherever the model for classification is trained victimization LS-SVM, and

(4) Attack recognition, wherever the trained classifier is employed to find intrusions on the check knowledge.

The fig.1 shows the proposed system architecture that describes how the system works below:

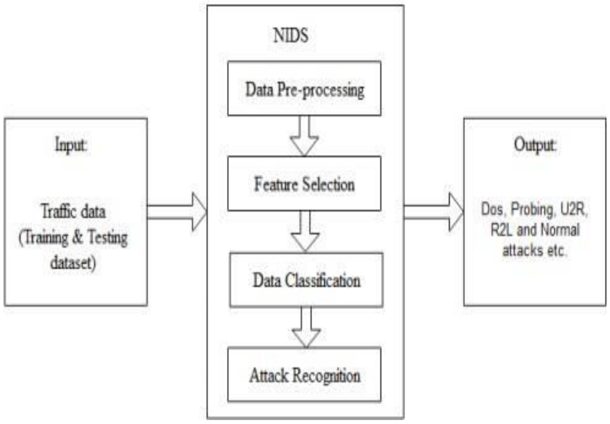


Fig 1.Proposed System Architecture

5.1] Data Collection:

Data assortment is that the 1st and a vital step to intrusion detection. The kind of information of supply and also the location wherever data is collected from square measure 2 determinate factors within the style and also the effectiveness of IDS. to produce the most effective suited protection for the targeted host or networks, this study proposes network-based IDS to check the planned approaches. The planned IDS run on the closest router to the victim(s) and monitor the incoming network traffic. During the training stage, the collected information samples square measure classified with relation to the transport/Internet layer protocols and square measure labeled against the domain data. However, the information collected within the take a look at stage square measure classified per the protocol varieties solely.

5.2] Data Preprocessing

The data obtained throughout the part of information assortment square measure 1st processed to get the essential options like those in KDD Cup ninety nine dataset.

5.3] Feature choice

Even though each affiliation during a dataset is diagrammatic by varied options, not all of those options square measure required to build IDS. Therefore, it's vital to spot the foremost informative options of traffic information to realize higher performance. The versatile Mutual data based mostly Feature choice (FMIFS) is meant to cut back the amount of features. The versatile linear coefficient of correlation based mostly Feature choice is meant to pick a feature that maximizes correlation and to eliminate moot and redundant options. However, the planned feature choice algorithms will solely rank options in terms of their connectedness however they can not reveal the most effective range of options that square measure needed to coach a classifier. To do so, the technique 1st utilizes the planned feature choice algorithmic program to rank all features supported their importance to the classification processes.

5.4] Classification

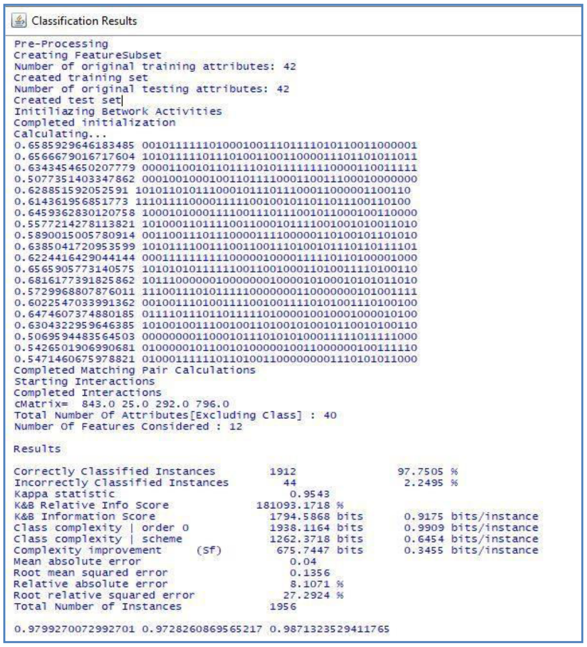
The optimum set of options is chosen, for all categories, five LS-SVM classifiers got to be applied on testing dataset. Each categoryifier distinguishes one class of records from the others. As an example the categoryifier of traditional class distinguishes traditional information from non-Normal (All varieties of attacks). The DoS category distinguishes DoS traffic from nonDoS information (including traditional, Probe, R2L and U2R instances) so on. The five LS-SVM classifiers square measure then combined to make the intrusion detection model completely differentiate to

tell apart} all different categories. The planned algorithmic program liquidator is applied on testing dataset to cut back the time of intrusion classification.

Advantages:

1. The experimental results square measure consistent and cozy.
2. It doesn't contain duplicate information within the coaching and testing set. So the results square measure additional correct than the strategy which has higher detection rates victimization frequent records.
3. The intrusion classification rates vary per completely different machine learning algorithmic program. Therefore it's helpful for efficient and correct detection of various learning technique

6. RESULTS



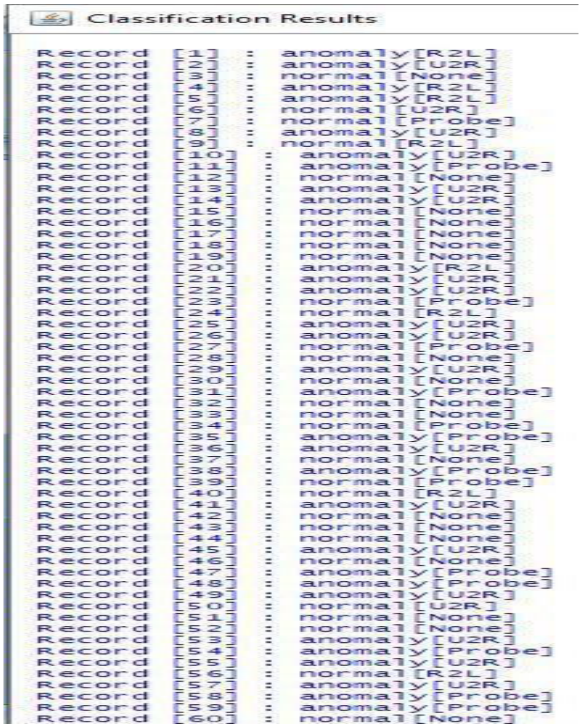
```
Pre-Processing
Creating FeatureSubset
Number of original training attributes: 42
Created training set
Number of original testing attributes: 42
Created test set
Initiliazing Betwork Activities
Completed initialization
Calculating...
0.6585929646183485 0010111111010001001110111101011001100001
0.6566679016717604 1010111110111010011001100001110110101011011
0.634345465020777 9 000011001011011110101111111100001100111111
0.5077351403347862 000100100010011011110001100111000100000000
0.628851592052591 1010110101110001011101110001100000110011 0
0.614361956851773 111011110000111110010010110110111001101 00
0.6459362830120758 1000101000111100111011100101100010011 0000
0.5577214278113821 1010001101111001100010111100100101001 1010
0.5890015005780914 0011001110111000011110000011010010110101 0
0.6385041720953599 1010111100111001100111010010111011011 1101
0.6224416429044144 0001111111111100000100001111101101000010 00
0.6565905773140575 1010101011111100110010001101001111010011 0
0.6816177391825862 1011100000010000000100001010001010101101 0
0.5729968807876011 1110011101011111000000011000000001010011 11
0.6022547033991362 0010011101001111001001111010100111010010 0
0.6474607374880185 0111101110110111110100001001000100001010 0
0.6304322959646385 1010010011100100110100101001011001010011 0
0.5069594483564503 0000000011000101110101010001111101111100 0
0.5426501906990681 0100000101100101000001001100000001001111 10
0.5471460675978821 0100011111101101001100000000111010101100 0
Completed Matching Pair Calculations
Starting Interactions
Completed Interactions
cMatrix=  843.0 25.0 292.0 796.0
Total Number Of Attributes[Excluding Class] : 40
Number Of Features Considered : 12

Results

Correctly Classified Instances        1912               97.7505 %
Incorrectly Classified Instances        44                2.2495 %
Kappa statistic                          0.9543
K&B Relative Info Score             181093.1718 %
K&B Information Score                 1794.5868 bits      0.9175 bits/instance
Class complexity | order 0            1938.1164 bits      0.9909 bits/instance
Class complexity | scheme             1262.3718 bits      0.6454 bits/instance
Complexity improvement     (Sf)        675.7447 bits      0.3455 bits/instance
Mean absolute error                      0.04
Root mean squared error                  0.1356
Relative absolute error                  8.1071 %
Root relative squared error             27.2924 %
Total Number of Instances             1956

0.9799270072992701 0.9728260869565217 0.9871323529411765
```

Fig1. Statistical Analysis of Data: Normal or Anomaly



```
Record  [1]  :  anomaly[R2L]
Record  [2]  :  anomaly[U2R]
Record  [3]  :  normal[None]
Record  [4]  :  anomaly[R2L]
Record  [5]  :  anomaly[R2L]
Record  [6]  :  normal[U2R]
Record  [7]  :  normal[Probe]
Record  [8]  :  anomaly[U2R]
Record  [9]  :  normal[R2L]
Record  [10]  :  anomaly[U2R]
Record  [11]  :  anomaly[Probe]
Record  [12]  :  normal[None]
Record  [13]  :  anomaly[U2R]
Record  [14]  :  anomaly[U2R]
Record  [15]  :  normal[None]
Record  [16]  :  normal[None]
Record  [17]  :  normal[None]
Record  [18]  :  normal[None]
Record  [19]  :  normal[None]
Record  [20]  :  anomaly[R2L]
Record  [21]  :  anomaly[U2R]
Record  [22]  :  anomaly[U2R]
Record  [23]  :  normal[Probe]
Record  [24]  :  normal[R2L]
Record  [25]  :  anomaly[U2R]
Record  [26]  :  anomaly[U2R]
Record  [27]  :  normal[Probe]
Record  [28]  :  normal[None]
Record  [29]  :  anomaly[U2R]
Record  [30]  :  normal[None]
Record  [31]  :  anomaly[Probe]
Record  [32]  :  normal[None]
Record  [33]  :  normal[None]
Record  [34]  :  normal[Probe]
Record  [35]  :  anomaly[Probe]
Record  [36]  :  anomaly[U2R]
Record  [37]  :  normal[None]
Record  [38]  :  anomaly[Probe]
Record  [39]  :  normal[Probe]
Record  [40]  :  normal[R2L]
Record  [41]  :  anomaly[U2R]
Record  [42]  :  normal[None]
Record  [43]  :  normal[None]
Record  [44]  :  normal[None]
Record  [45]  :  anomaly[U2R]
Record  [46]  :  normal[None]
Record  [47]  :  anomaly[Probe]
Record  [48]  :  anomaly[Probe]
Record  [49]  :  anomaly[U2R]
Record  [50]  :  normal[U2R]
Record  [51]  :  normal[None]
Record  [52]  :  normal[None]
Record  [53]  :  anomaly[U2R]
Record  [54]  :  anomaly[Probe]
Record  [55]  :  anomaly[U2R]
Record  [56]  :  normal[R2L]
Record  [57]  :  anomaly[U2R]
Record  [58]  :  anomaly[Probe]
Record  [59]  :  anomaly[Probe]
Record  [60]  :  normal[None]
```

Fig2. Attack Recognition: Type of Attack (Dos, U2r, R2l, Probing)

7. CONCLUSION AND FUTUURE WORK

The proposed feature selection algorithm is computationally efficient when it's applied to the LSSVM-IDS. The building (training) and test times consumed by the detection model using HFSA compared with the detection model using all features is less? A supervised filter-based feature selection algorithm has been proposed, namely HFSA. HFSA is an improvement over MIFS and MMIFS. FMIFS suggests a modification to Battiti’s algorithm to scale back the redundancy among features. The proposed LSSVM-IDS + HFSA has shown comparable results with other state-of the-art approaches when using the Corrected Labels sub dataset of the KDD dataset and tested on Normal, DoS, and Probe classes; it outperforms other detection models when tested on U2R and R2L classes. In future scope, the system is going to be implemented using another standard dataset either real time or non-real time. By reducing the offline time interval overhead, the web time interval is going to be minimized.

8. REFERENCES

[1] M. Roesch. Snort – Lightweight Intrusion Detection for Networks. Proceedings of USENIX LISA’99, November 1999.

[2] Cristianini, Nello, ShaweTaylor, John; An Introduction to Support Vector Machines and other kernel-based learningmethods, Cambridge University Press, 2000.

[3] Nitin, Mattord, Verma. Principles of Information Security. Course Technology. pp. 290–301, 2008.

[4] Anderson, Ross. Security Engineering: A Guide to Building Dependable Distributed Systems. New York: John ley & Sons. pp. 387–388, 2001.

[5] M. Mahoney, A Machine Learning Approach to Detecting Attacks by Identifying Anomalies in Network Traffic, Ph.D Dissertation, Florida Institute of Technology, 2003.

[6] S. Bridges and R. Vaughn, Fuzzy data mining and genetic algorithms applied to intrusion detection, Proceedings twenty third National Information Security Conference, October 1–19, 2000.

[7] M. Glickman, Balthrop and S. Forrest. A machine learning evaluation of an artificial immune system. Evolutionary Computation, 13(2):179–212, 2005.

[8] Sara Matzner Chris Sinclair, Lyn Pierce, An application of machine learning to network intrusion detection in Proceedings of the 15th Annual Computer Security Applications Conference, pages 371–377, Phoenix, AZ, 1999.

[9] T. Lane and C. Brodley, Temporal sequence learning and data reduction for anomaly detection. ACM

[10] Transactions on Information and System Security, 2(3), August 1999.

[11] T. Pietraszek and A. Tanner Data mining and machine learning towards reducing false positives in intrusion detection. Inform
Secur Tech Rep; 10(3):169–83, 2005.

[12] R. Lippmann, et al. The DARPA Off-Line Intrusion Detection Evaluation, Computer Networks 34(4) 579–595, 2000.

[13] V. Vapnic. The Nature of Statistical Learning Theory, Springer, New York, 1995.

[14] C. Cortes and V. Vapnik, Support-vector network, Machine Learning, vol. 20, pp. 273–297, 1995.

[15] C. C. Chang and C.-J. Lin, LIBSVM: A Library for Support Vector Machines 2001.

[16] DARPA intrusion detection evaluation, http://www.ll.mit.edu/IST/ideval/data/data index.html

[17] KDD Classifier Learning Contest http://cseweb.ucsdedu/∼elkan/clresultshtml, 1999.

[18] C. Thomas and N. Balakrishnan, Usefulness of DARPA data set in Intrusion Detection System evaluation,

[19] Proceedings of SPIE International Defense and Security Symposium, 2008.