International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

Hybrid Approch of KNN+Euclidean Distance to Detect Intrusion within Cloud Based Systems

Upasna Khanna[1], Prabhdeep Singh[2]

[1]Research Scholar of M.Tech, Department of Computer Science and Engineering, Global Institute of Management & Emerging Technologies, Amritsar, Punjab, India

[2]Assistant Professor, Department of Computer Science and Engineering, Global Institute of Management & Emerging Technologies, Amritsar, Punjab, India

***Abstract*— *K-Nearest-Neighbors is one of the least complex yet effective characterization strategies. The center calculation behind it is to figure the distance from a question indicate the greater part of its neighbors and to pick the nearest one. The Euclidean distance is the most successive decision making strategy implemented in proposed approach for intrusion detection in cloud. This paper investigates a basic yet compelling likeness definition inside Nearest Neighbors for IDS applications. This novel similitude lead is quick to process and accomplishes an extremely accurate execution on the ID benchmark.*

***Keywords*— *K-Nearest-Neighbors, Euclidean Distance, Intrusion Detection System.*

I. Introduction

In this paper we utilize a machine learning way to deal with plan what's more, actualize an interruption discovery framework. The framework will enable the PC to guard against security dangers, while exceptionally little PC security information is required. This interruption identification framework is upheld by the smart machine learning and information mining calculations as opposed to the very refined PC security advancements. It recognizes the vindictive exercises utilizing a K Nearest Neighbors (KNN) characterization technique.(Science & Mangalore, 2016)KNN is a technique for grouping objects in light of the information gained from the preparation information. One of the key difficulties is to think of a legitimate approach to compute the separate between information focuses that speak to network movement logs.(Science & Mangalore, 2016) This paper will show another approach to process closeness inside KNN that is extremely straightforward and proficient to execute, what's more, is by all accounts extremely appropriate to the kind of issue we are handling. We accomplish best in class execution on two arrange interruption recognition datasets.

II. Related Work

Analysis

In nowadays a solitary server handles the various solicitations from the client. Here the server needs to handle the every one of the solicitations from the clients at the same time, so the preparing time will be high. This may prompts loss of information and parcels might be postponed and debased.(Singh & Tiwari, 2015) On doing this the server can't handle the inquiry from the client in a legitimate way. So the preparing time gets expanded. It might prompts movement and blockage. To defeat these issues we are going for the idea called distributed computing. In this distributed computing we will execute the Proxy server to maintain a strategic distance from these issues. However, in this framework Data Efficiency is enhanced yet not the information security. At whatever point we talk about information effectiveness we ought to talk about information security additionally, in light of the fact that in the distributed computing we don"t know from which cloud the information is coming, so in the current framework there is no framework to discover the information security. The framework in light of the new design has better versatility and adaptation to internal failure. A group comprises of a solitary server and different intermediary servers and is gotten to by numerous customers.(K. J. Chabathula, C. D. Jaidhar, & M. A. Ajay Kumara, 2015) Intermediary servers stores information on neighborhood plates and read or compose information determined by a server. The server keeps up the list for all record put away in various intermediaries. At the point when a customer needs to download a few information, it will initially send a demand to the Server and the Server at that point divert the demand to a relating intermediary that have the required information and thus the information will be sent to the customer. With the blend of Cloud and Grid figuring ideas, the information demand can be productively adjusted in an auspicious way. The real piece of the Project is Security, so previously mentioned stage talks about Cloud and Grid Technology, yet not about security. The Security execution is accomplished by two stage, in particular Behavioral – Knowledge.

Behavior Analysis

(Haddad Pajouh, Javidan, Khayami, Ali, & Choo, 2016)Utilizing this technique, we have to perceive expected conduct (honest to goodness utilize) or an extreme conduct deviation. The system must be effectively prepared to productively distinguish interruptions. For a given interruption test set, the system figures out how to recognize the interruptions. Be that as it may, we concentrate on recognizing client behavioral examples and deviations from such examples. With this system, we can cover a more extensive scope of obscure assaults.

Knowledge Analysis

(Onik, Haq, & Mustahin, 2015) Utilizing a specialist framework, we can depict a vindictive conduct with a run the show. One preferred standpoint of utilizing this sort of interruption recognition is that we can include new guidelines without adjusting existing ones. Interruption identification (ID) is a sort of security administration framework for PCs and systems. An ID framework assembles and breaks down data from different territories inside a PC or a system to recognize conceivable security ruptures, which incorporate both interruptions (assaults from outside the association) and abuse (assaults from inside the association). ID utilizes powerlessness evaluation (in some cases alluded to as examining), which is an innovation created to survey the security of a PC framework or system. Intrusion location capacities include:

Observing and investigating both client and framework exercises.

- Analyzing framework designs and vulnerabilities
- Assessing framework and record uprightness
- Related existing techniques

Intrusion Detection for Grid and Cloud Computing

(Canbay & Sagiroglu, 2015)Cloud and Grid processing are the most powerless focuses for intruder‟s assaults because of their appropriated condition. For such conditions, Intrusion Detection System (IDS) can be utilized to upgrade the safety efforts by a deliberate examination of logs, designs and system activity. Customary IDSs are not appropriate for cloud condition as system based IDSs (NIDS) can't recognize scrambled hub correspondence, likewise have based IDSs (HIDS) are not ready to locate the concealed assault trail. Kleber, schulter et al. have proposed an IDS benefit at cloud middleware layer, which has a review framework intended to cover assaults that NIDS and HIDS can't recognize. The design of IDS administration incorporates the hub, benefit, occasion inspector and capacity. The hub contains assets that are gotten to through middleware which characterizes get to control arrangements.(Xie & Hu, 2013) The administration encourages correspondence through middleware. The occasion inspector screens and catches the system information, additionally breaks down which control/approach is broken. The capacity holds conduct based (correlation of late client activities to regular conduct) and learning based (known trails of past assaults) databases. The examined information is sent to IDS benefit center, which investigates the information and alert to be an interruption. The creators have tried their IDS model with the assistance of reenactment and discovered its execution agreeable for realtime usage in a cloud domain. In spite of the fact that they have not examined the security arrangements consistence check for cloud specialist organization and their announcing methods to cloud clients.

Intrusion Detection in the Cloud

(Huijun, Hong, & Hong, 2013)Interruption identification framework assumes an imperative part in the security and steadiness of dynamic safeguard framework against gatecrasher unfriendly assaults for any business and IT association. IDS usage in distributed computing requires an effective, versatile and virtualization-based approach. In distributed computing, client information and application is facilitated on cloud benefit provider‟s remote servers and cloud client has a restricted control over its information and assets. In such case, the organization of IDS in cloud turns into the duty of cloud supplier. Despite the fact that the manager of cloud IDS ought to be the client and not the supplier of cloud administrations. (Behrozinia, Azmi, Keyvanpour, & Pishgoo, 2013) have proposed a joining answer for focal IDS administration that can consolidate and incorporate different prestigious IDS sensors yield gives an account of a solitary interface. The interruption identification message trade organize (IDMEF) standard has been utilized for correspondence between various IDS sensors. The creators have recommended the sending of IDS sensors on particular cloud layers like application layer, framework layer and stage layer. Cautions produced are sent to „Event Gatherer‟ program. Occasion gatherer gets and change over ready messages in IDMEF standard and stores in occasion information base storehouse with the assistance of Sender, Receiver and Handler modules. The investigation part breaks down complex assaults and introduces it to client through IDS administration framework. The creators have proposed a compelling cloud IDS administration design, which could be observed and directed by the cloud client. They have given a focal IDS administration framework in view of various sensors utilizing IDMEF standard for correspondence and checked by cloud client.

Security Issues in Cloud Computing

Cloud data confidentiality issue

(Daneshpazhouh & Sami, 2013)Classification of information over cloud is one of the glaring security concerns. Encryption of information should be possible with the customary systems. Be that as it may, encoded information can be secured from a vindictive client yet the protection of information even from the executive of information at administration provider‟s end couldn't be covered up. Looking and ordering on encoded information remains a state of worry all things considered. Previously mentioned cloud security issues are a couple and dynamicity of cloud design are confronting new difficulties with fast usage of new administration worldview

Network and Host Based Attacks on Remote Server

(Weiming & Hongzhi, 2013)Host and system interruption assaults on remote hypervisors are a noteworthy security worry, as cloud sellers utilize virtual machine innovation. DOS and DDOS assaults are propelled to refuse assistance accessibility to end clients.

Cloud Security Auditing

(Daneshpazhouh & Sami, 2013)Cloud evaluating is a troublesome undertaking to check consistence of all the security strategies by the merchant. Cloud specialist organization has the control of delicate client information and procedures, so a computerized or outsider evaluating system for information trustworthiness check and legal investigation is required. Protection of information from outsider inspector is another worry of cloud security.

Lack of Data Interoperability Standards

It comes about into cloud client information secure state. On the off chance that a cloud client needs to move to other specialist co-op because of specific reasons it would not have the capacity to do as such, as cloud user‟s information and application may not be good with different vendor‟s information stockpiling organization or stage. Security and classification of information would be in the hands of cloud specialist co-op and cloud client would be reliant on a solitary specialist co-op.

III. Proposed Approach

Similarity Definition

The K-nearest neighbor classifier works in light of the remove ascertained between the question point and every information focuses in the preparation dataset. At that point we pick the K nearest focuses and take a vote of their class marks to choose the name of the question point. Consequently, ostensibly, the essence in K-nearset neighbor arrangement is the meaning of a distance (or proportionately, a likeness). Include determination is now part of characterizing similitude since it figures out which elements are taken into account in the comparability calculations. Presently it stays to characterize in which approach to consider the chose highlights. The most widely recognized distance work utilized as a part of K-closest neighbor grouping is the Euclidean distance. The reason for this is just that it is the idea of distance we are utilized to the distance one would quantify with a ruler. Be that as it may, with regards to high dimensional complex information sorts, the Euclidean distance is no longer reasonable or not material by any stretch of the imagination.

To get around this issue, there has been distributed work in the writing on different distance measurements for blended components, e.g. (Gopal, Yang, Salomatin, & Carbonell, 2011)Here we characterize and utilize another and to a great degree straightforward approach to compute the similitude between two such information objects, which we depict in the spin-off. We characterize the closeness of M and N as the total of ‘matches' between their relating highlights. For one highlight of M and the relating highlight of N, on the off chance that they are indistinguishable or situated in a similar interim then we characterize their comparability to be 1, generally the similitude squares with 0. At that point we include all these 1-0 esteems to get the comparability between these two information focuses. A portion of the components are clear cut, others are whole numbers, and for these it is anything but difficult to think about whether they are indistinguishable or not. Some different elements have nonstop esteems. For these it is not sensible to just take a gander at whether their qualities are indistinguishable – more often than not they are not indistinguishable but rather still have a similar importance. For instance, when the term for a association is 9999 seconds in one information point and 9998 seconds in the other, we could state that the lengths of these two associations are same. In this way, for components that take consistent esteems.(Mafra, Moll, Da Silva Fraga, & Santin, 2010) That is, the genuine esteems are put into a few interims to begin with, and afterward we apply the possibility of comparability by coordinating the interims instead of the first esteems. Further, some the consistent esteemed elements go from "0" to "1379963888" for instance. In the event that we straightforwardly analyze two computerized values which could go from 0 to 1379963888, it is unlikely to discover two components with a similar esteem. That implies the likeness for the two information purposes of this specific highlight would be quite often 0. Putting these qualities into interims can take care of this sort of issue as well. Having set up the value of binning the component values, we have to choose about the quantity of containers. For illustration, a component with qualities from "0" to "100", keeping in mind the end goal to get a more precise outcome, we isolate it by 10 and place it into 10 distinctive interims. (Yu, Chan, Ng, & Yeung, 2010)The aggregate number of interims to consider very relies on upon the scope of the qualities and the genuine significance of the element. We can assign these qualities into 5 interims, 10 interims, 50 interims, 100 interims or whatever other number. On the off chance that the element is exceptionally delicate to the progressions of the esteem, at that point we may put these qualities in a bigger number of interims and the other way around. There is dependably an exchange off between an excessive number of interims and excessively couple of interims. Judgment skills and some underlying (profoundly non-thorough) experimentation guided us in this decision.

Choosing the Number of Nearest Neighbors

(Wang et al., 2016)In the K-closest neighbor classifier, the class mark of the target information point is chosen by its nearest neighbors. The precision of the grouping exceedingly relies on upon the quantity of closest neighbors picked. Picking excessively numerous neighbors would get some uproarious information that is immaterial to the question point, thus the classifier's expectation is a total of all these insignificant information focuses. In different words, the classifier under-fits. Then again, picking excessively few neighbors can likewise adversely influence the grouping precision from that point forward the expectation depends on excessively couple of information focuses, and the classifier over-fits. A little esteem of K implies that commotion information could affect the outcome. Be that as it may, with our closeness definition we find for this sort of issue that picking K is substantially less demanding. This is since by development our similitude might be thought of as characterizing a limited number of concentric circles with various spans whose inside is the unlabelled inquiry point. The range of a hover equivalents to the distance (opposite closeness) between the inquiry point and any of the preparation focuses that have the same closeness score. These focuses sit on a similar circle. Utilizing this definition, the conceivable estimations of K range from 1 to the quantity of elements chosen by the framework and the comparability scores go from 0 to the greatest of number of highlights chose and they are all whole number esteems. For guaranteed comparability, it will discover a gathering of information focuses which share a similar closeness. For the case we gave before, the K could be 1, 2, 3, 4 or 5 and the distance could be 0, 1, 2, 3, 4, or, on the other hand 5. Subsequently, we find when setting K=1, there is an entirety accumulation of neighbors that have the same biggest comparability with the question information point. Actually, K=1 ends up being sufficient. As far as we can tell, expanding K will rapidly bring about a immense gathering of neighbors, which does not help any further. Hence, we found that 1 is the best an incentive for K in this assignment.

Euclidean Distance

(Mafra et al., 2010)The Euclidean distance or Euclidean metric is the "conventional" straight-line remove between two focuses in Euclidean space. With this distance, Euclidean space turns into a metric space. The related standard is known as the Euclidean standard. More established writing alludes to the metric as Pythagorean metric. A summed up term for the Euclidean standard is the L ^{2} standard.

$$D(p,q) = d(q,p) = \sqrt{(q_1 - p_1) + (q_2 - p_2)^2 \ldots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)}^2$$

Example

Give us a chance to consider measuring the distances between our 30 tests in Exhibit 1.1, utilizing just the three consistent factors contamination, profundity and temperature. What might happen on the off chance that we connected recipe (4.4) to quantify remove between the last two specimens, s29 and s30.

$$D_{s29,s30} = \sqrt{(6.0-1.9)^2 + (51-99)^2 + (3.0-2.9)^2}$$

$$= \sqrt{16.81 + 2304 + 0.01} = \sqrt{2320.82} = 48.17$$

Hybrid Approach

By combining KNN with Euclidean distance intrusion from within cloud computing can be detected. KNN specifies number of neighbors to be analyzed and Euclidean distance gives closest distance between the neighbors. In case neighbors detected are more than threshold specified, nodes could be labeled as malicious. The algorithm for the same is listed as under

```
Algorithm IDSKNNE
1. Initialize Cloud with nodes as vms within datacenters.
2. Declare initial neigbours as threshold for source and destination.
3. Initialize dataset for analysis.
4. Specify value of K.
5. Calculate Euclidean distance for specified value of K
       $ED_i = \sqrt{(Y_i - Y_{i+k})^2 + (X_i - X_{i+k})^2}$
6. Compare $ED_i$ with threshold value specified
   If($ED_i$<Threshold_Neighbour)
       Declare $ED_i$ as Neighbour of source
       $Count_i$=$Count_i$+1
   End of if
   If($Count_i$>Threshold_Intruder)
       Declare $Node_i$ as Intruder
   End of if
7. Repeat step for every node in Cloud
8. Output result in the form of prediction accuracy
```

The result and performance analysis is listed in next section

IV. Result and Performance Analysis

The result in terms of time consumption and number of intruders detected along with prediction accuracy is given. The time consumption associated with proposed and existing approaches is listed as under

TABLE 1. Showing error rate.

ERROR_RATE	
KNN WITHOUT EUCLIDEAN DISTANCE	KNN WITH EUCLIDEAN DISTANCE
6.67	4.56
6.99	5.11
7.94	5.99
8.46	6.99
9.98	7.99

Error rate indicates difference in actual and predicted values. Proposed approach gives better accuracy since error rate is minimized.

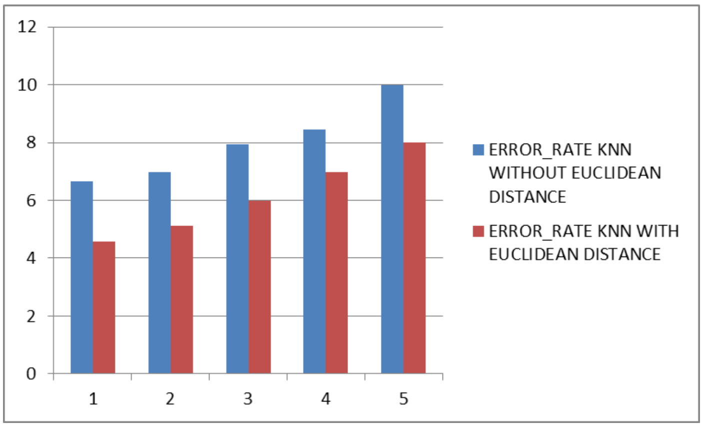

Fig. 1. Showing ERROR_RATE.

TABLE 2. Showing PREDICTION ACCURACY.

PREDICTION ACCURACY	
KNN WITHOUT EUCLIDEAN DISTANCE	KNN WITH EUCLIDEAN DISTANCE
93.33	97.56
92.11	94.19
91.15	93.59
90.89	91.99
88.28	90.56
86.15	88.2

Prediction accuracy is achieved with the help of error rate. Minimum error gives maximum accuracy. Accuracy enhanced proves worth of the study.

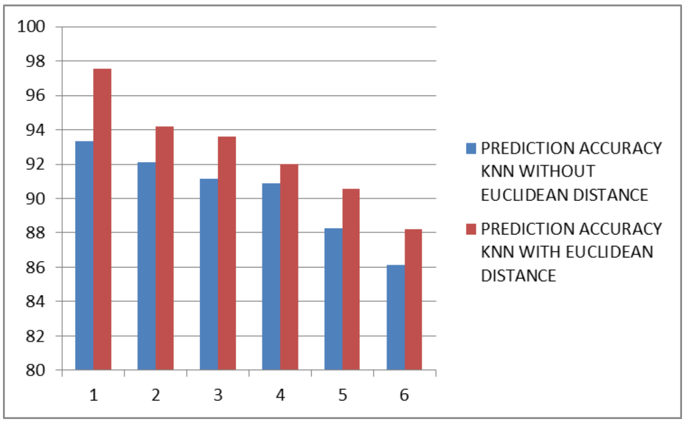

Fig. 2. Showing PREDICTION ACCURACY.

TABLE 3. Showing TIME_CONSUMED.

TIME_COMSUMED	
KNN WITHOUT EUCLIDEAN DISTANCE(us)	**KNN WITH EUCLIDEAN DISTANCE(us)**
7.11	5.1
6.99	4.29
6.54	3.89
5.89	3.03
5.17	2.46

Time consumed indicates total time consumed in terms of micro seconds. Time consumed through proposed approach is minimized.

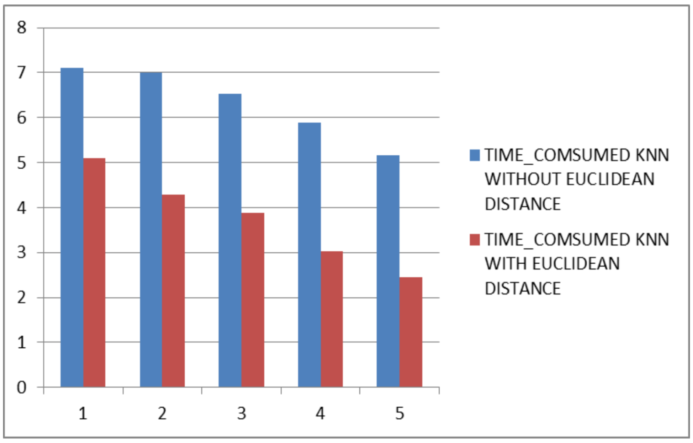

Fig. 3. Showing TIME_CONSUMED.

V. Conclusion and Future Scope

Proposed approach uses hybridization of KNN and Euclidean distance to detect intrusion if any within cloud environment. Results obtained in terms of time consumed and accuracy. Nearest neighbor in existing approach uses Manhattan distance approach which is changed to Euclidean distance in proposed approach for intrusion detection. Performance comparison of existing and proposed approach is compared to prove worth of the study.

In future performance analysis of SVM with KNN can be used for intrusion detection in Cloud Environment.

References

[1] S. Behrozinia, R. Azmi, M. R. Keyvanpour, and B. Pishgoo, "Biological inspired anomaly detection based on danger theory," *5th Conference on Information and Knowledge Technology (IKT)*, pp. 102–106, 2013.

[2] Y. Canbay and S. Sagiroglu, "A hybrid method for intrusion detection," *IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pp. 156–161, 2015.

[3] A. Daneshpazhouh and A. Sami, "Semi-supervised outlier detection with only positive and unlabeled data based on fuzzy clustering," *5th Conference on Information and Knowledge Technology*, pp. 344–348, 2013.

[4] S. Gopal, Y. Yang, K. Salomatin, and J. Carbonell, "Sctatistical learning for file-type identification," *Proceedings - 10th International Conference on Machine Learning and Applications, ICMLA 2011*, 1(DiiD), pp. 68–73, 2011.

[5] H. Haddad Pajouh, R. Javidan, R. Khayami, D. Ali, and K.-K. R. Choo, "A two-layer dimension reduction and two-tier classification model for anomaly-based intrusion detection in IoT backbone networks," *IEEE Transactions on Emerging Topics in Computing*, vol. PP, issue 99, pp. 1–1, 2016.

[6] C. Huijun, S. Hong, and Z. Hong, "Early recognition of Internet service flow," *Proceedings - 2013 Wireless and Optical Communications Conference (WOCC)*, pp. 464–468, 2013.

[7] K. J. Chabathula, C. D. Jaidhar, and M. A. Ajay Kumara, "Comparative study of Principal Component Analysis based Intrusion Detection approach using machine learning algorithms," *3rd International Conference on Signal Processing, Communication and Networking (ICSCN)*, pp. 1–6, 2015

[8] P. M. Mafra, V. Moll, J. Da Silva Fraga, and A. O. Santin, "Octopus-IIDS: An anomaly based intelligent intrusion detection system," *Proceedings - IEEE Symposium on Computers and Communications*, pp. 405–410, 2010.

[9] A. R. Onik, N. F. Haq, and W. Mustahin, "Cross-breed type Bayesian network based intrusion detection system (CBNIDS)," *18th International Conference on Computer and Information Technology (ICCIT)*, pp. 407–412, 2015.

[10] Divyatmika and M. Sreekesh, "A two-tier network based intrusion detection system architecture using machine learning approach," *International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, pp. 42–47, 2016.

[11] P. Singh and A. Tiwari, "An efficient approach for intrusion detection in reduced features of KDD99 using ID3 and classification with KNNGA," *Proceedings 2nd IEEE International Conference on Advances in Computing and Communication Engineering (ICACCE)*, pp. 445–452, 2015.

[12] Z. Wang, F. Zou, B. Pei, W. He, L. Pan, Z. Mao, and L. Li, "Detecting malicious server based on server-to-server realation graph," *IEEE First International Conference on Data Science in Cyberspace (DSC)*, pp. 698–702, 2016.

[13] Z. Yongli, Z. Yungui, T. Weiming, and C. Hongzhi, "An improved feature selection algorithm based on MAHALANOBIS distance for network intrusion detection," *International Conference on Sensor Network Security Technology and Privacy Communication System (SNS & PCS)*, pp. 69–73, 2013.

[14] M. Xie and J. Hu, "Evaluating host-based anomaly detection systems: A preliminary analysis of ADFA-LD," *6th International Congress on Image and Signal Processing (CISP)*, pp. 1711–1716, 2013.

[15] H. Yu, P. P. K. Chan, W. W. Y. Ng, and D. S. Yeung, "Apply randomization in KNN to make the adversary harder to attack the classifier," *International Conference on Machine Learning and Cybernetics (ICMLC)*, pp. 179–183, 2010.