International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 1 | ISSN : 2456-3307

Kernel Based Intrusion Detection Using Data Mining Techniques

Shivalingari Bhanu Sree

PG Scholar, Department of IT, VNR Vignan Jyothi Institute of Engineering and Technology, Hyderabad , TS, India

ABSTRACT

From the onset of internet arrangement, protection menaces normally recognized as intrusions has return to be a highly important and demanding issue in internet arrangements, knowledge and information system. In this system to overcome these menaces every time a detection arrangement was requested because of extreme development in networks. within the growth of the arrangement, attackers came to be stronger and every single amount compromises the protection of the network.Hence a requirement of the Intrusion Detection arrangement came to be a very important and important instrument in net security. Detection and hindrance of such aggressions loud intrusions typically depends on the talent and potency of Intrusion Detection Arrangement (IDS). In this apporch number of component has been directed for using the methods, these systems have their own advantages and deficiencies". Here mainly focusing on different classification methods.

Keywords : Intrusion Detection, Anomaly Detection, Misuse Detection, KDD Cup99.

I. INTRODUCTION

From past decades aboard fast progress within the web established data, new request spans for computer web have emerged. At the alike period, expansive vary progress within the LAN and WAN request spans in company, commercial, industry, protection and healthcare sectors made us extra reliant on the computer networks. All of these request spans made the web an associate appealing target for the mistreatment and huge vulnerability for the community. An interesting work to do or an attempt to realize action for a touch folks came to be a foul dream for the others. In unnumerable occurrence malicious deeds created this nightmare to come back to e a reality.

In addition to hacking, new entities like worms, Trojans and viruses gave extra panic into the net-worked society.As the present scenario could be a moderately new development,web armaments are weak.Though, due to the recognition of computer net, their property and however, manufacturing dependency on them, knowledge of the menace will have desecrating consequences. Safeguarding such an important groundwork has return to be the priority one scrutiny span for unnumerable researchers.

In this project,mainly focus on the research of the present course in Intrusion Detection Arrangements (IDS) and to look at a compact present issue that continues during this survey space. In analogy to a compact mature and well -stayed survey spans, IDS could be active earth of analysis. Though, due to its duty important nature, it's enticed significant attention regarding itself. Density of scrutiny on this subject is continually rising and everyday extra researchers are unit concerned during this earth of work. The menace of a replacement wave of cyber or net aggressions isn't simply a chance that need to be believed, still it's an accept undeniable fact that will emerge at anytime. The present trend for the IDS is

reserved from a reliable protecting arrangement, although the main believed is to create it probable to note novel web attacks.

One of the main concerns is to create positive that just in case of associate intrusion endeavor, the arrangement is in a position to note and to report it. when the detection is reliable, the subsequent pace should to be to guard the online (response). In additional, the IDS arrangement will be upgraded to an associate Intrusion Detection and Reply Arrangement (IDRS). Though, no segment of the IDS is presently at a completely reliable level. Even nevertheless, researchers are unit at the same time involved in working each detection and answer factions of the system. A main blow within the IDS is that the promise for the intrusion detection. A main issue in the IDS is to agreement for the intrusion detection.This is the cause why in numerous cases IDS employ the intrusion detection.During this methodology, IDS is really helping the online protection master and it is not valid enough to be sure in its own. The reason is the in- ability of IDS arrangements to notice the new or modified attack patterns. Even supposing the most recent method of the detection ways has enhanced the detection rate.However, there is an extended methodology to go.

There are 2 main ways in which for identifying intrusions, signature-based and anomaly-based intrusion detection. In the early method, attack outlines or the activities of the intruder is modeled(attack signature is modeled). Here the arrangement can gesture the intrusion when a match is detected. Though, within the subsequent neither method normal actions of the online is modeled. During this method, the arrangement can raise the alarm when the act of the online doesn't match aboard its traditional behavior. there's an added Intrusion Detection (ID) method that's loud specification-based intrusion detection. During this method traditional deeds (expected behavior) of the host is enumerated and after sculpturesque. . In this way, manage worth for protection, freedom of procedure for the host is limited. In this paper, these ways will be briefly debated and compared.

The believed of possessing associate entrant accessing the arrangement lacking even having the ability to note it's the worst nightmare for every net protection officer. As the present ID data isn't correct masses to enhance a reliable detection, heuristic methodologies can be a way out. As for the last line of protection, and so as to cut the quantity of unseen intrusions, heuristic ways like Honey Jars (HP) may be deployed. Honey jars may be installed on every arrangement and act as mislead or decoy for a resource.

Another main issue during this survey span is that the speed of detection. Computer webs have a dynamic nature in a truly sense that knowledge and knowledge within them are unceasingly changing. Therefore, noticing associate intrusion exactly and duly, the arrangement has got to add real time. Working in the real period isn't merely to per-form the location in real period; but is to change to the new elements within the system. Real period operating IDS is associate alert scrutiny span pursued by countless researchers. Most of the scrutiny works are unit aimed to familiarise the foremost amount effectual methodologies. The aim is to create the requested methods suitable for the real period implementation.

From a disparate outlook, 2 ways in which may be envisaged in requesting IDS. In this relation, IDS can be whichever host established or net based mostly. In the host established IDS, ordering will merely defend its own innate machine (its host). On the supplementary hand, within the net established IDS, the ID procedure is somehow distributed aboard the net- work. . In this way, whereas the agent established knowledge is extensively requested, a distributed arrangement will protect the web as a whole. In this design IDS can control or monitor web

firewalls, web routers or web switches as well as the client machines.

The paper focuses on the detection portion of the intrusion detection and reply drawback. Researchers have pursued disparate ways in which or a mixture of disparate ways in which to resolve this drawback. Every single way has its own theory and presumptions. This is so because their no precise activity ideal for the legitimate user, the intruder or the online itself.

II. RELATED WORK

S. Duque and Omar [2] proposed a K-Mean clustering on NSL-KDD dataset. "The calculation is connected on various five groups. The best outcomes are acquired when 22 bunches were utilized. Likewise K-Mean grouping is utilized as a part of mixture approaches", similar to B. Sharma and H. Gupta [3] utilizes two systems affiliation run and grouping. "Apriori and K-Mean is utilized to recognize the interruptions. The test is done on KDD'99 dataset. The CPU Utilization (74%),implementation measures are execution time (120ms) and memory use (54%)".

Ravale and Nilesh et al. [4] proposed half and half approach of K-Mean and RBF kernel capacity of SVM. "The exactness aftereffect of the half and half approach is 93% and identification rate is 95%.Where, Chao and Wen et al. [5] proposed crossover approach of K-Mean and K-NN. The precision result is better i.e. 99% in this work. Both crossover approaches utilizes KDD'99 dataset".

Liang and Nannan et al. [7] proposed a framework "which is blend of K-Mean and Fluffy C Mean (FCM) calculations to dispose of false positive from the dataset DARPA 2000. The finish of the work is the impact of FCM calculation is superior to anything that of K-Mean grouping". Zhengjie and Yongzhong [8] proposed approach the "K-Mean and molecule Swarm Advancement strategy (PSO-KM). The recognition rate of known assaults is 75.82% and of obscure assaults is 60.8%".

"To enhance the execution of SVM, Horng and Yang,et al. [9] one and half SVM with various leveled bunching.The BRICH progressive bunching calculation is utilized for addition choice system to dispose of immaterial highlights from dataset with the goal that SVM characterize the information all the more precisely. The precision rate of proposed framework is 95.72% and false positive rate is 0.7%".

Classifier	Method	Parameters	Advantages	Disadvantages
Support Vector Machine	A SVM may be a discriminative classifier formally defined by a separating hyperplane.In alternative words,given labeled traning information.The algorithm outputs an optimal	The viability of SVM lies in the resolution of kernel and delicate edge parameters. "For pieces, diverse sets of (C, γ) values are attempted and the one with the best cross-validation accuracy is picked. Attempting	1. Profoundly Exact 2. Ready to demonstrate complex nonlinear choice limits 3. Less inclined to over fitting than different techniques	1.Highly recursive involution and board memory requirements of the specified quadratic Programming in extensive tasks. 2. The decision in the part is troublesome 3.

	hyperplane which categories new examples.	exponentially developing groupings of C is a pratical technique to distinguish good parameters".		The speed both in preparing and testing is moderate.
K-Nearest Neighbour	With the majority of the votes a object is defined as its neighbor,which is carried out by the class.here k is a positive integer.if k=1 then the object is allotted to the class of the nearest neighbours.	Two limitations have examined to optimize the execution of the kNN, the number k of closest neighbor and the element space modification.	1. Scientifically tractable. 2. simple for execution 3. Uses neighborhood data, which can yield exceptionally versatile, conduct 4. Loans itself effectively to parallel usage.	1. Vast capacity necessities. 2. Highly defenseless to the curse of spatiality 3. easy in ordering test records.
Artificial Neural Network	An ANN is a versatile system that change its structure based on outer or inner data that goes through the network during the training stage.[12]	ANN utilizes the cost work C is an essential concept On learning, because it may be an evaluation of however distant a specific arrangement is from an optimum solution for the issue to be resolved.	1.Needs less formal measurable preparing. 2. Ready to certainly recognize complex nonlinear connections amongst subordinate and autonomous factors.	1. "Black box" nature. 2. Great process load 3. vulerability to over fitting. 4.Involues long training time.
Bayesian Method	Algorithm attempts to estimate the conditional probabilities of classes using the joint probabilities of sample observations and classes based on the	In Bayes,class priors and attribute probability distributions parameters will be approximated with corresponding frequencies from the training set.	1. Naive Bayesian classifier clarifies calculations. 2. Display high exactness and speed when connected to extensive databases.	1.The suppositions made in class restrictive autonomy. 2. Absence of accessible likelihood information.

	rule.[13]			
Decision tree	Decision tree constructs a binary organization tree. "Every hub compares to a binary establish on one feature; one branch relates to the positive samples of the found and the other to the negative cases".[11]	Decision Tree Induction utilizes parameters like a set arrangement qualities and a property determination technique.	1. Development doesn't need any domain information. 2.Modify with high dimensional information. 3. Portrayal is straightforward. 4. Set to process both numerical and all out information.	1. Yield quality should be straight out. 2. Constrained to one yield quality. 3. Choice tree calculations are insecure. 4. Trees made from numeric datasets can be mind boggling.

III. IMPLEMENTATION

In proposed approach,the major objective is the combination of supervised and unsupervised learning.K-mean clustering method and kNN classification method should provides solution for identify the anomalous data. Data mining techniques is to identify the intrusions and for each approach as different accuracy,false alarm rate and detection rate. Following figure shows the depiction of proposed work.

3.1. DATASET :

Dataset consists of 41 attributes in each record and it contains of labels either normal or anomalous.There are 5 unique categories, where one shows standard conduct and also the rest demonstrate assaults. Assaults are ordered as Denial of service(DoS), Probing(Prob), Root2Local(R2L), and User2Root(U2R).

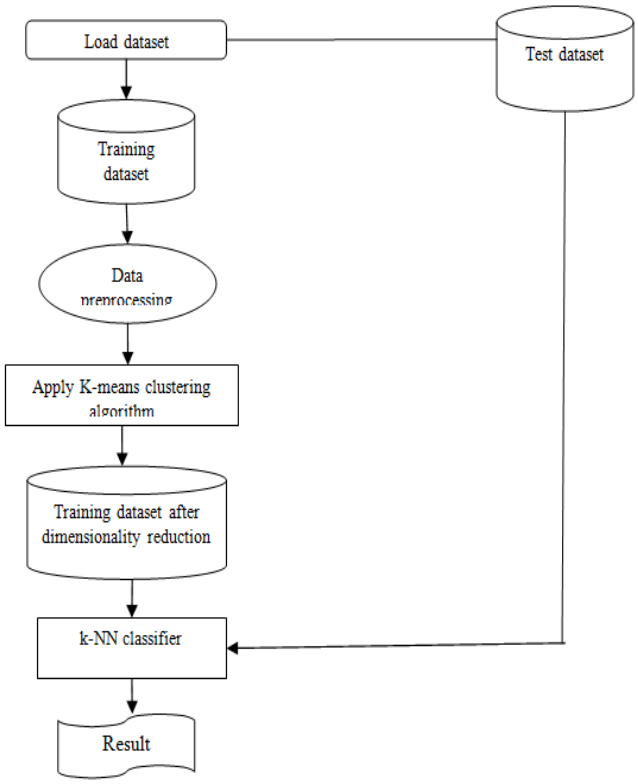

Figure 1: System architecture

Table 1 shows the highlights, and additionally their sorts and number.

Table 1: Feature of KDD cup-99 Dataset and their types and numbers

Type	Features with their Numbers
Nominal	Protocol_type(2),service(3),flag(4)
Binary	Land(7),logged_in(12),root_shell(14),su_attempted(15), is_host_login(21),is_guest_login(22)
Numeric	duration(1),src_bytes(5),dst_bytes(6),wrong_fragment_urgent(9), host(10),num_failed_logins(11),num_compromised(13),num_root(16), num_file_creations(17),num_shells(18), num_access_files(19), num_outbound_cmds(20),count(23),srv_count(24),serror_rate(25), srv_serror_rate(26),rerror_rate(27),srv_rerror_rate(28),same_srv_rate(29)diff_srv_rate(30),srv_diff_host_rate(31),dst_host_count(32), dst_host_srv_count(33),dst_host_same_srv_rate(34), dst_host_diff_srv_rate(35), dst_host_same_src_port_rate(36),dst_host_srv_diff_host_rate(37), dst_host_serror_rate(38),dst_host_srv_serror_rate(39), dst_host_rerror_rate(40),gst_host_srv_rerror_rate(41)

In preprocessing method,different between the configuration of information, it is important to preprocess to convert the character information into numeric information. In KDD dataset, three features are symbolic. These are:

1. Protocol_type: Characterises the convention utilized as a part of the association (e.g. TCP, UDP).

2. Service: Characterises which goal arranges benefit utilized (e.g. Telnet, FTP).

3. Flag: Characterises the status of the connection(e.g. SF, REJ).

3.3. K-MEAN CLUSTERING:

K-Mean Cluster [2] [3] [4], "is a techniques which cluster the related information based on the conduct .K-Mean is an unsupervised learning, i.e. information doesn't indicate what we tend to attempting to learn K-Means clustering used to detect the anomalous information by the researchers. In proposed framework, K-Means cluster functions as a pre-arrangement stage which groups objects in view based on the attributes value into number of disjoint clusters".

$$J(v) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (||x_i - v_j||)^2$$

where,

'||x_i - v_j||' is the Euclidean distance between x_i and v_j.
'c_i' is the number of data points in i^{th} cluster.

'c' is the number of cluster centers.

Algorithmic steps are:

Stage 1:select the quality of centroids objects from dataset because underlying centroids.

Stage 2:Compute the Euclidean separation between every datum and therefore the centroids. Stage 3: If the information point is nearest to the centroid, at that point abandon it and don't roll out any improvement in its position. In any case, if the information point isn't nearest to the centroid, at that point move to its nearest one.

Stage 4: Calculate the centroid of both adjusted groups.

Stage 5: Repeat stage 3 till the point that we get the relentless centroids.

Here euclidean distance between the data point.

Euclidean distance is:

$$Dist((x,y), (a,b)) = \sqrt{(x-a)^2+(y-b)^2}$$

3.4. k-Nearest Neighbour:

k-NN is a simple and basic classification techniques.k-NN is also referred as lazy learning and instance based learning.k-NN is employed within the classification and regression on the applications of the method in many areas.In classification method,k-NN algorithm is a method for classifying the objects based on the closest training data.In regression,k-NN algorithm assigning the property value for the object to be the average of the values of the k-Nearest Neighbours.

Algorithm steps are:
Step 1:Find the K training instance that are nearest to unknown instance.
Step 2:choose the most commonly happening classification for the k instance.

IV. RESULTS

The experimental outcomes are assessed from the proposed system in figure1 on KDD dataset.
In the proposed system,we reduces the dimensions of the training dataset by applying the K-means clustering method to the data.where the number of clusters equal to the number of classes labels.After obtaining the clusters the another steps is to find the nearest neighbour between the data samples and the cluster centers with k-NN classifier then distance are computed.Test data is to estimate the class labels of the test dataset by assigning the label of the neighbour by k-NN classifier.Calculating the performance of the intrusion detection by a accuracy,detection rate and false alarm rate.Hardware necessities used are 2 GB RAM, and 160 GB Hard disk. Fig.2 Shows perform of intrusion detection by the graphically representation.

Performance measures of proposed system are:

Accuracy = TP+TN/TP+TN+FP+FN
Detection Rate = TP/TP+FP
False Alarm = FP/FP+TN

Predicted→ Actual↓	Normal	Intrusions(Attacks)
Normal	TN	FP
Intrusions	FN	TP

True Positives(TP): The number of malicious recognized as malicious.
True Negatives(TN): The number of benign programs properly recognized as benign.
False Postive(FP): The number of benign programs incorrectly recognized as attacks.
False Negative(FN): The number of malicious dishonestly recognizes as malicious.

Table 2 : Sample KDD cup-99 dataset

Protocol	Flag	Dst_bytes	count	Srv_count	Dst_host_count	Serror_rate	Attacks
Udp	SF	146	1	1	255	0	R2l
Udp	SF	146	2	2	255	0	Dos
Udp	S3	146	12	4	187	0	Dos
Tcp	S2	146	22	12	196	0	U2r
Tcp	SF	0	5	21	71	0	Normal
Tcp	S0	185	2	13	3	0	Normal
Icmp	REJ	185	3	20	54	0	Prob
Icmp	SF	260	21	11	174	0	Prob
Udp	SF	146	15	15	255	0	R2l

Tcp	S3	329	2	23	255	0	R2l
Udp	S2	923	22	1	177	0	Dos
Icmp	S0	137	13	4	196	0	U2r
Icmp	RSTU	735	2	12	54	0	Normal
Udp	RSTU	260	1	2	255	0	Normal
Tcp	SF	185	3	13	255	0	Normal

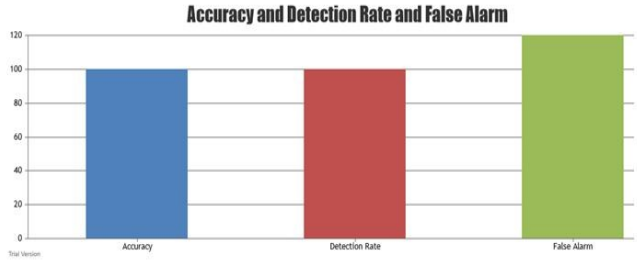

Figure 2. Graphical Analysis of results

V. CONCLUSION

In this work, main contribution is in defining the clusters and the nearest neighbor by combining. “The measure designed is Euclidean distance based on the distance between data samples.K-means approach is chosen for clustering using the distance measure to cluster both the training dataset. k-NN in supervised learning based interruption discovery effectively. Here, k-NN maps the system activity into predefined classes i.e. attack type or normal composes based on the training the label dataset. k-NN based IDS, detection rate and false alarm rate. In this investigation,we propose K-means clustering approach and k-NN approach based on IDS that take care of the issue on the network. It causes IDS to accomplish high Detection Rate,False Alarm Rate, enhance accuracy and thusly high intrusion detection ability”.

VI. REFERENCES

[1]. L. Dhanabal, S.P. Shantharajah, "A study of NSL-KDD Dataset for Intrusion Detection System based on Classification Algorithms", International Journal of Advanced Research in Computer and Communication Engineering, Vol.4, Issue 6, pp. (446-452), June 2015.

[2]. S. Duque, N.B Omar, "Using data Mining Algorithms for Developing a Model for Intrusion Detection System (IDS)", Proceedings of Science direct: Procedia Computer Science 61, pp. (46-51), 2015.

[3]. B. Sharma and H. Gupta, "A design and Implementation of Intrusion Detection System by using Data Mining", IEEE Fourth International Conference on Communication Systems and Network Technologies, pp.700-704, 2015.

[4]. U. Ravale, M. marathe, P. Padiya, "Feature Selection based Hybrid Anomoly Intrusion Detection System using K Means and RBF Kernal Function", Proceedings of Science Direct: International Conference on Advanced Computing Technologies and Applications (ICACTA), pp. 428-435, 2015.

[5]. W. C. Lin, S. W. Ke, C. F. Tsai, "CANN: An intrusion detection system based on combining cluster centers and nearest neighbors", Proceedings of Science direct: Knowledge-Based Systems, pp. 13-21, 2015.

[6]. J. Haque, K.W. Magld, N. Hundewale, "An Intelligent Approach for Intrusion Detection based on Data Mining Techniques", Proceedings of IEEE, 2012.

[7]. Liang Hu, Taihui Li, Nannan Xie, Jiejun hu, "False Positive Elemination in Intrusion Detection based on Clustering", IEEE International Conference on Funny System and Knowledge Discovery (FSKD), pp. 519-523, 2015.

[8]. Zhengjie Li, Yongzhong Li, Lei Xu, "Anomoly Intrusion Detection Method based on K-Means Clustering Algorithm with Particle Swarm Optimization", IEEE International Conference

of Information Technology, Computer Engineering and Management Sciences, pp. 157-161, 2011.

[9]. S. J. Horng, M.Y. Su, Y. H. Chen, T. W. Kao, R. J. Chen, J. L. Lai, C. D. Perkasa, "A novel intrusion detection system based on hierarchical clustering and support vector machines", Proceedings of Science direct: Expert Systems with Applications, pp. 306-313, 2011.

[10]. Baowei Song Chunxue Wei," Algorithm of Constructing Decision Tree Based on RoughSet Theory, " International Conference on Computer and Communication Technologies Agriculture Engineering, IEEE 2010.

[11]. Weiguo Yi, Jing Duan, Mingyu Lu "Optimization of Decision Tree Based on Variable Precision Rough Set" International Conference on Artificial Intelligence and Computational Intelligence IEEE 2010.

[12]. V. R. Dinavahi and S. C. Srivastava, "ANN based voltage stability margin prediction," in Proc. IEEE Power Eng. Soc. Summer Meeting, vol. 2, pp. 1275–1280, July 2001.

[13]. Chen, F. (2009), "Bayesian Modeling Using the MCMC Procedure," in Proceedings of the SAS Global Forum 2008 Conference,Cary NC: SAS Institute Inc