Anomaly Detection Approach using Hidden Markov Model

Enkhjargal Dorj* and Erdenebaatar Altangerel*
*MUST, CSMS/Computer Science, Ulaanbaatar, Mongolia
enkhjargal@csms.edu.mn, erka@must.edu.mn

Abstract— **Anomaly detection is an important problem that has been researched within diverse research areas. Numerous methods and approaches based on Hidden Markov Model regarding the anomaly detection have been proposed and reported in the literature. However, the potential applications using Hidden Markov Model classification based anomaly detection technique have not yet been fully explored and still in its infancy. This paper investigates the capabilities the use of Hidden Markov Model in anomaly detection for discrete sequences.**

Index Terms- Hidden Markov Model; Anomaly detection; Baum-Welch algorithm; Data discretization; Likelihood;

I. INTRODUCTION

Anomaly detection is an important problem that has been researched within diverse research areas and applications such as intrusion detection for cyber-security, fraud detection for credit cards, insurance or health care, and fault detection in safety critical systems. Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior [1]. Sequence data is found in a wide variety of application domains such as intrusion detection, bio-informatics, weather prediction, and system health management. Anomaly detection for discrete sequences is a challenging task and it concerns exploiting the sequential nature of data to detect anomalies [2].

The state transition, which is hidden in the Hidden Markov Model, can be used to characterize the difference between normal behavior and anomalous behavior of system [3]. Srivastava [4] highlighted that HMM allows for the rapid assimilation, analysis, and discovery of system anomalies. Joshi and Phoha [5] proposed anomaly detection system, a predictive model capable of discriminating normal and abnormal behavior of network traffic. HMMs have been applied for sequence anomaly detection [6]. Zang et al. [3] presented hierarchal HMM for anomaly detection in security of network. Chandola et al. [1] noticed that key strength of HMM based anomaly detection is that they can model complex models. However, HMM based anomaly detection techniques have not yet been fully explored and still in its infancy.

This paper presents an anomaly detection approach using Hidden Markov Model (HMM). In Section 2, the mathematical background to Hidden Markov Model will be discussed. Section 3 will develop the HMM based anomaly detection approach. Section 4 deals with the results which we have received after applying the HMM based anomaly detection approach. Finally, the conclusions of this paper will be provided.

II. MATEMATICAL BACKGROUND TO HIDDEN MARKOV MODEL

Hidden Markov Models are powerful finite state machines that are widely used for sequence modeling. Hidden Markov model is a statistical model in which the system is assumed to be a Markov process with unobserved or hidden states. In an HMM, the state is not directly visible, but the output that is dependent on the state is observable. In defining HMMs, we will use similar notation as used by Rabiner [7]. We define a system at any time as one of a set of K hidden states, $S_1, S_2, \dots, S_K$. Depending on the set of probabilities associated with a state (called transition probabilities), the system may undergo the change of state at a discrete time. The time instants associated with state changes are represented as $t = 1,2, \dots, T$, and we denote the state at time t as q_t. The hidden state sequence can be denoted as $Q = \{q_1 q_2 \dots q_T\}$. The state transition probability matrix can be represented as A=$\{a_{ij}\}$, where $a_{ij} = P(q_t = S_j | q_{t-1} = S_i)$; $1 \leq i, j \leq K$; and $a_{ij} \geq 0$. The number of distinct observation symbols per state is M. Observation symbols correspond to the physical output of the system being modeled, which at any time can be one of a set of M observation symbols, $\{v_1, v_2, \dots, v_M\}$. We denote the observation at time t as O_t. The observation sequence can be denoted as $O = \{O_1 O_2 \dots O_T\}$. The probability of the observation symbol v_m having been produced from state j is, $B = \{b_j(v_m)\}$, where $b_j(v_m) = P(O_t = v_m | q_t = S_j)$, $1 \leq j \leq K$, and $1 \leq m \leq M$. The initial state distribution vector $\pi = \{\pi_i\}$, where $\pi_i = P(q_1 = S_i)$, $1 \leq i \leq K$. In compact form, the complete parameter set of the HMM can be represented as shown in Eq. (1):

$$\lambda = \{A, B, \pi\} \quad (1)$$

A. Likelihood of the observation sequence

The forward variable $\alpha_t(i)$ and backward variable $\beta_t(i)$ of the forward and backward algorithm are defined to evaluate the likelihood of the observation sequence and re-estimate of the model parameters for each of the iterations of the training. The forward variable $\alpha_t(i)$ is the probability of generating the partial observation sequence $O_1 O_2 \dots O_t$ at time t when system is in state S_i. The forward variable at time $t = 1$, $\alpha_1(i)$, is defined by Eq. (2) [7]:

$$\alpha_1(i) = \pi_i b_i(O_1), \quad 1 \le i \le K \tag{2}$$

The forward variable at time $t + 1$, $\alpha_{t+1}(j)$, can be calculated as shown by Eq. (3) [7].

$$\alpha_{t+1}(j) = b_j(o_{t+1}) \sum_{i=1}^{K} \alpha_t(i)\, a_{ij}\ , \quad 1 \le j \le K,\ 1 \le t \le T - 1 \tag{3}$$

The likelihood of the observation sequence, $P(O|\lambda)$, is the sum of the terminal forward variables $\alpha_T(i)$, as defined in Eq. (4) [7].

$$P(O|\lambda) = \sum_{i=1}^{K} \alpha_T(i) \tag{4}$$

The backward variable $\beta_t(i)$ is the probability of generating partial observation sequence $O_{t+1}O_{t+2} \ldots O_T$ at time t and in state S_i, given the model. The backward variable at time T, $\beta_T(i)$, is defined as shown by Eq. (5):

$$\beta_T(i) = 1, \quad 1 \le i \le K \tag{5}$$

The backward variable at time t, $\beta_t(i)$, can be solved inductively as follows [7]:

$$\beta_t(i) = \sum_{i=1}^{K} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j), \quad 1 \le i \le K,\ t = T - 1, T - 2, \ldots, 1 \tag{6}$$

B. *Re-estimations of model parameters*

The model parameters A, B, π need to be re-estimated in each of the iterations of HMM training, since the Baum-Welch algorithm is an iterative learning algorithm. In order to update the model parameters, variables $\xi_t(i, j)$ and $\gamma_t(i)$ need to be identified first. The variable $\xi_t(i, j)$ is the probability of being in state S_i at time t and in state S_j at time $t + 1$, given the observation sequence $O_1 O_2 \ldots O_T$ and defined as follows [7]:

$$\xi_t(i, j) = P\left(q_t = S_i, q_{t+1} = S_j \middle| O, \lambda\right) \tag{7}$$

$\xi_t(i, j)$ can be estimated using forward and backward variables, as shown by Eq. (8):

$$\xi_t(i, j) = \frac{\alpha_t(i)\, a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^{K} \sum_{j=1}^{K} \alpha_t(i) \cdot a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)} \tag{8}$$

The variable $\gamma_t(i)$ is the probability of being in state S_i at time t, given the observation sequence $O_1 O_2 \ldots O_T$ and model λ, defined as [7]:

$$\gamma_t(i) = P(q_t = S_i | O, \lambda) \tag{9}$$

The probability $\gamma_t(i)$ can be expressed in terms of $\alpha_i(t)$ and $\beta_i(t)$ as [7]:

$$\gamma_t(i) = \frac{\alpha_t(i) \beta_t(i)}{\sum_{i=1}^{K} \alpha_t(i) \beta_t(i)} \tag{10}$$

Now re-estimation of model parameters can be defined using α, β, γ and ξ variables. Initial probabilities are can be re-estimated as follows [7]:

$$\overline{\pi_i} = \gamma_1(i) \tag{11}$$

Transition probabilities are can be re-estimated as follows:

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{12}$$

Emission probabilities are can be re-estimated as follows:

$$\bar{b}_j(V_m) == \frac{\sum_{t=1\ s.t\ O_t = v_m}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \tag{13}$$

C. *Convergence criteria*

The training steps including likelihood estimation and re-estimations of model parameters have to be repeated convergence is achieved or maximum number of iterations is reached. Hence, convergence coefficient or maximum number of iterations can be used as termination criteria to stop training process. Convergence is achieved when classification results do not change for two consecutive iterations (the change in the likelihood for two consecutive iterations is less than some threshold and convergence is declared) [8]. This convergence coefficient depends on level of accuracy desired, and the resources available (e.g., time, cost).

III.NOMALY DETECTION APPROACH USING HIDDEN MARKOV MODEL

This section develops anomaly detection approach using Hidden Markov Model. The developed anomaly detection approach has two distinct phases: training and detection. In the training phase, HMMs are formulated that best describes the normal and anomalous behaviors in the system. In the detection phase, the unknown test observation sequences are then categorized as normal or anomalous.

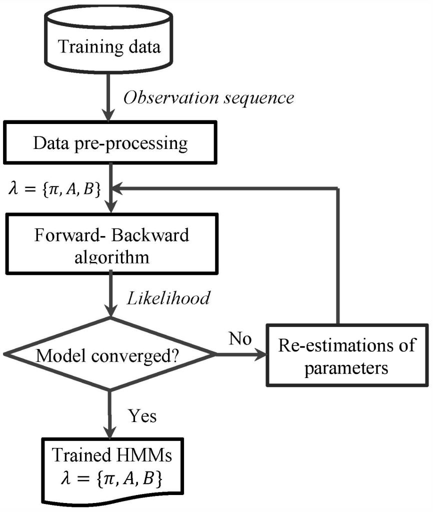
A. *Data pre-processing*

Hidden Markov Models are powerful finite state machines that are widely used for sequence modeling. Machine learning algorithms automatically extract knowledge from machinereadable information. Their success is usually dependent on the quality of the data that they operate on. Data pre-processing is an important step in the machine learning process. The pre-processing step is necessary to resolve several types of problems including noisy data, redundancy data, missing data values, etc [9][10]. Most of the existing machine learning algorithms are able to extract knowledge from data set store discrete features. If the features are continuous, the algorithms can be integrated with discretization algorithm which transforms them into discrete attributes.

In order to transfer continuous observation data into discrete sequences, data discretization is applied as a data pre-processing before training and detection phases. We discretized our data using the equal-width partitioning method that divides the range into M intervals of equal size, since the number of distinct observation symbols per state is M in our system. If X and Y are the lowest and highest values of the attribute, then the width of the intervals will be W = (Y - X)/M.

B. Training phase

In our developed approach, Baum-Welch algorithm is used to train HMMs that best describes the normal and anomalous behaviors of the system. Baum-Welch algorithm starts with an initial estimate, then uses forward and backward algorithm as cost effective estimation of likelihood and model converges to the nearest local maximum of the likelihood function. HMM learning algorithm is illustrated in Fig. 1.



Training data
Observation sequence
Data pre-processing
$\lambda = \{\pi, A, B\}$
Forward- Backward algorithm
Likelihood
Model converged?
No
Re-estimations of parameters
Yes
Trained HMMs $\lambda = \{\pi, A, B\}$

Fig.1. Flow chart of the HMM training algorithm

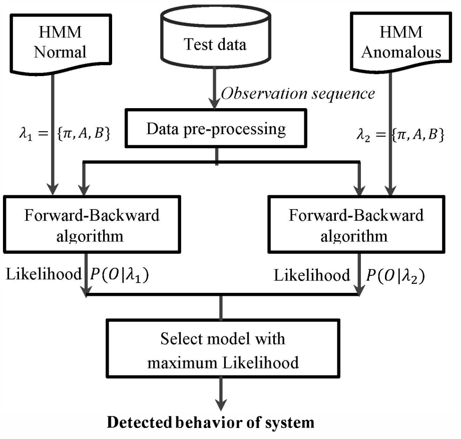
As illustrated in Fig.1, the training iterations including likelihood estimation and re-estimations of the model parameters should be repeated several times until model converges to optimal parameters.

Initialization is important issue for building HMMs for any classification problem. Different starting points lead a different optimum solution [11]. The likelihood P(O|λ), the probability of the observation sequence O given the model λ, used for training and detection phases, since detection phases based on Maximum Likelihood estimation (MLE). Eq. (4) was used for likelihood estimation.

C. Detection phase

In the detection phase, for test data we used same discretization scheme which used for discretizing training dataset. In the detection phase, the likelihoods for normal and anomalous HMMs are computed by the forward and backward algorithm with optimal parameters.

The model with the highest likelihood is considered to be the model that represents the type of behavior of the unknown observation sequence. The procedure for detecting the type of behavior of an unknown observation sequence is depicted in Fig. 2.



HMM Normal
Test data
HMM Anomalous
Observation sequence
$\lambda_1 = \{\pi, A, B\}$
Data pre-processing
$\lambda_2 = \{\pi, A, B\}$
Forward-Backward algorithm
Forward-Backward algorithm
Likelihood $P(O|\lambda_1)$
Likelihood $P(O|\lambda_2)$
Select model with maximum Likelihood
Detected behavior of system

Fig.2. Flow chart of anomaly detection using HMM

As illustrated in Fig.2, the detection result is based on maximum likelihood estimation.

IV.APPLICATION OF THE ANOMALY DETECTION APPROACH

An application of the developed Bayesian HMM-based anomaly detection approach was demonstrated using synthetic data. Two behaviors, normal, representing normal behavior, and anomalous, representing abnormal behavior, were monitored by our approach. Two HMMs representing the normal and anomalous states needed to be trained. 8 observation sequences labeled by normal and anomalous based on system-specific knowledge are used for training. The data pre-processing technique described in Section 3 was used to generate the training observation sequence. Equal-width partitioning was used to discretize data during data pre-processing phase.

TABLE1

DETECTION RESULT USING DEVELOPED HMM BASED ANOMALY DETECTION APPROACH

#	Test data	Real behavior	Detected behavior	Decision	Accuracy

1	Test data1	Normal	Normal	True	85.7%
2	Test data2	Anomalous	Anomalous	True	
3	Test data3	Normal	Normal	True	
4	Test data4	Normal	Normal	True	
5	Test data5	Anomalous	Anomalous	True	
6	Test data6	Normal	Normal	True	
7	Test data7	Anomalous	Normal	False	

While training the model, special attention has to be given to the initialization of A, B and π parameters and model selection issues [5]. In our developed approach, initial values of A, B and π are taken to be uniformly distributed. Both the normal and anomalous models need to have the same number of labels, M, in the observation sequence. The range of each observation sequence was divided into M intervals (partitions) of equal size. The number of labels considered in the observation sequence is 5 based on the complexity and size of the available data.

Choosing the number of hidden states is an important problem for HMMs. Number of states within the HMM affects generalization of the model. Too many hidden states can lead to over fitted model and poor performance [8]. The number of hidden states can be chosen based on the knowledge of the observation sequences, i.e., the different types of observation sequences resulting from different combinations of available symbols/labels [12]. In this paper, we chose the number of hidden states K for the HMMs to be 9 based on the different observation sequences available.

In the detection phase, 7 test sequences were used for anomaly detection using the normal and anomalous models built in the training phase, as shown in Table 1. The model with highest likelihood is considered to be the model which represents the behavior of the test sequence. The detection accuracy of the HMMs can be measured from Table 1. As shown in Table 1, there is 85.7 % accuracy using the developed anomaly detection approach. The efficiency of the approach can be improved by taking into account the optimal number of hidden state, type of HMM, and proper selection data pre-processing technique.

Further research is needed to improve the detection accuracy of the developed HMM by using high quality data from field. The design of the developed algorithm allows for extending current anomaly detection approach as approach of diagnostics by providing multiple labels in the observation sequence based on system-specific knowledge.

V.CONCLUSIONS

In this paper, we present a data-driven approach for anomaly detection using discrete Hidden Markov Model. The detection result showed that the developed approach can help detect anomalous behaviors in system with 85.7% accuracy. The developed anomaly detection approach has two distinct phases: training and detection. In the training phase, models are formulated that best describes the normal and anomalous behaviors in the system. In the detection phase, the likelihoods for normal and anomalous models are computed by the forward and backward algorithm using the optimal parameters of the trained models. The model with the highest likelihood is considered to be the model that represents the type of behavior of the unknown observation sequence. The learning step is typically done using standard Baum-Welch algorithm, while the likelihood estimation is done using the forward and backward algorithm. Discrete HMM allows implementation of online detection since it does not need to invoke any assumption about distribution of data. But discrete HMM can lead to a loss of information during data discretization, which can lead to false decisions regarding the normal and anomalous behavior of the system.

Although the developed approach proved to be successful for anomaly detection of univariate systems, it needs to be implemented using real world data and extended to multivariate systems.

REFERENCES

[1] V. Chandola, A.Banerjee, and V.Kumar, "Anomaly Detection: A Survey," *Journal of ACM Computing Surveys,* vol.41. no.3, article no.15, 2009.

[2] V.Chandola, A.Banerjee, and V.Kumar, "Anomaly Detection for Discrete Sequences: A Survey," *IEEE Transactions on Knowledge and Data Engineering*, vol.24. no.5, 2012.

[3] X.Zang, P.Fan, and Z.Zhu,"A new anomaly detection method based on hierarchical HMM," *Proceeding of the Fourth International Conference on Parallel and Distributed Computing, Applications and Technologies,* pp.249-252, Chengdu, Sichuan, PR China, 2003.

[4] A.N. Srivastava, "Discovering System Health Anomalies Using Data Mining Techniques," *Proceedings of the Joint Army Navy NASA Air Force Conference on Propulsion*,Charleston SC,2005.

[5] S.S.Joshi andV.V.Phoha, "Investigating Hidden Markov Models Capabilities in Anomaly Detection," *Proceeding of 43rd ACM Southeeast Conference,* Kennesaw, GA, March 18-20, 2005.

[6] Y.Qiao, X.W.Xin, Y.Bin, and , and S.Ge,"Anomaly intrusion detection method based on hmm," *Electronics letters,* vol.38, no. 13, pp.663-664,2002.

[7] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257-286, 1989.

[8] P. Baruah, and R.B Chinnam, "HMMs for diagnostics and prognostics in machining processes," *International Journal of Production research*, vol. 43. , no. 6, pp. 1275-1293, 15 March 2005.

[9] S.B. Kotsiantis and P.E. Pintelas, "Data preprocessing for Supervised Learning," *International Journal of Computer Science*, vol. 1, no. 2, pp. 111-117, 2006.

[10] H. Liu and H. Motoda, Instance selection and Construction for Data mining, New York; London: Springer, 2011.

[11] F. Camci, "Process monitoring, diagnostics and prognostics using support vector machine and hidden Markov models," PhD Thesis, Wayne State University, 2005.

[12] M.J.Beal,"Variational algorithms for approximate Bayesian inference," PhD Thesis, Gatsby Computational Neuroscience Unitm University College London,2003.