International Journal for Innovative Engineering and Management Research

A Peer Revieved Open Access International Journal

www.ijiemr.org

ELSEVIER

SSRN

Title: **PERFORMANCE EVALUATION OF INTRUSION DETECTION UTILIZING MACHINE LEARNING ALGORITHMS**

Volume 06, Issue 6, Pages: 268 - 277

Paper Authors

Suresh Kumar P.G.V , Shaheda Akthar

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

Performance Evaluation of Intrusion Detection Utilizing Machine Learning Algorithms

Suresh Kumar P.G.V [1], Shaheda Akthar[2]

[1] Research Scholar, Department of CSE, Acharya Nagarjuna University, Andhra Pradesh
[2] Lecturer, Department of Computer Science, Govt. College for Women, Guntur,
E-mail: pendemsuresh@gmail.com

Abstract:
There are numerous issues with conventional intrusion detection systems (IDS), for example, low recognition capacity against obscure organization assault, high bogus caution rate and lacking examination ability. Subsequently the significant extent of the examination in this area is to build up an interruption identification model with improved precision. In this paper, testing and famous NSL-KDD dataset for interruption identification is picked for performed tests, where grouping and four benchmark machine learning methods are utilized so as to decide ideal strategy for characterization space. This paper intends to group the NSL-KDD dataset regarding their metric information by utilizing the best four machine learning characterization calculations like Decision Tree, Naive Bayes, KNN and SVM to discover which calculation will have the option to offer all the more testing precision. NSL-KDD dataset has comprehended a portion of the inborn restrictions of the accessible KDD'99 dataset. The aftereffects of the led tests exhibit that Decision Tree performed viably in distinguishing assaults.

Keywords: *IDS, ML, Decision Tree, Naïve Bayes and KNN*

1. Introduction

Malicious intruders in the organization are expanding step by step because of the fast advancement of web. The intruders can get to, control and handicap the frameworks associated on the web. To shield different digital assaults and Computer infections, heaps of Computer security methods have been concentrated in a decade ago, which incorporate cryptography, firewalls and interruption identification framework (IDS) and so forth [4]. To characterize what an assault is, any activity which compromises the privacy, uprightness and accessibility is called an assault. The assaults by and large spotlight on the weaknesses of a client on the organization by unapproved admittance to a framework.

To forestall such dangers and assaults and identify any nosy exercises, security programming organizations built up an IDS. IDS take a shot at the rule that if the conduct of an ordinary client is unique, it may be an interruption endeavor. An interruption location framework utilizes a lot of strategies to recognize any dubious exercises on the organization level and host level.

This security component can be executed utilizing an IDSwhich can be portray as an assortment of programming or equipment gadget ready to gather, dissect and recognize any undesirable, dubious or malevolent traffic either on a specific Computer host or network [9].

The fundamental function of IDSs is basic since the organizations can be helpless against be assaulted by both inside and outer interlopers. The IDS has gotten one of the major segments of Computer security to distinguish these noxious dangers with the point of shielding frameworks from normal damages and gathering weaknesses [12]. In this way to accomplish its assignment, an IDS should utilize some factual or numerical strategy to peruse and decipher the data it gathers and accordingly reports any malignant action to the organization manager.

IDSare created to recognize unapproved endeavors to get to or control the PC frameworks. IDS has been grouped into two significant classes, in particular mark-based discovery and inconsistency-based location. In signature-based IDS, assault example of gatecrashers is displayed and the framework will inform once the match is recognized [13]. All realized assaults are related to diminished bogus positive rate. Mark information bases must be refreshed regularly in order to recognize the new assault design. Notwithstanding, inconsistency identification frameworks make a profile of ordinary movement. Any example that goes amiss from the ordinary profile is treated as an abnormality. Consequently, even obscure assault designs are distinguished with no manual mediation.

II Related Works

Numerous scientists have applied machine learning methods for the proficient plan of Network IDS. Machine Learning applications include millions or even billions of bits of information records. For instance, in the KDD Cup'99 dataset, there are in excess of 4 million and 3 million cases in the preparation set and test set, separately. Yet, a portion of the methods can't have any significant bearing on such bigger datasets because of the inadequate memory limit of the framework or time taken to complete the preparation. Here we had utilized the NSL-KDD dataset. NSL-KDD is an informational collection recommended to tackle a portion of the innate issues of the

KDD Cup'99 informational collection which are referenced in [10].

Herve Nkiama et al [3] proposed an element determination component which intends to wipe out non-important features just as distinguish the highlights which will add to improve the location rate, in light of the score each feature have set up during the choice cycle. To accomplish that objective, a recursive component disposal measure was utilized and connected with a choice tree put together classifier and later with respect to, the reasonable pertinent features were distinguished. This methodology was applied on the NSL-KDD dataset which is an improved rendition of the past KDD 1999 Dataset, scikit-discover that is an AI library written in python was utilized in this paper. Utilizing this methodology, important features were distinguished inside the dataset and the exactness rate was improved. These outcomes loan to help the possibility that features choice improve fundamentally the classifier execution.

L. M. Ibrahim et al. [7] utilized a solo ANN to build an IDS dependent on inconsistency identification. The framework utilized self-association map (SOM) ANNs for recognition and to recognize assault traffic from ordinary traffic.

Megha Aggarwal and Amrita [8] expressed that the Intrusion discovery framework manages enormous measure of information which contains different superfluous and repetitive features bringing about expanded preparing time and low recognition rate. In this way feature choice assumes a significant function in interruption discovery. They likewise proposed a relative investigation of various component choice strategies are introduced on KDDCUP'99 benchmark dataset and their exhibition are assessed regarding identification rate, root mean square blunder and computational time.

Mohammed A. Ambusaidi et al [9] proposed a common mutual information-based calculation that scientifically chooses the ideal component for characterization. This common mutual information-based component choice calculation can deal with straightly and nonlinearly subordinate information highlights. Its adequacy is assessed in the instances of organization interruption recognition. An Intrusion Detection System (IDS), named Least Square Support Vector Machine based IDS (LSSVM-IDS), is assembled utilizing the highlights chose by our proposed include determination calculation. The exhibition of LSSVM-IDS is assessed utilizing three interruption location assessment datasets, in particular KDD Cup 99, NSL-KDD and Kyoto 2006+ dataset.

Pavan Pongle and Chavan [11] propose a unified and conveyed engineering for a half and half IDS, which they executed dependent on reproduced situations and organizations. It centers around identifying directing assaults, for example, the wormhole assault.

III Machine Learning (ML)

IDS engineers utilize different methods for interruption discovery. One of these strategies depends on ML. ML strategies can foresee and distinguish dangers before they bring about significant security occurrences [1].

ML, a part of man-made reasoning, is a logical order worried about the plan and advancement of calculations that permit computers to advance practices dependent on observational information, for example, from sensor information or information bases. A significant focal point of ML research is to consequently figure out how to perceive complex examples and settle on astute choices dependent on information [5]. ML has a wide scope of uses, including web crawlers, clinical conclusion, text and penmanship acknowledgment, picture screening, load determining, promoting and deals finding, etc.

AI strategies can be utilized to discover and bring data by the methods for models which can't be distinguished effectively by human perception. These components are classifiers which characterize the organization information approaching into the framework to choose whether the movement is an assault or some ordinary action.

The model can be prescient to make expectations later on, or clear to pick up information from information. To play out a prescient or illustrative assignment, ML by and large utilize two primary methods: Classification and Clustering. In order, the program must foresee the most likely classification, class or name for novel perception into one or numerous predefined classes or name while grouping, the classes are not predefined during the learning cycle. Nonetheless if the reason for the IDS is to separate between typical or interruption traffic, arrangement is prescribed and in the event that we look to distinguish the kind of interruption, grouping can be more useful. To improve the interruption discovery framework and decrease the bogus negative and bogus positive, which can be tried by the utilization of various calculations. In this paper, Naive Bayes Classifier, Decision Tree Classifier, Random Forest Classifier, KNeighbours Classifier, Logistic Regression, SVM Classifier and Voting Classifiers are utilized for preparing information and testing it.

3.1 Decision Tree

Decision tree learning is one of the best procedures for administered arrangement learning. Decisiontrees are a straightforward recursive structure for communicating a successive characterization measure in which a case, depicted by a lot of qualities, is allotted to one of a disjoint arrangement of classes [2][5]. A Decision tree is a tree structure which groups an information test into one of its potential classes. Decision trees are utilized to separate information by settling on choice guidelines from the enormous measure of accessible data. A Decision tree classifier has a straightforward structure which can be minimally put away and that proficiently arranges new information.

Decision trees comprise of hubs and leaves. Every hub in the tree includes testing a specific quality and each leaf of the tree means a class. As a rule, the test contrasts a quality worth and a steady. Leaf hubs give a grouping that applies to all examples that arrive at the leaf, or a lot of characterizations, or a likelihood circulation over every conceivable arrangement. To group an obscure occurrence, it is directed down the tree as indicated by the estimations of the traits tried in progressive hubs, and when a leaf is reached, the case is characterized by the class allocated to the leaf.

3.2 Naive Bayes

Naive Bayes is one of the best and productive grouping calculations. NaiveBayes Classifier that is the probabilistic classifier dependent on the Bayes Theorem. Naive Bayes classifier expect that the impact of the qualities esteem on a given class is free on the estimation of different highlights [5]. The classifier just picks the mark with the most elevated likelihood, given the info highlights. The innocent segment of the classifier is that it accepts a solid autonomy between ascribes, basically it expects the probabilities for every one of the info highlights are autonomous of one another.

Leave H alone a theory and X is an information living in a specific C class. At that point P (H/X) is known as the back likelihood that communicates our certainty level on a speculation H after X information is given. P (H) speaks to the H earlier likelihood for all example information. P (H/X) is surely more enlightening than P (H). Bayes' hypothesis portrays the connection between P (H/X), P (H), and P (X) is appeared on condition 1 as follow:

$$P(H/X) = P(X/H) * P(H)/P(X)$$

3.3 Support Vector Machine (SVM)

The SVM is another sort of MLtechnique dependent on measurable learning hypothesis. Due to great advancement and a higher precision, SVM has become the examination focal point of the ML people group. SVMs are set of related administered learning strategies utilized for characterization and relapse [14]. A few late examinations have detailed that the SVM by and large are fit for conveying better as far as grouping exactness than the other information arrangement calculations. SVM is based on factual learning hypothesis by Vapnik et al proposed another learning technique, which is based on a set number of tests in the data contained in the current preparing text to get the best order results.

An exceptional property of SVM will be, SVM at the same time limit the experimental grouping blunder and boost the mathematical edge. So SVM called Maximum Margin Classifiers. SVM depends on the Structural danger Minimization. SVM map input vector to a higher dimensional space where a maximal isolating hyperplane is built. Two equal hyperplanes are built on each side of the hyperplane that different the information. The isolating hyperplane is the hyperplane that augment the separation between the two equal hyperplanes. A supposition that is made that the bigger the edge or separation between these equal hyperplanes the better the speculation mistake of the classifier [6].

3.4 K nearest neighbor (KNN)

KNN is a famous characterization calculation showing great execution qualities and a brief time of preparing time. KNN is straightforward, generally well known, profoundly proficient and compelling calculation for design acknowledgment. KNN is a straight forward classifier, where tests are ordered dependent on the class of their closest neighbor [5].

The KNN is a non-parametric order strategy, which is straightforward yet powerful much of the time [13]. For an information record d to be grouped, its K closest neighbors are recovered, and these structures an area of d. Lion's share casting a ballot among the information records in the area is normally used to choose the grouping for 'd' with or without thought of separation-based weighting. Be that as it may, to apply KNN we have to pick a suitable incentive for K, and the achievement of order is a lot of subject to this worth. One might say, the KNN strategy is one-sided by K. There are numerous methods of picking the K esteem, yet a basic one is to run the calculation ordinarily with various K esteems and pick the one with the best presentation.

IV Experimental Results

The target of this segment is to assess execution of four AI calculations regarding exactness, accuracy and review of the NSL-KDD informational collection, which is a reconsidered variant of KDD'99 informational index [10]. The purpose behind utilizing NSL-KDD dataset for our analyses is that the KDD'99 informational index has countless repetitive records in the preparation and testing informational collection. For paired grouping, the NSLKDD characterizes the organization traffic into two classes, specifically, ordinary and abnormality.

4.1 Dataset

The information comprises of traffic records to construct an Intrusion Detection System and a model which could anticipate if there is an assault or an interruption endeavor or is it a typical association. The dataset comprises of web traffic records caught by an interruption discovery framework. These are viewed as the inconspicuous records caught by a genuine interruption identification framework. There is an aggregate of 42 ascribes out of which 41 credits are the traffic information and the other one is assault class (typical or an assault) and the other property is the score which signified the seriousness of traffic information subtleties. The analyses were performed on full preparing informational collection having

125973 records and test informational collection having 22544 records.

For every perception in the NSL KDD dataset, there are 41 features,3 is ostensible, 4 are double and the staying 34 are constant factors. It has 23 traffic classes in the preparation dataset and 30 in the test dataset. These assaults can be bunched into four fundamental classifications DOS, examining, U2R and R2L. The highlights are characterized into 3 wide sorts 1) essential highlights, 2) content-based highlights and 3) traffic-based highlights. The assault data of the NLS-KDD dataset is recorded in Table-1.

Table-1: Details of NSL KDD data

Training Data (1,25,973)		**Testing Data (22,544)**	
Type of Attack	**Total No. of Instances**	**Type of Attack**	**Total No. of Instances**
Normal	67343	Normal	9711
DOS	45927	DOS	7456
Probe	11656	Probe	2421
R2L	52	R2L	200
U2R	995	U2R	2756

4.2 Performance Metrics

So as to approve the forecast consequences of the examination of the four well known information mining methods and the 10-crease hybrid approval is utilized. The k-overlap hybrid approval is normally used to decrease the blunder came about because of arbitrary examining in the correlation of the accuracies of various forecast models. The whole arrangement of information is arbitrarily separated into k folds with similar number of cases in each crease. The preparation and testing are performed for k times and one overlay is chosen for additional testing while the rest are chosen for additional preparation. The current investigation partitioned the information into 10 folds where 1 overlap was for trying and 9 folds were for preparing for the 10-crease hybrid approval. The preparation information is chosen from the entire dataset arbitrarily and straightforwardly took care of into the proposed mining approach.

Execution of every classifier is measure as far as disarray network, accuracy, precision and recall. These measurements are generally characterized for a parallel characterization task with positive and negative classes. That is:

Accuracy: Accuracy is a measure which decides the likelihood that how much outcomes are precisely grouped.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \qquad (1)$$

Precision: Precision speaks to how exact the classifier expectations are since it shows the measure of genuine positives that were anticipated out of all sure marks relegated to the occasions by the classifier. Exactness is the extent of positive expectations that are right

$$\text{Precision}= \frac{TP}{TP+FP} \qquad (2)$$

Recall: Recall is the extent of positive examples that are effectively anticipated positive. it shows the measure of really anticipated positive classes out of the measure of complete genuine positive classes.

$$\text{Recall}=\frac{TP}{TP+FN} \qquad (3)$$

Where,

- True positive (TP) = number of positive examples effectively anticipated.

- False negative (FN) = number of positive examples wrongly anticipated.

- False positive (FP) = number of negative examples wrongly anticipated as sure.

- True negative (TN) = number of negative examples accurately anticipated.

These qualities are frequently shown in a disarray network as be introduced in Table-2. Arrangement Matrix shows the recurrence of right and off base forecasts. It analyzes the real qualities in the test dataset with the anticipated qualities in the prepared model.

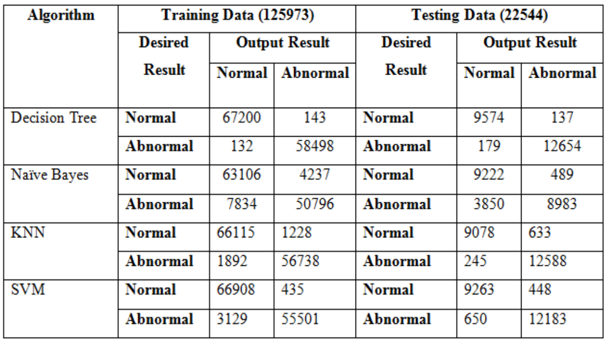
Table-2: Confusion Matrix of classification

		Predicted	
		Normal	**Abnormal**
Actual Class	**Normal**	TP	FN
	Abnormal	FP	TN

4.3 Results

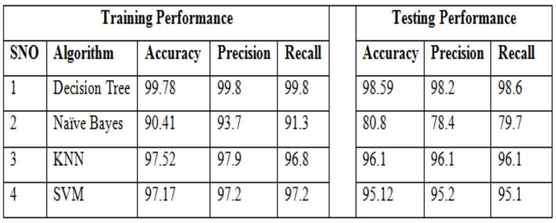
The disarray lattice of every Classification technique is introduced in Table-3; the qualities to gauge the exhibition of the strategies (for example accuracy, precision and recall) are gotten from the disarray network and appeared in Table-4 for preparing information and table-5 for testing information. From the above tables we locate that most elevated exactness of Classification model is Decision Tree (97.59%) in both preparing and testing information as appeared in figure-1and figure-2.

Table-3: Confusion Matrix of NSL KDD dataset



Algorithm	Training Data (125973)			Testing Data (22544)		
	Desired Result	Output Result		Desired Result	Output Result	
		Normal	Abnormal		Normal	Abnormal
Decision Tree	**Normal**	67200	143	**Normal**	9574	137
	Abnormal	132	58498	**Abnormal**	179	12654
Naïve Bayes	**Normal**	63106	4237	**Normal**	9222	489
	Abnormal	7834	50796	**Abnormal**	3850	8983
KNN	**Normal**	66115	1228	**Normal**	9078	633
	Abnormal	1892	56738	**Abnormal**	245	12588
SVM	**Normal**	66908	435	**Normal**	9263	448
	Abnormal	3129	55501	**Abnormal**	650	12183

Table-4: Training and Testing Performance of ML Algorithms



Training Performance						Testing Performance		
SNO	Algorithm	Accuracy	Precision	Recall		Accuracy	Precision	Recall
1	Decision Tree	99.78	99.8	99.8		98.59	98.2	98.6
2	Naïve Bayes	90.41	93.7	91.3		80.8	78.4	79.7
3	KNN	97.52	97.9	96.8		96.1	96.1	96.1
4	SVM	97.17	97.2	97.2		95.12	95.2	95.1

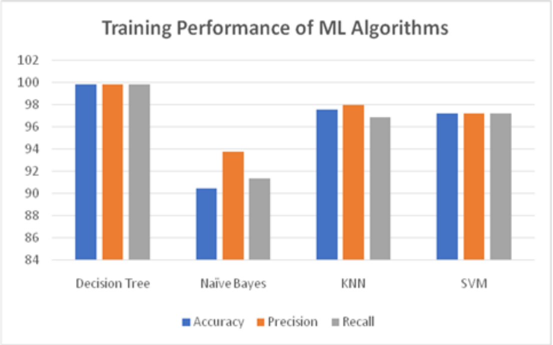


Figure-1: Training performance of ML algorithms

We evaluate our four models using different execution estimations like exactness, Precision and Recall, the Experimental results are showed up in the table-4 and table-5 and same showed up in the figure-1and figure-2. We find in the Figure-1 for our preparation information, the introduction of the Decision Tree estimation has accomplished 99.78% Accuracy, Naïve Bayes has 90.41%, KNN has accomplished 97.52% and SVM model has achieved 97.17%. We find in the Figure-2 for our testing information, the introduction of the Decision Tree count has accomplished 98.59% Accuracy, Naïve Bayes has 80.8%, KNN has accomplished 96.1% and SVM model has achieved 95.12%.

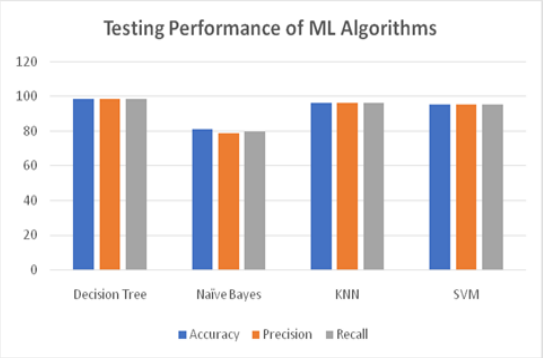


Figure-2: Testing Performance of ML Algorithms

As the result from assessment among the four figurings, we find that most vital precision of Classification model is Decision Tree for both preparing and testing (99.78% and 98.59). Precisely when veered from exactness and survey are also higher in the Decision Tree model when appeared differently in relation to other three models.

V. Conclusion

As organization assaults have expanded in number and seriousness in the course of recent years, IDS is progressively turning into a basic segment to make sure about the organization. Because of huge volumes of security review information just as intricate and dynamic properties of interruption practices, streamlining execution of IDS turns into a significant open issue that is accepting increasingly more consideration from the examination network. In this paper, four unique sorts of ML models were applied in particular Decision Tree, Naïve-Bayes, K Nearest Neighbors and Support Vector Machine for the interruption identification framework. The exhibition of all these ML models were watched and looked at dependent on changed standard assessment boundaries, for example, Accuracy, Precision and Recall of the test information. Our test has been done with four distinctive grouping calculations for the dataset and in that decision, tree shows a high exactness for both testing and preparing information contrasted with every other calculation. It was seen that the Decision Tree Classifier calculation performed in a way that is better than different models creating an exactness of 98.59%.

References

1. D. Hand, H. Mannila, P. Smyth.: Principles of Data Mining. The MIT Press, 2001
2. G. Ravi Kumar, Venkata Sheshanna Kongara & Dr. G. A. Ramachandra, "An

Efficient Ensemble Based Classification Techniques for Medical Diagnosis", International Journal of Latest Technology in Engineering, Management and Applied Sciences, Volume II, Issue VIII, PP: 5-9, ISSN-2278-2540,2013

3. Herve Nkiama, Syed Zainudeen Mohd Said and Muhammad Saidu "A Subset Feature Elimination Mechanism for Intrusion Detection System", International Journal of Advanced Computer Science and Applications, Vol. 7, No. 4, PP:148-157, 2016
4. H. J. Liao, C. H. R. Lin, Y.-C. Lin, and K.-Y. Tung, ‘‘Intrusion detection system: Acomprehensive review", J. Netw. Comput. Appl., vol. 36, no. 1, pp. 16–24, 2013.
5. J. Han and M. Kamber," Data Mining concepts and Techniques", the Morgan Kaufmann series in Data Management Systems, 2nd ed. San Mateo, CA; Morgan Kaufmann, 2006.
6. L. M. Ibrahim, D. T. Basheer, and M. S. Mahmod, ‘‘A comparison study for intrusion database (KDD99, NSL-KDD) based on self organization map (SOM) artificial neural network,’’ J. Eng. Sci. Technol., vol. 8, no. 1, pp. 107–119, 2013.
7. Megha Aggarwal and Amrita, "Performance Analysis of Different Feature Selection Methods in Intrusion Detection", International Journal of Scientific & Technology Research, Volume 2, ISSUE 6, ISSN 2277-8616, June 2013
8. Mohammed A. Ambusaidi, Xiangjian He, Priyadarsi Nanda and Zhiyuan Tan, "Building an Intrusion Detection System Using a Filter-Based Feature Selection Algorithm", IEEE Transactions on Computers, Volume:65, Issue:10, PP:2986 - 2998, Oct. 1 2016
9. M. P. K. Shelke, M. S. Sontakke, and A. D. Gawande, "Intrusion Detection System for Cloud Computing," Int. J. Sci. Technol. Res., vol. 1, no. 4, pp. 67–71, 2012.
10. "Nsl-kdd data set for network-based intrusion detection systems." Available on: http://nsl.cs.unb.ca/KDD/NSL-KDD.html, March 2009.
11. Pavan Pongle and Gurunath Chavan. Real time intrusion and wormhole attack detection in internet of things. International Journal of Computer Applications, 121(9), 2015.
12. S. Mukherjee and N. Sharma, "Intrusion Detection Using Naive Bayes Classifier with Feature Reduction," Procedia Technology, vol. 4, pp. 119–128,2012
13. Tsai, C.F.; Lin, C.Y. A triangle area based nearest neighbors' approach to intrusion detection. Pattern Recognit, 43, 222–229,2010
14. V. Vapnik. The Nature of Statistical Learning Theory. NY: Springer-Verlag. 1995.