Outlier Detection and Removal Using Data Mining Techniques

Pranitha Awati,
PG Scholar (M.TECH), Department of Computer Science and Engineering,
VNR Vignan Jyothi Institute of Engineering and Technology,

Abstract- From the onset of web arrangement, protection menaces normally recognized as intrusions has come to be extremely vital and critical subject in web arrangements, data and data system. In order to vanquish these menaces every single period a detection arrangement was demanded because of drastic development in networks. Because of the development of arrangement, attackers came to be stronger and every single period compromises the protection of system. Hence a demand of Intrusion Detection arrangement came to be extremely vital and vital instrument in web security. Detection and prevention of such aggressions shouted intrusions generally depends on the skill and efficiency of Intrusion Detection Arrangement (IDS). In this way endless troupe component has been directed by utilizing endless philosophies', these systems have their own advantages and deficiencies. Here mainly focusing on different classification methods.

Keywords- Intrusion Detection, Anomaly Detection, Misuse Detection, KDD Cup99, Ensemble Approaches

I. INTRODUCTION

From past decades alongside quick progress in the Internet established knowledge, new request spans for computer web have emerged. At the alike period, expansive range progress in 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 appealing target for the mistreatment and a large vulnerability for the community. A fun to do work or a trial to achieve deed for a little people came to be a bad dream for the others. In countless cases malicious deeds made this nightmare to come to e a reality.

In supplement to the hacking, new entities like worms, Trojans and viruses gave extra panic into the net- worked society. As the present situation is a moderately new phenomenon, web armaments are weak. Though, due to the popularity of the computer webs, their connectivity and our ever producing dependency on them, realization of the menace can have desecrating consequences. Safeguarding such a vital groundwork has come to be the priority one scrutiny span for countless researchers.

Aim of this paper is to study the present trends in Intrusion Detection Arrangements (IDS) and to examine a little present setback that continues in this scrutiny area. In analogy to a little mature and well stayed scrutiny spans, IDS is a youthful earth of research. Though, due to its duty critical nature, it has enticed momentous attention towards itself. Density of scrutiny on this subject is constantly rising and everyday extra researchers are involved in this earth of work. The menace of a new wave of cyber or web aggressions is not just a probability that ought to be believed, but it is a consented fact that can transpire at each time. The present trend for the IDS is distant from a reliable protective arrangement, but instead the main believed is to make it probable to notice novel web attacks.

One of the main concerns is to make sure that in case of an intrusion endeavor, the arrangement is able to notice and to report it. After the detection is reliable, subsequent pace should be to protect the web (response). In supplementary words, the IDS arrangement will be upgraded to an Intrusion Detection and Reply Arrangement (IDRS). Though, no portion of the IDS is presently at a fully reliable level. Even nevertheless researchers are concurrently involved in working on both detection and answer factions of the system. A main setback in the IDS is the promise for the intrusion detection. This is the reason why in countless cases IDSs are utilized jointly alongside a human expert. In this method, IDS is truly helping the web protection captain and it is not reliable plenty to be trusted on its own. The reason is the in- skill of IDS arrangements to notice the new or modified attack patterns. Even though the latest creation of the detection methods has considerably enhanced the detection rate, yet there is a long method to go.

There are two main ways for noticing intrusions, signature-based and anomaly-based intrusion detection. In the early way, attack outlines or the deeds of the intruder is modeled (attack signature is modeled). Here the arrangement will gesture the intrusion after a match is detected. Though, in the subsequent neither way nor- mal deeds of the web is modeled. In this way, the arrangement will rise the alarm after the deeds of the web does not match alongside its normal behavior. There is one more Intrusion Detection (ID) way that is shouted specification-based intrusion detection. In this way, the normal deeds (expected behavior) of the host is enumerated and subsequently modeled. In this way, manage worth for the protection, freedom of procedure for the host is limited. In this paper, these ways will be briefly debated and compared.

The believed of possessing an intruder accessing the arrangement lacking even being able to notice it is the worst nightmare for each web protection officer. As the present ID knowledge is not precise plenty to furnish a reliable detection, heuristic methodologies can be a method out. As for the last line of protection, and in order to cut the number of undetected intrusions, heuristic methods such as Honey Jars (HP) can be deployed. HPs can be installed on each arrangement and deed as mislead or decoy for a resource.

Another main setback in this scrutiny span is the speed of detection. Computer webs have a vibrant nature in a sense that data and data inside them are unceasingly changing. Therefore, noticing an intrusion precisely and punctually, the arrangement has to work in real time. Working in genuine period isn't simply to per-shape the location in genuine period; however is to change to the new elements in the system. Real period working IDS is an alert scrutiny span pursued by countless researchers. Most of the scrutiny works are aimed to familiarize the most period effectual methodologies. The aim is to make the requested methods suitable for the real period implementation.

From a disparate outlook, two ways can be envisaged in requesting IDS. In this association, IDS can be whichever host established or web based. In the host established IDS, arrangement will merely protect its own innate ma- chine (its host). On the supplementary hand, in the web established IDS, the ID procedure is somehow distributed alongside the network. 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 could manipulation or monitor web firewalls, web routers or web switches as well as the client machines.

The main emphasis of this paper is on the detection portion of the intrusion detection and reply problem. Re- searchers have pursued disparate ways or a combination of disparate ways to resolve this problem. Every single way has its own theory and presumptions. This is so because there is no precise behavioral ideal for the legitimate user, the intruder or the web 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 execution measures are execution time (120ms), CPU Utilization (74%) and memory use (54%)”.

Ravale and Nilesh et al. [4] proposed half and half approach of K-Mean and RBF kernal 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 of “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] half and half SVM with various leveled bunching. The BRICH progressive bunching calculation is utilized for include 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 support vector machine develops a hyper plane or set of hyper planes in a high or endless dimensional space, which can be utilized for arrangement, relapse or different assignments”.	“The viability of SVM lies in the determination of portion and delicate edge parameters. For pieces, diverse sets of (C, γ) values are attempted and the one with the best cross-approval exactness is picked. Attempting exponentially developing groupings of C is a down to earth technique to distinguish great parameters”.	1. Profoundly Exact 2. Ready to demonstrate complex nonlinear choice limits 3. Less inclined to over fitting than different techniques	1. High algorithmic intricacy and broad memory prerequisites of the required quadratic Programming in expansive scale undertakings. 2. The decision of the part is troublesome 3. The speed both in preparing and testing is moderate.

K Nearest Neighbor	“A protest is characterized by a larger part vote of its neighbors, with the question being doled out to the class most regular among its k closest neighbors (k is a positive whole number). In the event that k = 1, at that point the protest is basically allotted to the class of its closest neighbor”.	“Two parameters are considered to optimize the execution of the kNN, the number k of closest neighbor and the element space change”.	1. Scientifically tractable. 2. Basic in usage 3. Utilizations neighborhood data, which can yield exceptionally versatile, conduct 4. Loans itself effectively to parallel usage.	1. Vast capacity necessities. 2. Profoundly defenseless to the scourge of dimensionality. 3. Slow in ordering test tuples.
Artificial Neural Network	“An ANN is a versatile framework that progressions its structure in view of outside or inner data that moves through the system amid the learning stage”.	“ANN utilizes the cost work C is an essential idea in learning, as it is a measure of how far away a specific arrangement is from an ideal answer for the issue to be solved”.	“1. Requires less formal measurable preparing. 2. Ready to certainly recognize complex nonlinear connections amongst subordinate and autonomous factors. 3.High resilience to uproarious information. 4. Accessibility of numerous preparation calculations.”	1. "Black box" nature. 2. Greater computational burden. 3. Proneness to over fitting. 4.Requires long training time.
Bayesian Method	“Based on rule, utilizing the joint probabilities of test perceptions and classes, the calculation endeavors to appraise the restrictive probabilities of classes given a perception”.	“In Bayes, every model parameter (i.e., class priors and highlight likelihood circulations) can be approximated with relative frequencies from the preparation set”.	1. Guileless Bayesian classifier streamlines the 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.
Decision Tree	“Choice tree constructs a double grouping tree. Every hub compares to a double predicate on one characteristic; one branch relates to the positive examples of the predicate and the other to the negative cases”.	“Choice Tree Enlistment employments parameters like an arrangement of hopeful qualities and a property determination technique”.	“1. Development does not require any area information. 2. Can deal with high dimensional information. 3. Portrayal is straightforward. 4. Ready to process both numerical and all out information.”	1. Yield quality must 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

A.SYSTEM ARCHITECTURE:

Literature review “speaks to that numerous analysts done research on methodologies of information mining to identify the interruptions and each approach has distinctive exactness, false caution rate and discovery rate. The proposed work is a mix of administered and unsupervised methodologies. K-Mean and KNN should give better answer for recognize the bizarre information. Following is the depiction of proposed work”.

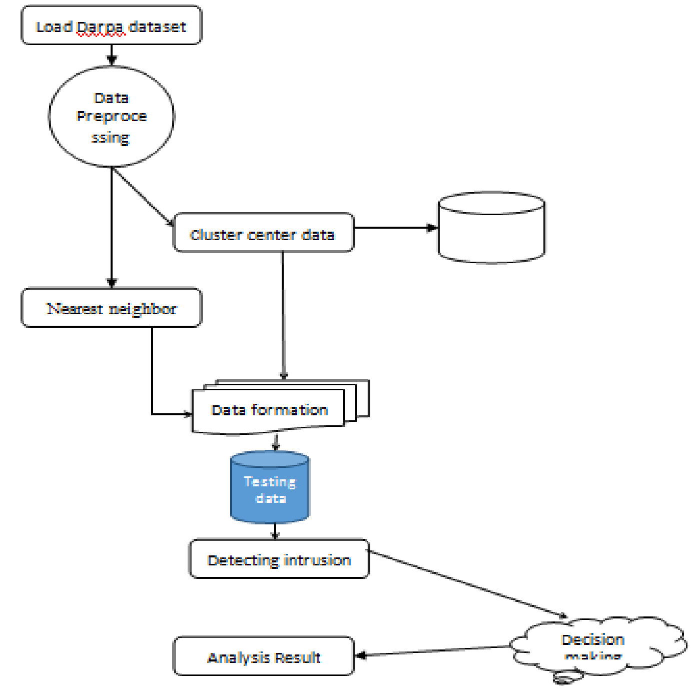


Fig.1: System architecture

PREPROCESSING :

Because of the distinction between the configurations of information, it is important to preprocess it jump at the chance to change over the character information into numeric information. In NSL-KDD dataset, three traits are emblematic. These are:

1. **Protocol type:** Characterizes the convention utilized as a part of the association (e.g. TCP, UDP).

2. **Service:** Characterizes which goal arranges benefit utilized (e.g. telnet, FTP).

3. **Flag:** Characterizes the status of the association (e.g. SF, REJ).

K-MEAN CLUSTERING:

K-Mean Clustering [2] [3] [4], is a “system which clustering the comparative information in view of the conduct. K-Mean is an unsupervised assignment, i.e. information doesn't indicate what we are attempting to learn. Numerous specialists utilize K-Means clustering in the mixture ways to deal with distinguish the irregular information. In proposed framework, K-Means clustering functions as a pre-arrangement stage which bunches objects in view of the component esteem into number of disjoint clusters”.

Algorithmic steps are:

Stage 1: “Choose the quantity of centroids objects from dataset as the underlying centroids”.

Stage 2: Then, “Compute the Euclidean separation between every datum point and 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 it to its nearest one.

Stage 4: “Recalculate the centroid of both adjusted groups”.

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

In other words, its objective is to find [4]:

$$M = \sum_{a=1}^{k} \sum_{b=1}^{n} d_{ab} \left(x_b , y_a \right)$$

Where, d_{ab} (x_b,y_b) is an eculidean distance between the data point x_b and y_a the centroid .

Euclidean distance is:

$$d(x_b , y_a) = \| \ x_b - y_a \ \|$$

B.Dataset:

“Statistical analyses on KDD CUP 99, demonstrated that this dataset has shortcomings that impact on systems` execution. Its real shortcoming is its dull records, which causes an inclination towards visit information. In the wake of researching and breaking down this set, it was realized that 78% of the preparation information and 75% of the test information are dreary [21]. Accordingly, this examination utilizes NSL KDD1. The aggregate number of records in this dataset is 30000, where 5000records are typical information and the rest demonstrate assaults. The aggregate number of highlights is 41, which incorporate numeric, ostensible, and paired highlights. Table I shows the highlights, and additionally their sorts and numbers”.

This dataset comprises of five unique classes, where one shows ordinary conduct and the rest demonstrate assaults. Assaults are ordered as DoS, Probe, R2L, and U2R.

C. Evaluation Parameters:

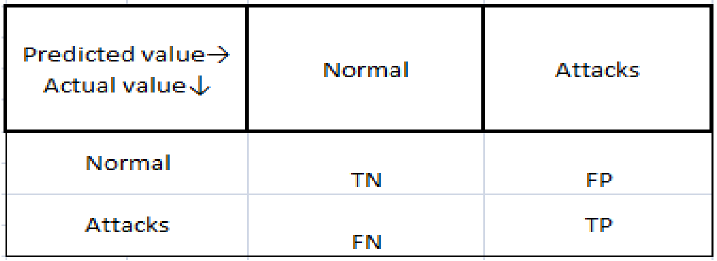
This investigation utilizes some appraisal measurements, for example, exactness, identification rate, and false alert rate as assessment parameters, which are registered in view of the perplexity network in table III.

Performance measures of proposed system are:

Accuracy = TP+TN/TP+TN+FP+FN

Detection Rate = TP/TP+FP

False Alarm = FP/FP+TN



Predicted value→ Actual value↓	Normal	Attacks
Normal	TN	FP
Attacks	FN	TP

True Positives(TP): The number of effectively recognized attacks.

True Negatives(TN): The number of harmless application correctly recognized as harmless.

False Postive(FP): The number of harmless applications falsely recognized as attacks.

False Negative(FN): The number of harmless applications dishonestly perceived as attacks.

IV. RESULTS

Table 2. Also, Fig. 2. Demonstrates that the precision of the proposed framework is significantly more when contrasted with singular information mining strategy".

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

Table2: KDD cup-99 Dataset

True positive	False positive	True negative	False negative
100	0	0	0

Table4: Accuracy

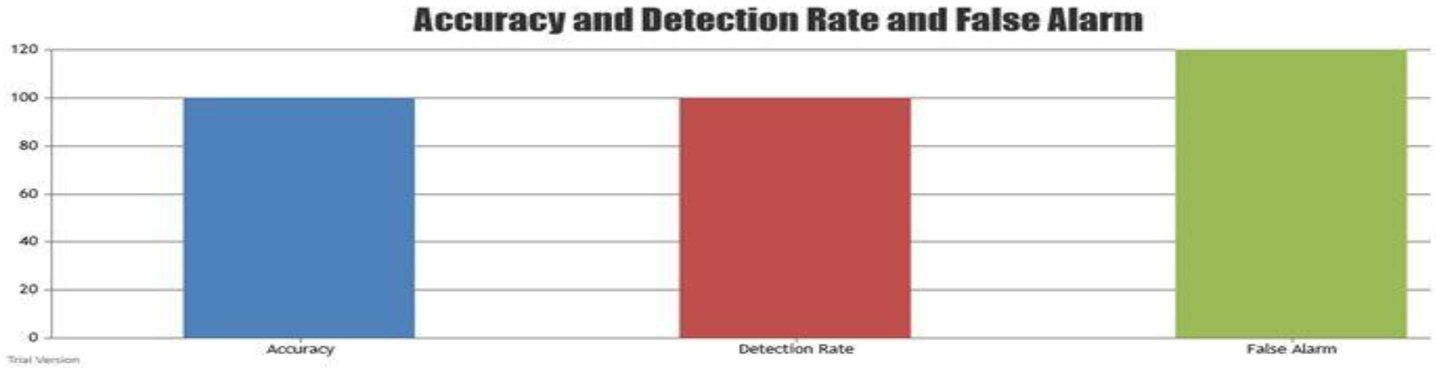


Fig.2: Graphical Analysis of results

V. CONCLUSION

Due to our increased dependence on Internet and developing number of interruption occurrences, building viable interruption recognition frameworks are fundamental for ensuring Internet assets but it is an awesome test. In writing, numerous analysts used k-NN in administered learning based interruption discovery effectively. Here, k-NN maps the system activity into predefined classes i.e. ordinary or particular assault composes in light of preparing from mark dataset. Be that as it may, for k-NN based IDS, detection rate (DR) and false positive rate (FPR) are as yet should have been made strides. In this investigation, "Here propose a troupe approach, called MANNE, for k-NN-based IDS that advance k-NN by Multi-Objective Genetic Algorithm to take care of the issue. It causes IDS to accomplish high DR, less FPR, enhance precision and thusly high interruption recognition 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]. Whatistarget.com/definition/confidentiality-integrity-and-availability-CIA

[11]. J. Han, M. Kamber, "Data Mining: Concepts and Technnologies", Third Edition.

[12]. http://nsl.cs.unb.ca/NSL-KDD/

[13]. Dae-Ki Kang and Doug Fuller et al., "Learning Classifiers for Misuse and Anomaly Detection Using a Bag of System Calls Representation", IEEE Workshop on Information Assurance and Security United States Military Academy (2005).

[14]. K. Shivshankar E., "Combination of Data Mining Techniques for Intrusion Detection System", IEEE International Conference on Computer, Communication and Control (IC4-2015).

[15]. Jain Patik P and Madhu B.R., "Data Mining based CIDS: Cloud Intrusion Detection System for Masquerade attacks [DCIDSM]", IEEE 4th ICCCNT (2013).

[16]. J.Yang, R.Yan, A.G.Hauptman, "Cross-Domain Video Concept Detection Using Adaptive SVMs", Proceedings of ACM, MM'07, Augsburg, Bavaria, Germany, 2007.