ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

Data Mining Model For Network Intrusion Detection Using Boyer-Moore Algorithm

[I]Kollu Vijaya Kumar, [II]B Veerendranath
[I]M.Tech Student, [II]Asst. Professor
[I,II]Dept. of CSE, Kakinada Institute of Technology and Science, Divili, A.P. India

Abstract
Nowadays, as information systems are more open to the Internet, the importance of protected networks is largely developed. New intelligent Intrusion Detection Systems which are based on sophisticated algorithms rather than popular signature based detections are used frequently in well developed applications. Network based intrusion detection system monitor network activities. Intrusion detection is the process of monitoring the events occurring in a computer system or network and analyzing them for signs of possible incidents, which are violations or imminent threats of violation of computer security policies, acceptable use policies. In data mining-based intrusion detection system, we should make use of specific domain knowledge in relation to intrusion detection in order to effectively extract relative rules from large amounts of records. Intrusion detection systems (IDS) are primarily focused on identifying possible incidents, logging information about them, and reporting them to security administrators. Knowledge-Discovery is one of the hot topics in the field of knowledge extraction from database. Data mining is used to automatically learn patterns from large quantities of data mining can efficiently discover useful and interesting rules from large collection of data. This paper proposes new ensemble approach for Boyer-Moore Algorithm. Detrimental results how's better results for detecting intrusions as analyze to others existing methods.

Keywords
Knowledge Discovery, BOYER-MOORE Algorithm, Data Mining for Decision Tree based, Ensemble approach, Network Intrusion Detection System, Denial of Service.

I. Introduction

Being widely used and rapidly developed in recent years, network technologies have provided us with new life and shopping experiences, especially in the fields of business-learning and e-money. But along with network development, there has come a huge increase in network crime. It not only greatly affects our everyday life, high believes heavily on networks and Internet technologies, but also damages computer systems that serve our daily activities, including business, learning, and entertainment and so on. Besides of this internal hacking is difficult to detect because firewalls and Intrusion Detection Systems usually only defend against outside attacks. Intrusion Detection Systems an essential detection used as a counter measure to maintain data integrity and system availability from attacks. Intrusion Detection Systems (IDS) is a combination of software and hardware that set about to perform intrusion detection.

Intrusion detection is a process of colleting intrusion related knowledge happening in the process of observing the events and balancing them for sign or intrusion. It raises the alarm when a executable intrusion pass in the system. The network data obtain of intrusion detection consists of macro amount of textual information, which is delicate to comprehend and analyze. Many IDS can be described fundamental functional components. Information Obtain, Analysis, and Response. Different obtains of information and events based on

information are collected to decide whether intrusion has taken place. This information is collected at various levels like system, host, application, etc. Based on analysis of this data, we can cite the intrusion based on two general practices. Misuse detection and Anomaly detection. Issue detection is based on extensive knowledge of patterns associated with known attacks provided by human experts. Pattern matching, data mining, and state transition analysis are some of the approaches for Misuse detection and Anomaly detection is based on profiles that represent normal behaviour of users, hosts networks, and detecting struggle of significant deviation these profiles. Statistical Methods, expert system are some of the methods for intrusion detection based on Anomaly detection.

The primary desire behind using intrusion detection in data mining [5, 10, 12, 13] is automation .guide of the natural nature and pattern of the intrusion can be computed using data mining. To apply data mining techniques in intrusion detection, first, the collected monitoring data needs to be pre-processed and altered to the format suitable for mining processing. Next, the reformatted data will be used to develop a clustering or classification model. The classification model can be rule-based, decision-tree based, association-rule based, Bayesian-network based, or neural network based. Intrusion Detection mechanism based on IDS are not only automated but so provide for a significantly elevated carefulness and productivity. Unlike manual techniques, Data Mining ensures that no intrusion will be missed while checking real time records on the network. Credibility is essential in every system.IDS are now becoming essential part of our security system, and its credibility also adds value to the whole system Data mining techniques can be applied to gain insightful knowledge of intrusion prevention mechanisms. They can help detect new vulnerabilities and intrusions, discover previous unknown patterns of attacker natures, and provide decision support for intrusion management. The proposed paper organized as, Section2 explains about data mining. Section3 introduces methodology used in this paper. Experiment and result included in Section4 with concluding conclusion in section 5.

II. Data Mining

Data mining, also called Knowledge-Discovery and Data Mining, is one of the hot topic in the field of knowledge extraction from database. Data mining is used to automatically learn patterns from large quantities of data Mining can efficiently discover useful and Interesting rules from large collection of data. It is a fairly recent

topic in computer science but utilizes many older computational techniques from statistics, information retrieval, machine learning and pattern recognition. Data mining is disciplines works to finds the major relations between collections of data and enables to is cover a new and anomalies nature. Data mining based on intrusion detection techniques practically fall into one of two categories; misuse detection and anomaly detection In misuse detection, each instance in a data set is labelled as 'normal 'or 'intrusion' and a learning algorithm is trained over the labelled data. These techniques are able to automatically retrain intrusion detection models on different input data that include new types of attacks, as long as they have been labelled appropriately Data mining are used in different field such as marketing, financial affairs and business organizations in general and proof it is success.

The main approaches of data mining that are used including classification which maps a data item into one of several predefined categories. This approach normally output "classifiers" has ability to classify new data in the future, for example, in the form of decision trees or rules. An ideal application in intrusion detection will be together sufficient "normal" and "abnormal" audit data for a user or a program. The second essential approach is Clustering which maps data items into groups according to similarity or distance between them. Anomaly detection techniques thus identify new types of intrusions as deviations from normal usage [7,8]. In statistics-based outlier detection techniques [4]the data points are modelled using a stochastic distribution and points are determined to be outliers depending upon their relationship with in model. However ,with increasing dimensionality, it becomes increasingly difficult and in-accurate to estimate the multidimensional distributions of the data points[1]. However, recent outlier detection algorithms that we utilize in this study are based on computing the full dimensional distances of the points from one another[9,16] as well as on computing the densities of local neighbourhoods [6].Classifier construction is another essential research challenge to build efficient IDS.

Nowadays, many data mining algorithms have become very popular for classifying intrusion detection datasets such as decision tree, Naïve Bayesian classifier, neural network, genetic algorithm, and support vector machine etc. However, the classification truthfulness of most existing data mining algorithms needs to be improved, because it is very difficult to detect several new attacks, as the attackers are continuously changing their attack patterns Anomaly network intrusion detection models are now using to detect new attacks but the false positives are usually very high. The performance of an intrusion detection model depends on its detection rates and false positives .Ensemble approaches [14] have the advantage that they can be made to adopt the changes in the stream more accurately than single model techniques. Several ensemble approaches have been proposed for classification of evolving data streams. Ensemble classification technique is advantageous over single classification method. It is combination of several base models and it is used for continuous learning. Ensemble classifier has better sharpness over single classification technique. Bagging and boosting are two of the most well-known ensemble learning methods due to their theoretical performance guarantees and strong experimental results. Boosting has attracted much attention in the machine learning community as well as in statistics mainly because of its excellent performance and computational attractiveness for large datasets.

III. Methodology

This proposed model uses The Boyer-Moore Algorithm i.e. Decision tree based classification techniques to increase performance of the intrusion detection system.

The Boyer-Moore Algorithm:-Although the above algorithm quite knows, it doesn't help that much unless the strings you are exploring involve allot of come again at terns. It'll still depend upon you to go all along the document (s1) to be searched in. For too text editor type applications, the average case complexity is petty better than the naive algorithm (O(N), where N is the length of s1.)(The worst case for the KMP is N+M comparisons - much better than naive, so it's useful in certain cases). The Boyer- Moore algorithm is significantly better, and works by searching the target string s2 from right to left, while moving it left to right along s1.

The following example illustrates the general idea:

'the caterpillar' Match fails:
'pill' There's no space (' ') in the search string, so move it
^ right along 4 places
'the caterpillar' Match fails. There's no e either, so move along 4
'pill'
^
'the caterpillar' 'l' matches, so continue trying to match right to left
'pill'
^
'the caterpillar' Match fails. But there's an 'i' in 'pill' so move along
'pill' to position where the 'i's line up.
^
'the caterpillar' Matches, as do all the rest..
'pill'

The Boyer-Moore algorithm is greatly better, and works by exploring the target strings two from right to left, while affecting it left to right along s1. This still only wants knowledge of the second string, but we have need an array following an indication, for each advisable character that may appear, where it chance in the search string and hence how much to move along. So, index['p']=0, index['i']=1, index['l'] = 3 (index the rightmost 'l' where repetitions) but index['r']=-1 (let the value be -1 for all characters not in the string). When a union fails at a position i in the document, at a character C we move along the find string to a position where the present character in the document is above the index[C]th character in the string (which we know is a C), and start coordinating again at the right hand end of the string. (This is only achieved when this really results in the string being moved right, otherwise the string is just moved up another place, and the find started again from the right hand end) The Boyer-Moore algorithm in fact combines this process of skipping over characters with a method similar to the KMP algorithm (useful to improve efficiency after you have partially doubled a string). However, we'll just assume the normal version that skips based on the environment of a character in the find string.

4. Experimental Results

The proposed The Boyer-Moore algorithm is tested on KDDCup'99 dataset and compared to that of a Naïve Bayer's, KNN, eClass0 [1], eClass1 [1] and the Winner (KDDCup'99).

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

Survey of Anomaly Detection There are common types of two attacks in network intrusion detection (NDS):
The attacks that involve single connections and the attacks that involve multiple connections (bursts of connections). The standard metrics in Table 1 treat all types of attacks similarly thus failing to provide sufficiently generic and systematic evaluation for the attacks that involve many network connections.

Table 1: Metrix for Evaluation of Intrusion Detection

Confusion Matrix		Predicted Class	
		Normal	Intrusion / Attack
Actual Class	Normal	True Negative	False Positive
	Intrusion / Attack	False Negative	Correctly Detected

Interleaved Test-Then-Train

In this method each individual example can be used to test the model before it is used for principles and from this the definiteness scan be incrementally updated. The intension behind using this method is that, the model is always being tested on examples it has not seen. The advantage over holdout method being that holdout set is not needed for testing and ensures a smooth plot of carefulness over time as each individual example will become increasingly less significant to the overall average's, Assessment on KDDCup'99 Data Set.
The experiment is set up on a intrusion detection real data stream which has been used in the Knowledge Discovery and Data Mining (KDD) 1999 Cup competition.InKDD99 dataset the input data flow hold the implements of the network connections, such as protocol type, connection duration, login type etc. Each data sample in KDD99 dataset corresponds attribute value of a class in the network data flow, and each class is labelled either as normal or as an attack with exactly one specific attack type. In total, 41 features have been used inKDD99 dataset and each connection can be categorized into five main classes as one normal class and four main intrusion classes as DOS, U2R, R2L and Probe. There are 22 different types of attacks that are grouped into the four main types of attacks DOS, U2R, R2L and Probe tabulated in Tables.
The experimental putting is for the KDD99 Cup, filling 10% of the whole real raw data stream (494021 data samples) and 12 features are selected as per proposed algorithm.
Figures 1(a) -1(c) show graphical comparison of `Boyar-Moore algorithm with the Winner (KDDCup'99), eClass0, eClass1, KNN, C4.5 and Naïve Bays in terms of closeness or detection rate.

Table 2: Types of Attacks

Main Attack Classes	22 different attack types
DOS- Denial of Service	back, land, neptune, pod, smurf, teardrop
U2R- User to Root	buffer_overflow, loadmodule, perl, rootkit
R2L- Remote to User	ftp_write, guess_password, imap, multihop, phf, spy
Probe	ipsweep, nmap, portsweep, satan

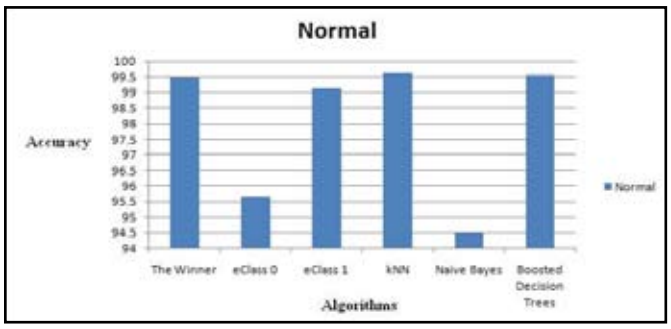

Fig. (a) .Normal with 41 features

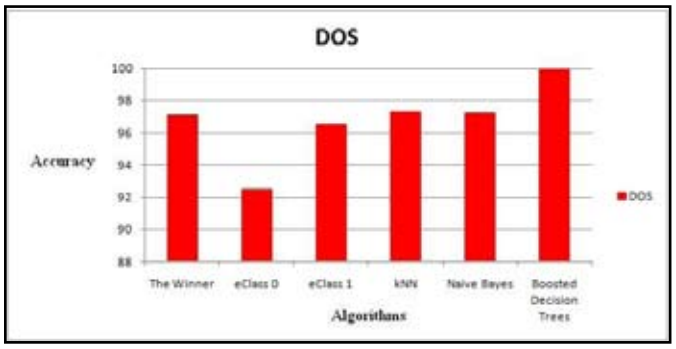

Fig.: (b) DOS attack with 41 features

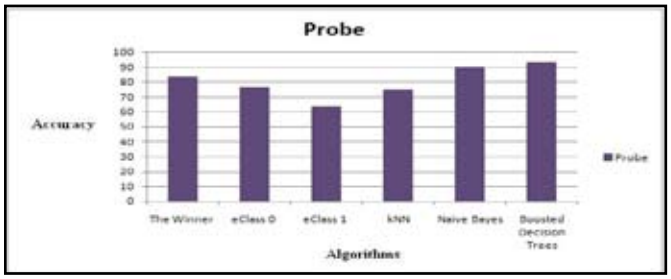

Fig.: (c) Probe attack with 41 features

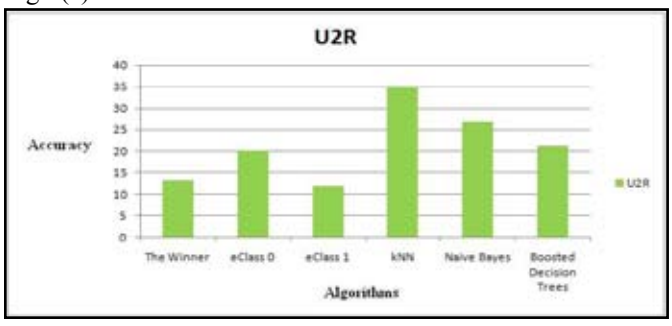

Fig.: (d) U2R attack with 41 features

V. Conclusion

This paper introduced a network intrusion detection model using The Boyer-Moore: a learning technique that allows combining several decision trees to form a classifier which is obtained from a weighted majority vote of the classifications given by individual trees.
The observation closeness of the Boyer-Moore has compared with Naive Bayesian, KNN, eClass0, eClass1 and the WinnerKDDCup'99). Boyer-Moore algorithm outperformed the compared algorithms on real world intrusion dataset, KDDCup'99. On the basis of these results, it can be concluded that The Boyer-Moore Algorithm may be a competitive alternative to these techniques in intrusion detection system.

VI. AcknowledgEment

The authors would like to thank the Management, Principal Dr. M.V. Ramakrishna Rao of Kakinada Institute of Technology and Science, Divili for their moral support.

References

[1] S. Kumar, "Classification and detection of computer *intrusions",Ph.D. thesis, Purdue Univ., West Lafayette, IN, 1995.*

[2] D. E. Denning, "An Intrusion Detection Model", IEEE Trans-actions on Software Engineering. SE13, pp. 222-232, 1987.

[3] R.P.Lippmann,D.JFried,IGraf,J.W.Haines,K. P.Kendall,D.McClung,D.Weber,S.E.Webster,. Wyschogrod,R.K.Cunningham, M.A.Zissman, "Evaluating IntrusionDetectionSystems:The1998DARPAOffline Intrusion Detection Evaluation. Proceedings. DARPA Information Survivability Conference and Exposition (DISCEX) 2000", Vol 2, pp. 12--26, IEEE Computer SocietyPress, Los Alamitos, CA, 2000.

[4] P. Nowak, "System wykrywania włamań iinformowania oawariach serwisów internetowych", Master Thesis, Technical University of Lodz, July 2006.

[5] P. P. Angelo, X. Zhou, "Evolving fuzzy rule based Classifiers from data streams", IEEE Transaction on Fuzzy Systems, Vol 16, No. 6, pp. 1462-1475, 2008.

[6] E.Amoroso.Sieci:Wykrywanieintruzów .dawnictwo RM, 1998.

[7] M.Wójtowski,B.Sakowicz,P.Mazur, Kompleksowy system wysokiej dostępności", Mikroelektronika Informatyka,Łódź 2005,pp. 211-216,ISBN 83-922632-0-0.

[8] B.Sakowicz,J.Wojciechowski,K.Dura. Metody.budowania wielowarstwowych aplikacji lokalnych i rozproszonych w oparciuo technologic Java 2 Enterprise Edition", Mikroelektronika I Informatyka, maj 2004, KTMiI P.Ł. , pp. 163-168, ISBN 83-9192895-0.

[9] B.Foote, "Integrating Java with C + + ", JavaWorld.com, 1996

[10] Aggrawal, P. Yu, "Outlier Detection for High Dimensional Data",Proceedings of the ACM SIGMOD Conference, 2001.

[11] B. Caswell, J. Hewlett, "Snort users manual", 2003.

[12] R.Bane, N.Shivsharan, "Network intrusion detection system (NIDS)", pp. 1272-1277, 2008.

[13] V.Barnett, T. Lewis, "Outliers in StatisticalData", John Wiley and Sons, NY, 1994.

[14] R. G. Byrnes D. J. Barrett, R. E. Silverman, "Linux. Bezpieczeństwo. Receptury.", O'Reilly, 2003.