Anomaly Detection Using Statistical Learning In Cloud Server System

Dhanashri A. Hase
Dept.of Computer Engineering,
Amrutvahini College of Engineering,
Sangamner,India.

Shrinivas K. Sonkar
Dept.of Computer Engineering,
Amrutvahini College of Engineering,
Sangamner,India.

***Abstract*— As a prime strategy to make certain the safety of IT infrastructure, anomaly detection plays a greater vital position in cloud computing platform which hosts the entire applications and information. On pinnacle of the traditional Markov chain version, we proposed on this paper possible multi-order Markov chain based totally framework for anomaly detection. In this approach, both the high-order Markov chain and multivariate time collection are followed to compose a scheme described in algorithms at the side of the schooling manner in the form of statistical mastering framework. To cut down time and area complexity, the algorithms are designed and carried out with non-zero cost table and logarithm values in initial and transition matrices. For validation, the collection of system calls and the corresponding return values are extracted from conventional Defense Advanced Research Projects Agency (DARPA) intrusion detection assessment facts set to shape a -dimensional take a look at input set.In this paper using spatio-temporal for they have been analyzed for various types of human activity including execution of political actions, disaster management, crime prevention, emergency services, etc . Such data can be regarded as a collection of spatio-temporal documents.**

Keywords—Kth-order Markov chain, multivariate time series, anomalydetection, statistical learning.

I. INTRODUCTION

Cloud computing comes with in dispensable dependency on networked laptop system. Unfortunately, while everyone knows there is no guarantee of its wellness; we have a tendency to definitely ignore this painful idea. An increasing number of academicals and commercial customers are beginning to rely completely on cloud computing servers that hostess entire packages and garage. In fact, those cloud computing and offerings within the form of dispensed and open shape end up apparent targets for ability threads. Thus, looking after both business and personal facts, servers exposed essential protection and availability problems. Their invulnerability are of degree significance to each people and the society. However, at some point of catastrophic failures inclusive of intrusion, crash or breakdown, the anomaly should be first determined earlier than any real remedy could come to its useful resource. Being recessive on the early stage, such problems timber not shows off wonderful trends and regularly leads to behind schedule responses and irrecoverable effects. Huge quantities of statistics with each spatial and temporal information (e.g., geo-tagged tweets) are being generated, and are frequently used to proportion and unfold personal updates, spontaneous thoughts, and breaking information. We talk over with such records as spatio-temporal documents. It is of splendid interest to discover subjects in a group of spatio-temporal documents. we have look at the trouble of effectively mining topics from spatio-temporal files within a person detailed bounded place and timespan, to offer customers with insights approximately events, tendencies, and public concerns inside the exact place and time length. We advocate a unique algorithm that is capable of efficaciously integrate two pre-skilled subject matter fashions learnt from two file sets with a bounded error, based on which we develop an green method to mining topics from a massive range of spatio-temporal documents within a area and a timespan. Our experimental effects display that our technique is capable of enhance the runtime through at least an order of importance in comparison with the baselines. Meanwhile, the effectiveness of our proposed approach is near the baselines.

II. LITERATURE SURVEY

Dorothy E. Denning works on A version of a real-time intrusion-detection professional machine able to detecting break-ins, penetrations, and other kinds of laptop abuse is described. The model is based totally at the speculation that security violations can be detected by using tracking a machine's audit facts for extraordinary patterns of system utilization. The model includes profiles for representing the conduct of subjects with respect to items in terms of metrics and statistical fashions, and guidelines for acquiring know-how approximately this behavior from audit information and for detecting anomalous conduct. The version is independent of any particular machine, software surroundings, system vulnerability, or kind of intrusion, thereby providing a framework for a fashionable-motive intrusion-detection expert system.

Nong Ye works on An anomaly detection technique to discover intrusions into pc and community structures. In this technique, a Markov chain version is used to represent a temporal profile of normal behavior in a pc and network gadget. The Markov chain version of the norm profile is found out from ancient information of the gadget's ordinary behavior. The observed conduct of the machine is analyzed to infer the possibility that the Markov chain model of the norm profile supports the located conduct. A low opportunity of aid indicates an anomalous conduct that might also end result from intrusive sports. The approach turned into carried out and examined on the audit statistics of a Sun Solaris gadget. The checking out results confirmed that the method virtually distinguished intrusive activities from ordinary sports within the checking out data.

Wenyao Sha, Yongxin Zhu, Tian Huang, Meikang Qiu works on a possible multi-order Markov chain based scheme for anomaly detection in server structures. In our method, each the high-order Markov chain and multivariate time collection are taken into consideration, along side the designated layout of training and checking out algorithms. To compare its effectiveness, the Defense Advanced Research Projects Agency (DARPA) Intrusion Detection Evaluation Data Set is used as stimuli to our version, through which device calls and the corresponding go back values form a -dimensional input set. The calculation end result shows that this approach is able to produce several effective indicators of anomalies. In addition to absolutely the values given by way of an individual single-order model, we also note a novelty remarkable earlier than, i.e., the modifications in rating positions of outputs

from special-order ones also correlate carefully with ordinary behaviours. Moreover, the analysis and alertness proves our method's efficiency in eating reasonable cost of time and garage.

III. PROPOSED METHODOLOGY

A. The Model Of Markov Chain

From the schooling sequence Wn ∈ W,n=1,2,….Nw, we can derive both the initial opportunity distribution matrix Q, wherein each element qi represents the preliminary possibility of the corresponding state Xi:

Q =[q1…qi…. qm} ; in which qi =P(Wn =Wi) ; and the one-step transition possibility distribution matrix P, wherein each element pij represents the transition opportunity from Xi to Xj:

P=[pij] m× m ; where pij=[Wn=Wj | Wn-1 = Wi) The version is completely specified as soon as P and Q is given, for this reason we denote it as π(P,Q)

Statistical Learning Framework :

State space W ={W_1,W_2,….,W_m} construct first order markov chain transition space W^1 :

W^1={[W_1 W_1],[W_1 W_2],[W_1 W_3],……[W_m W_m]},

binary state space Z:

Z={ normal , abnormal }

So the whole set of enter (input) data I can be represented as coordinates mapping from space W^1 to space Z

I = {(w^1_1,z_1),(w^1_2,z_2)….,(w^1_{Nw}-1,z_{Nx-1})}

Where w_n^1 ∈ W^1 and z_n ∈ Z

Where series w_n^1

$$w_n^1 = [x_x \; x_{n+1}] \in W^1, n=1,2,3,....,N_x-1.$$

From the enter dataset I given the model π(P,Q)

P(I | π(P,Q))

Probability of dataset I the model π(P,Q) and it can calculated

$$P(I \mid \pi(P,Q)) = \{q_{w1}, \quad Nw = 1$$
$$\{q_{w1} \pi^{Nw}_{n=2} P_{wn-1}{}^{wn} \quad N_x > 2, \ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

Example:

Training and Testing is shown step by step.

10:09:20	10:09:21	10:09:22	10:09:24	10:09:26
kill()	fork()	kill()	fork()	open()
success	failure	success	failure	failure
		normal		

Fig. 1. System call series of Virtual Machine 1 (training set).

10:09:20	10:09:21	10:09:22	10:09:24	10:09:26
fork()	fork()	kill()	open()	open()
failure	failure	success	failure	success
		normal		

Fig. 2. System call series of Virtual Machine 2 (training set).

Training data set I, system calls and return values shown in fig.1 and fig.2

TABLE 1

Mapping from [RV_1 SC_1 RV_2 SC_2] to W

[RV_1 SC_1 RV_2 SC_2]	W
[a_1 b_1 a_1b_1]	W_1
[a_1 b_1 a_1b_2]	W_2
[a_1 b_1 a_1b_3]	W_3
...	...
[a_2 b_3 a_2 b_1]	W_{34}
[a_2 b_3 a_2 b_2]	W_{35}
[a_2 b_3 a_2 b_3]	W_{36}

The two state space A and three state space B obtained as:

A={success , failure} = {a1,a2},

B={open(),kill(),fork()}={b1,b2,b3}.

The schooling sequences of return values and system calls derived immediately from fig.1 and fig.2 R_1,S_1,R_2,S_2 ,:

$R_1 = a_1,a_2,a_1,a_2,a_2$

$S_1=b_2,b_3,b_2,b_3,b_1$

$R_2= a_2,a_2,a_1,a_2,a_1$

$S_2=b_3,b_3,b_2,b_1,b_1$

we create the desk for mapping from the input multivariate collection to a univariate one as proven in Table 1,and assemble a univariate equivalent collection wn

w_n = { a_1,b_2,a_2,b_2} , { a_2,b_3,a_2,b_3},
{ a_1,b_2,a_1,b_2} , { a_2,b_3,a_2,b_1}

{ a_2,b_1,a_1,b_1}=$W_{12},W_{36},W_8,W_{34},W_{19}$

Mapping in Table2:

w^*_n={ W_{12},W_{36}},{ W_{36},W_8},{ W_8,W_{34}},{ W_{34},W_{19}}
=$W^*_{432},W^*_{1268},W^*_{286},W^*_{1207}$

w^k_n=[$W^*_{432,}$ W_8],[$W^*_{1268,}$ W_{34}],[$W^*_{286,}$ W_{19}]

Intial matrix Q^* with four nonzero elements

($q^*_{286} = q^*_{432} = q^*_{1,207} = q^*_{1,268} = 1/4$):

Q^*=[… ¼ … ¼ …1/4 … ¼ ..]$_{1\times1,296}$

TABLE 2
Mapping from [x_n x_n] to **W**

[x_n x_n]	**W**
[W_1 W_1]	W^*_1
[W_1 W_2]	W^*_2
[W_1 W_3]	W^*_3
.....	
[W_{36} W_{34}]	W^*_{1294}
[W_{36} W_{35}]	W^*_{1295}
[W_{36} W_{36}]	W^*_{1296}

10:19:30 fork() failure	10:19:31 kill() success	10:19:32 fork() failure

Fig. 3. System call series of Virtual Machine 1 (test set 2).

10:19:30 fork() failure	10:19:31 kill() success	10:09:32 open() failure

Fig. 4. System call series of Virtual Machine 1 (test set 1).

Transition matrix P^* three non-zero element
($P^*_{432,8} = P^*_{1,268.34} = P^*_{286,19}=1$):

$$P^* = \begin{bmatrix} 1 & \cdots & ... \\ ... & 1 & ... \\ ... & \cdots & 1 \end{bmatrix}_{1,296\times36}$$

$R'_1 = a_2,a_1,a_2$

$S'_1=b_3,b_2,b_3$

$R'_2= a_2,a_1,a_2$

$S'_2=b_3,b_2,b_1$

Similarly,

$w'_n = \{ a_2,b_3,a_2,b_3\} , \{ a_1,b_2,a_1,b_2\}, \{ a_2,b_3,a_2,b_1\}$
$=W_{36}, W_{36},W_8$

$w'^*_n=\{ W_{36},W_8\},\{ W_{36},W_8\}= W^*_{1268,}\ W^*_{286}$

$w'^k_n=[W^*_{1268,}\ W_{34}]$

10:59:30	10:59:31	10:59:32
open()	fork()	kill()
success	success	failure

Fig. 5. System call series of Virtual Machine 2 (test set 1).

10:59:30	10:59:31	10:59:32
kill()	fork()	open()
failure	success	success

Fig. 6. System call series of Virtual Machine 2 (test set 2).

According to equation (1),the probability of W'^* given model $\pi(P^*,Q^*)$:

$$P(I' \mid \pi(P^*,Q^*)) = q^*_{1,268} \times P^*_{1268,34}=1$$

On the alternative hand, for newly-found series proven in
Figs. 5 and 6, by applying the identical methodology as take a look at set 1 and a couple of, their probability of occurrence given the model $\pi(P^*,Q^*)$equals 0. As a result, the latter test set is anticipated to be greater "atypical" than the previous one since the latter check set has a decrease opportunity of assist than the former one does, which means that choice can be made in step with the opportunity of support. In addition, the ordinal indices of the brand new country area could comply with any mapping aside from Tables 1 and 2 as long as it keeps uniqueness.

C. Architecture

In this paper proposed a multi-order Markov chain based anomaly detection system. By observing the relative relations between results from the different-order models which gives another powerful indicator of anomalies. In general, because of the regular and periodical behaviors of cloud server systems, if the probability of test set given the lower-order model surpasses that given the higher-order one, it is inferred that irregular events may have happened in the system and further considerations or activities would be vital. In addition, combining multi-dimensional inter related sequences as a multivariate one into a single model would be another attainable way to deal with enhance the affectability of detection. As appeared, the return value series can be a valuable supplement to the system call series utilized the conventional behavior. Likewise, with both time and space effectiveness of the Training and Testing algorithm, this methodology limits the possibility of turning into the anomalies itself and is completely equipped for on the web (or constant) recognition. The time utilization of the training stage takes close to 15 seconds for a training set as vast as 1.6 million, and for models up to the third-order combined. To additionally enhance effectiveness there are different ways, for example, proportionate space development by including artificial state, or binary representation for sparse matrix could essentially relieve the space complexity issue. Further efforts may include the time-in homogeneous Markov chain, which may have the capacity to build transition matrix independently for a particular time period, as the time homogeneous assumption stated here could appear to be excessively restrictive for time-sensitive systems and anomalies. Model aggregations techniques may also be significant in combining multi-order restrictive probabilities into one single value for quicker decision-making. The Proposed framework comprise of four modules depicted as follows:

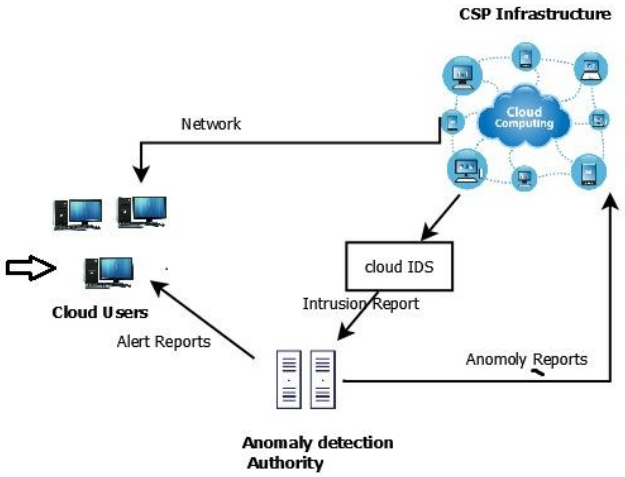


Figure 1: Anomaly Detection in Cloud Server System

1. Anomaly Detection

In these methodologies, anomaly or intrusion detection systems usually watch behaviors of watched items to involve statistical distributions as a set of trained profiles during the training stage. These systems at that point apply the set of trained profiles by comparing them against another set of profiles of watched objects during the detection stage. An anomaly or intrusion is identified if these two sets of profiles don't coordinate.

2. Anomaly Prevention

The anomaly prediction model combines attribute value prediction and multivariate anomaly characterization to raise early alerts. Most past learning-based anomaly detection schemes treat an distributed application as one solid entity, which experience two major issues. Initially, it can't recognize which segments are credited to the performance anomaly. The prediction accuracy of one solid model is fundamentally more regrettable than that of per-component model since the attribute value prediction mistakes collect as it incorporate the attributes of all segments into one model.

3. Validation Process Based On Markov Chain Model

As a real-world application of the order Markov chain anomaly detecting system which take in the great Defense Advanced Research Projects Agency (DARPA) Intrusion Detection Evaluation Data Set by the Cyber Systems and Technology Group of MIT Lincoln Laboratory to check this scheme. The DARPA dataset were gathered since June 1998 when DARPA led a seven week simulation of TCP attacks (anomalies) out of sight stream of typical and normal user activities. Real outcomes for experiment were accomplished with this data set. The rest data sets in past writing were either inaccessible to the general public or even older than DARPA data set. Consequently, it simply pick DARPA data set.

4. Attribute based File Sharing

This strategy for secure file sharing utilizing Attribute based File Sharing. Enterprises in general store information in internal storage and introduce firewalls to secure against gatecrashers to access the information. They additionally standardize information access strategies to prevent insiders to reveal the data without permission. In this strategy it will check the properties of the users whether the receiver have the same characteristics as the sender referenced. It will stay away from the unauthorized users or programmers. The sender gives the attributes of the recipient while sending the record to the receiver; the file gets encrypted according to the given characteristics. The receiver gets the encrypted file, and he has given the attributes, if its right, the first file gets decrypted for the receiver.

D.What Type of Anomaly is Detected

Anomaly detection means watch the behavior of the observed users with trained profiles if these two sets of profiles do not match then detected behavior is abnormal.

E. Algorithms

To formulate our set of rules, we follow the method of anomaly detection work flow comprising of schooling stage and take a look at level. On pinnacle of the standard training and checking out ranges, we similarly tackle the problems of model order choice, zero probability event remodels and sparse matrix garage and many others. Our answers to those issues can be defined earlier than the whole algorithm is formulated.

The purpose of algorithm training () and testing() the first in training phase train the load data set and in testing phase in loaded data set finding the abnormal behavior or anomaly.

Algorithm 1.Training()

Input: Training sequence $w_n \in W$ (n=1→Nw), Number of states m, Order K

Output: Initial probability distribution matrix Q, Transition probability distribution matrix P

1: Initialize()

2: for n = K →Nw do

3: x*n←[wn-K+1 wn-K+2 wn-K +3 … wn]

// Construct new sequence in space W*
4: end for
5: for n = K →Nw-1 do
6: $w^k_n \leftarrow [w_n^* \ w_{n+1}]$
// Construct new sequence in space WK
7: end for
8: for n = K →Nw do
9: if w_n= wi then
10: Increase Initial Matrix(i)
// Get the index of wn and increase q_i in Q by 1
11: end if
12: end for
13: for n = K →Nw do
14: if w^K_n=[W^*_i W_j] then
15: Increase Transition Matrix(i,j)
// Get the index of wKn and increase p_ij in P by 1/
16: end if
17: end for
18: Normalize Initial Matrix()
19: Normalize Transition Matrix()

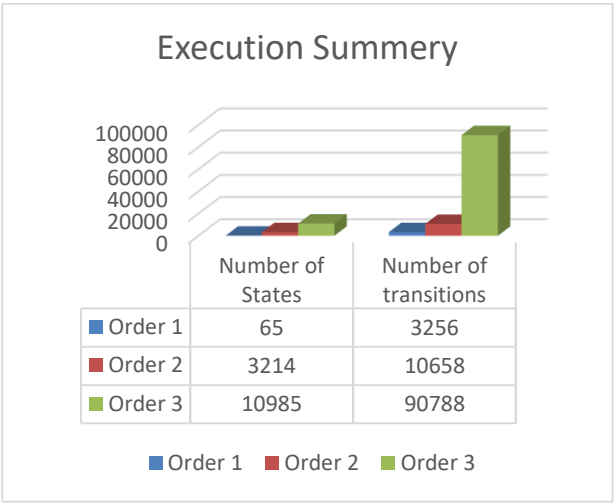
Algorithm 2.Testing()

Input: Testing sequence $z_n \in$ W,Q, P
Output: Probability of zn given Q and P
1: for n = K →Nz do
2: $z^*_n \leftarrow$[zn-K+1 zn-K+2 zn-K +3 … zn]
3: stop for
4: for n = K →Nz -1 do
5: $z^k_n \leftarrow [z_n^* \ z_{n+1}]$
6: quit for
7: for n = K →Nz do
8: if Zn = Xi then
9: probability ← Get Initial Matrix(i)
10: cease if
11: stop for
12: for n = K →Nz-1 do
13: if z^k_n =[W_i^* W_j] then
14: probability ← possibility × Get Transition Matrix(i,j)
15: end if
16: cease for
17: Output : probability

IV. RESULT AND DISCUSSIONS

In proposed system, experiment summery shows the summery of data which can be processed to detect the anomaly on cloud. While processing the data on cloud, user's system call series can be considered to detect the anomaly on cloud.

Using the two algorithm Training and Testing in training phase train the load data set and in testing phase in loaded data set finding the abnormal behavior or anomaly.



V. CONCLUSIONS

In this paper anomaly detection scheme based totally on multi order Markov chain by means of staring at that the relative positions between consequences from models of different orders provide a brand new powerful indicator for anomalies and one more concept is spatio-temporal for finding the location. In general, because of the normal and periodical behaviors of cloud server systems, if the probability of test set given the lower-order model surpasses that given the higher-order one, it is suggested that uncommon events may have happened in the system and further considerations or activities would be vital. In addition, combining multi-dimensional between related sequences as a multivariate one into a single model would be another feasible way to enhance the affectability of detection. Moreover, with both time and space productivity of the Preparing and Testing algorithm, this methodology limits the possibility of getting to be the source of anomalies itself and is completely equipped for on the web (or continuous) discovery.

REFERENCES

[1] Denning, D.E., "An intrusion-detection model," IEEE Transactions on Software Engineering. pp. 222 - 232, Feb. 1987
[2] Yihua Liao, V.Rao Vemuri, "Use of K-nearest Neighbor Classifier for Intrusion Detection," Computers & Security. 21(5): 439-448, 2002
[3] S. Saravanakumar, Amruth Kumar.A, Anandaraj.s, s.Gowtham, "Algorithms Based on Artificial Neural Networks for Intrusion Detection in Heavy Traffic Computer Networks," Proc. of Int'l Conf. on Advancements in Information Technology, pp.6-23, 2011
[4] C.V. Raman, Atul Negi, "A Hybrid Method to Intrusion Detection Systems Using HMM," Distributed Computing and Internet Technology.pp. 389 - 396, 2005
[5] Nong Ye, "A Markov Chain Model of Temporal Behavior for Anomaly Detection," Workshop on Information Assurance and Security. West Point, NY, June 2000.
[6] Nong Ye, Xiangyang Li, Qiang Chen, Syed Masum Emran, Mingming Xu, "Probabilistic Techniques for Intrusion Detection Based on Computer Audit Data," IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans. 31(4):266-274, July 2001
[7] Nong Ye, Yebin Zhang, Connie M. Borror, "Robustness of the Markov Chain Model for Cyber-Attack Detection," IEEE Transactions on Realiability.53(1):116-123, March 2004
[8] Robert Gwadera, Mikhail Atallah, "Markov Models for Identification of Significant Episodes," Proceedings of the 5th International Conference on Data Mining. 2005
[9] Wen-Hua Ju, Yehuda Vardi, "A Hybrid High-Order Markov Chain Model for Computer Intrusion Detection," Journal of Computational and Graphical Statistics. 10(2): 277-295, 2001
[10] Chuanhuan Yin, Shengfeng Tian, Shaomin Mua, "High-order Markov kernels for intrusion detection," Neurocomputing. 71(16-18):3247-3252,2008
[11] Markus Stowasser, "Modelling Rain Risk: A Multi-order Markov Chain Model Approach", The Journal of Risk Finance, 13(1):45 - 60, 2011
[12] Stephanie Forrest, Steven A, Hofmeyr, Anil Somayaji, Thomas A,Longstaff, "A Sense of Self for Unix Processes," Proceedings of IEEE Symposium on Security and Privacy, pp. 120 - 128, 1996
[13]W. Sha, Y. Zhu, T. Huang, M. Qiu, Y. Zhu, and Q. Zhang, "A multi-order markov chain based scheme for anomaly detection,"in Proc. IEEE 37th Annu. Comput. Softw. Appl. Conf.Workshops,2013,pp.83–88