Study on HMM Based Anomaly Intrusion Detection Using System Calls

SHI Shang-zhe, SUN Mei-feng

Information and Engineering College, Yangzhou University, Yangzhou Jiangsu 225127

Keyword: anomaly detection; system call; HMM; hidden state

Abstract. In order to improve the detection accuracy, we study on the HMM model based on system calls anomaly detection. We starting from the program semantics issued system call, analysis that the state hidden behind system calls is the program execution state. Then put forward that hidden state must greater than the number of unique system calls when training HMM. And observation probability can be as 01 vector form. HMM trained in our way is better than other models on detection accuracy.

Introduction

It is urgent needs anomaly detection technology that captures of new intrusions in time, because of intrusions rapidly developed and updated frequency. But there is a problem that anomaly detection has a high false alarm rate, so we need an express model of normal behavior. In the kinds of anomaly detection technology, establish normal model using system calls put forward by Forrest in 1996[1] has been focus on. It is distinguish intrusions by monitor the sequence of system calls caused by process. The sequence of system calls caused by process is stable because the code of process is determinate, but it will be changed by intrusions. So it has good detection result to detection intrusions by monitored system calls. There is many methods of model been suggested, HMM is one of them. [2]

HMM is an extension of Markov chain, which includes double stochastic process: one is Markov process express by states which are hidden; the other is observations sequence associate with states. In each state, there will be observations in accordance with the specific probability distribution which can be observed.[3] HMM has been successfully used in the research of biomedical and speech recognition because of its powerful ability of expression.[4,5] Literature [2] introduced HMM into anomaly detection for the first time, which used sequence of system calls to build normal model of process, and experiments show that HMM achieved best detection result in short sequence model, frequency model and data mining. From that on, HMM used in anomaly detection with system calls has been a hotpot of research until now. [6~13]

According to the theory of HMM, it has to set the number of states before training. Literature [2] proposed that this number should be the number of unique system calls in sequence, for that has better result in experiment. Literature [6] used 10~40 training model many times to find a best result, and proposed that: ”we found that the number of states is a sensitive parameter after experiment, but there is still not a theory that how to set this parameter, only way to determine this parameter is through many experiments”. Follow-up study on this has that accordance with Literature [2] used the number of unique system calls in sequence, [7,8] Literature [9] take 5~15 as the number of states, Literature [10] take the length of sliding window as the number of states and used 2 in their paper. In the recent years, many researches focus on the hybrid method to training model, so there is still not good solution on this problem.

This paper research on the semantic of program that system calls caused by, try to explain the meaning of hidden states in HMM based on anomaly detection using system calls. So we can determine the number of hidden states and to build a better HMM.

In the second section we introduce the concept of HMM and correlation algorithm. In the third section we propose a model of HMM which has been discussed how to confirm the number of hidden states. In the fourth section we detective anomaly with that HMM and analysis the

Published by Atlantis Press, Paris, France.

experiment result. The fifth section is conclusion and future work.

Concept of HMM and algorithm

Definition.HMM can be denoted as triples $\lambda = (A, B, \pi)$:

(1) Transition probability $a_{ij} = P(q_j \mid q_i)$; which is the probability of state q_i shift to state q_j. If there are N states, there would be N×N transition probabilities. A is matrix of them.

$$A = \lfloor a_{ij} \rfloor \text{ AND } \sum_{j=1}^{N} a_{ij} = 1$$

(2) Observations probability $b_j(k) = P(o_k \mid q_j)$; which is the probability of Observation o_k under the state of q_j. If there are M observations, there would be N×M probabilities. B is matrix of them.

$$B = \lfloor b_j(k) \rfloor \quad \text{AND} \quad \sum_{k=1}^{M} b_j(k) = 1$$

(3) Initial state probability π: which means the probability of occurrence state q_i (i=1,2,…,N) when t=0.

Three classic questions of HMM:

(1) Evaluation: HMM $\lambda = (A, B, \pi)$ and the sequence of observations $O = [o_1, o_2, ..., o_T]$ has been known, beseech the probability of this sequence of observations produced by this HMM, which is $P(O \mid \lambda)$.

(2) Decoding: HMM $\lambda = (A, B, \pi)$ has been known, beseech the sequence of hidden states which can be produce a specific sequence of observations.

(3) Learning: training a best HMM with a sequence of observations has been known.

Algorithm. It is involves two questions when we used HMM in anomaly detection in this paper: Evaluation and Learning.

(1) Algorithm of Evaluation:

This is a recursive algorithm, which has forward recursive, backward recursive and forward-backward recursive. Table 1 is algorithm of forward recursive:

Table 1 forward recursive

Take $\alpha_t(\mathrm{i})$ as the probability of part sequence of observations $o_1 \sim o_t$ produce by model λ at moment t under state q_i:

$\alpha_t(i) = P(o_1, o_2, ..., o_t, s_t = q_i \mid \lambda)$

First step: initialization

$\alpha_1(i) = \pi_i b_i(o_1)$

Second step: recursive

$for..t = 1..to..T - 1$

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij} \right] b_j(o_{t+1})$$

Third step: end

$$P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

(2) Algorithm of Baum-Welch(Learning)

Algorithm of Baum-Welch is more complicated, it is need a variable of backward recursive $\beta_t(j)$ except variable of forward recursive $\alpha_t(i)$ used before. $\beta_t(j)$ means the probability of part sequence of observations $o_{t+1} \sim o_T$ produce by model λ at moment t under state q_j.: $\beta_t(j) = P(o_{t+1}, ..., o_T \mid s_t = q_j, \lambda)$.

There are need other two auxiliary variables:

(1) $\zeta_t(i,j) = P(i_t = q_i, i_{t+1} = q_{i+1} \mid O, \lambda) = \dfrac{P(s_t = q_i, s_{t+1} = q_j, O \mid \lambda)}{P(O \mid \lambda)}$

$$= \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}$$

(2) $\gamma_t(i) = \sum_{j=1}^{N} \zeta_t(i,j)$

It can be got that the probability of state i to state j and the expected probability of in state i when cumulative the two auxiliary variables with time t. After that, we can get a new HMM $\overline{\lambda} = (\overline{A}, \overline{B}, \overline{\pi})$:

$$\overline{\pi_i} = \gamma_1(i) \tag{1}$$

$$\overline{a_{ij}} = \frac{\sum_{t=1}^{T-1} \zeta_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{2}$$

$$\overline{b_j(k)} = \frac{\sum_{t=1 \text{且} o_t = v_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \tag{3}$$

Then compare $P(O \mid \lambda)$ with $P(O \mid \overline{\lambda})$ to discuss which HMM is better and judge that model need for revaluation or not.

HMM based anomaly detection using system calls

It is take process behavior as monitor objects in anomaly detection using system calls. Although process behavior expressed as system calls and intrusions also achieved by system calls, but process behavior is formed by the operation of the program in fundamentally. It is take operating position, type of operation and operation environment as program execution states in OS, so we consider that process behavior is the shift of program execution states. Program execution states cannot be seen, but system calls associated can be observed. So the process behavior has the feature of HMM, and we can take program execution states as the hidden states in HMM.

Hidden states. According to our analysis, take program execution states as the hidden states in HMM when based anomaly detection using system calls. One program execution state can cause only one system call, but one system call can be caused by many program execution states. So it can be inferred that the number of hidden states should be more than the number of unique system calls in sequence. It can be assumed that the HMM will be most accurate when the number of hidden states matched with program execution states, and it will reduce the precision when the number of hidden states get away from the number of program execution states.

Figure 1 is our experiment result. The data set of experiment is system calls data set from university of New Mexico (http://www.cs.unm.edu/~immsec/systemcalls.htm). We used sequence of lpr system calls. This sequence contains 2398 system calls and belongs to 9 processes, and the number of unique system calls is 37.

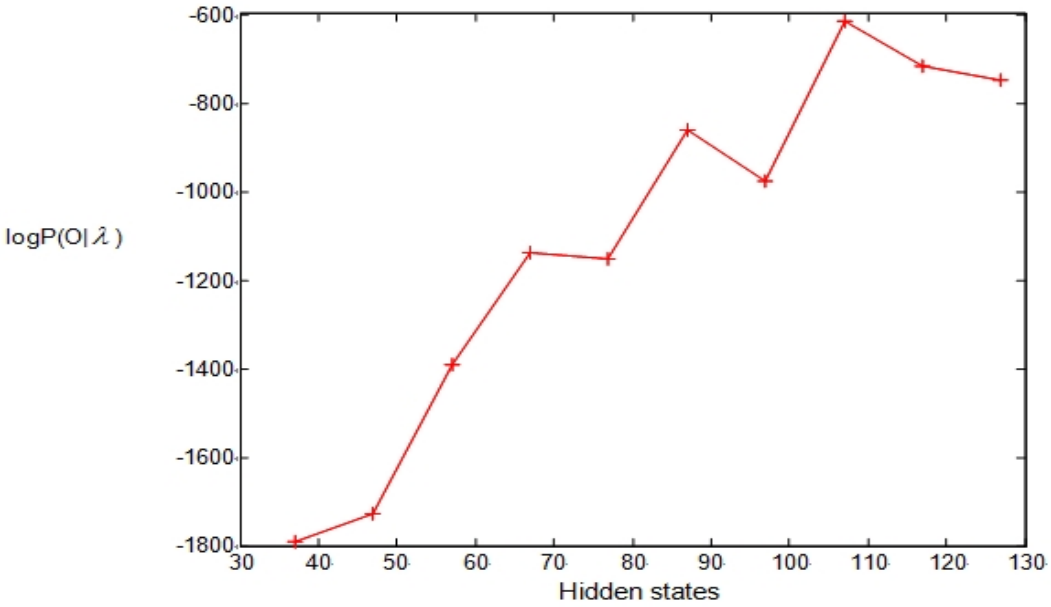

Figure 1 $\log P(\mathrm{O} \mid \lambda)$ change with different hidden states

In the experiment, we training HMM from take the number of unique system calls as the number of hidden states and calculate the $\log P(\mathrm{O} \mid \lambda)$, then we increase the number to training HMM again and calculate the $\log P(\mathrm{O} \mid \lambda)$ again until after a number that reduce the $\log P(\mathrm{O} \mid \lambda)$ twice. So we chose this number as the best number of hidden states, and HMM under this condition is the model we need.

Probability of observations. When the number of hidden states is the number of program execution states, one hidden state correspond one system call. So the probability of observations under a state is the form like: (0,0,0,…,1,…,0,0,0). When we training HMM restrict matrix B as ((0,0,0,…,1,…,0,0,0),…, (0,0,0,…,1,…,0,0,0)). It can be reduce the training time under this condition.

Establishment of classifier. After training the HMM, we used a slide window which length is k to slip the normal sequence of system calls, then calculated the range of normal probability.

In the detection phase, using the same slide window to slip the monitored sequence of system calls, and then calculated the probability of the k length sequence, if this probability is out of the range of normal probability, it is consider that an exception occurs.

Experiments

It is used the normal data and intrusion data from Professor Stephanie Forrest in department of computer science in university of New Mexico. We use 80% of normal data to training HMM, the rest 20% and intrusion sequence as test data. 20% normal data to detective the rate of false positive, and intrusion sequence for detection rate.

In our experiment, we compare the detection result with HMM trained used the number of unique system calls as the number of hidden states to our HMM. The comparative results described with ROC（receiver operating characteristic）[14], just as figure 2. Every point in ROC represents detection and false alarm rate under a specific threshold. It will alarm when there are 2 exceptions occurs in frame of window in detection phase. Detection rate is the number of alarms divided by the number of frame of intrusion data. And false alarm rate is the number of false alarms divided by the number of frame of normal test data. It can be get a ROC when change the threshold for many times. Obviously, if our HMM is more accurate, the detection result will be better.

Published by Atlantis Press, Paris, France.

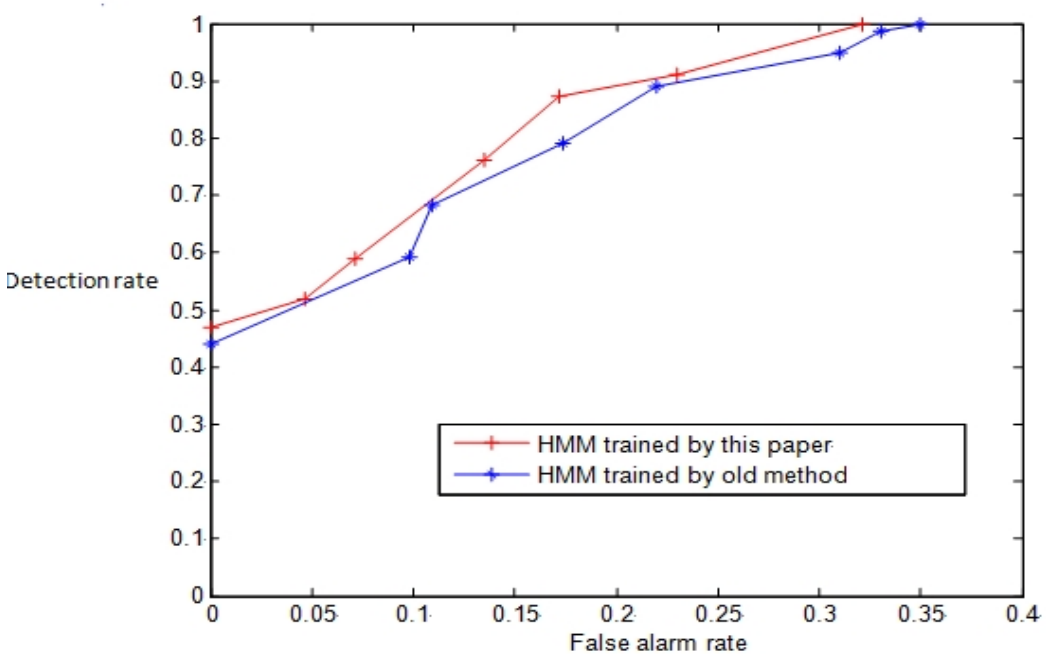

Figure 2 detection and false alarm rate under different HMM

Conclusions

The number of hidden states is a sensitive parameter when HMM used in anomaly detection. Different from the traditional approach to set this parameter as the number of unique system calls in sequence, we propose that this parameter should greater than the number of unique system calls in sequence by analysis the program semantic. And the probability of observations under states should be as the form as((0,0,0,…,1,…,0,0,0),…, (0,0,0,…,1,…,0,0,0)). The experiment prove that the HMM under our condition has better detection result.

References

[1] Forrest S., Hofmeyr S. A., Somayaji A., et al. A sense of Unix processes[C].Proceedings of the 1996 IEEE Symposium on Research in Security and Privacy. Oakland, California,1996.

[2] Christina Warrender, Stephanie Forrest and Barak Pearlmutter. Detectiong Intrusion Using System Calls: Alternative Data Models[C]. 1999 IEEE Symposium on Security and Privacy. IEEE Computer Society, 1999.

[3] L. R. Rabiner and B. H. Juang. An introduction to hidden Markov models. IEEE ASSP Mag, pp4-16, Jun 1986.

[4] Burrat C,Hughey R,Karplus K. Scoring hidden Markov models [J]. Computer Application in Bioscience, 1997,13:191-199.

[5] Cohen A. Hidden Markov models in biomedical signal processing[C]. 20th International conference EM-BS/IEEE. Hongkong,1998.

[6] Bin, Y., Qiao, Y., Xin, X.W., et al. Anomaly intrusion detection method based on HMM[J]. IEEE Electronic letters Online No: 20020467. 2002,38: 663–664.

[7] Bo Gao, Hui-Ye Ma,Yu-Hang Yang. HMMs based on anomaly intrusion detection method[C]. Proceedings of First International Conference on Machine Learning and Cybernetics. Beijing, 2002.

[8] Hu, J., Hoang, X.D., and Bertok, P. A multi layer model for anomaly intrusion detection using program sequences of system calls[C]. IEEE International Conference on Networks. 2003.

[9] Sung-Bae Cho and Hyuk-Jang Park. Efficient anomaly detection by modeling privilege flows using hidden Markov[J].Computers & Security, 2003, 22 (1): 45-55.

[10] C.V. Raman and Atul Negi. A Hybrid Method to Intrusion Detection Systems Using HMM[C].Distributed Computing and Internet Technology: Second International Conference, ICDCIT 2005. Bhubaneswar,India,2005.

[11] Wael Khreich, Eric Granger, Robert Sabourin,et al. Combining Hidden Markov Models for Improved Anomaly Detection[C]. IEEE International Conference on communications. Dresden, Germany,2009.

[12] Jiankun HU, Xinghuo Yu, D. Qiu,et al. A Simple and Efficient Hidden Markov Model Scheme for Host-Based Anomaly Intrusion Detection[J]. IEEE Network, 2009,23: 42-47.

[13] Duan Xuetao, Jia Fuchun,Liu Chunbo. Intrusion detection method based on hierarchical hidden Markov model and variable-length semantic pattern [J]. Journal of China Institute of Communacaitions,2010，31(3):109~114.

[14] Fan J,Upadhye S,Worster A. Understanding receiver operating characteristic(ROC) curves[J].Canadian Journal of Emergency Medicine,2006,8(1):19-20.

Published by Atlantis Press, Paris, France.