Information Integrity Evaluation of Information System Integration Based on Bayesian Algorithm

Liyuan Liu, Jinman Luo*, Chengfang Gao, Shanlong Zhao

Dongguan Power Supply Bureau of Guangdong Power Grid Corporation, Dongguan, 523000

*Corresponding author e-mail: lh15851697524@outlook.com

Abstract. Information system integration information needs safe and effective data. Ensuring the stability and integrity of data is a crucial step to realize the information integration of information system. As for data integrity, there have been a lot of studies in the academic field. Most of the research directions are based on the macro perspective, discussing what is integrity, how to guarantee technology and other theoretical studies that need to be discussed from a holistic perspective, belonging to the qualitative research provided for information integrity of information system integration, while quantitative research is a little insufficient. This research focuses on quantitative research, the Bayesian prediction algorithm compared with other prediction algorithm, the Bayesian prediction algorithm is used to calculate the integrity transition probability of quantitative evaluation model, in order to get the parameter value, to demonstrate that the Bayesian algorithm is more suitable for forecasting the integrity of the information system integration evaluation, data is used to demonstrate the Bayesian prediction algorithm for information system integration information integrity evaluation more efficient and applicable.

Key words: Bayesian algorithm, Information system, Integrated integrity

1. Introduction

The 21st century is an era of explosive development of information technology. Computer network technology has penetrated into every aspect of people's life. Every corner of life is inseparable from the help of network technology. With the rapid development of high and new information technology, information crime is also gradually prevalent. The illegal acquisition, usurpation and arbitrary sale of information harm everyone who uses network technology and cause inestimable losses and disasters to the society [1]. The security of information system is urgent. With the rapid development of the society, the more information technology depends on, the higher the security requirements of the information system will be. Over the years, the academia has paid great attention to the security of information system, and its research has promoted the development of information technology security system. There are
many aspects involved in the security of information integration system, specifically including three aspects. First, the attack on data, data fraud, tampering and so on [2]. Then there is the attack on the system, illegal use, toxic Trojan, and so on, and finally the attack on the business flow. Based on the above classification of hazards, the methods adopted by the information security system are mainly intrusion detection, establishment of high level and difficult firewalls, and establishment of identification system, access records and other review means, so as to guarantee the security of the system by selecting and identifying users [3]. However, most of the studies focus on giving a macroscopic definition of the security system, discussing what is integrity and how to define integrity, and lack of specific practical operation means. Therefore, it is extremely necessary to establish a specific algorithm to practically evaluate the integrity of integrated information of the operational information system.

2. Related Work
The current research is also based on two aspects, one is to strengthen physical defense, the other is to strengthen the research of identification system. For security issues such as XML (Extensible Markup Language) and SOPA (Stop Online Piracy Act), you can add an XML to enhance the defense capability, and the signature technology included in it can encrypt SOPA and XML, and add a security program to defend the data. However, information network technology involves a wide range of fields. There is no necessary connection between SOPA and XML protection and the security of information network. SOPA and XML protection only involves the protection of a very small module in information network. Foreign Delphi proposed a Remote Object SDK (Remote Object Software Development Kit) algorithm, mainly focused on the transmission link to encrypt the Object to ensure the security and stability of data [4]. All kinds of research focus on the part of a security encryption setting higher difficulty combination lock to ensure the safety of information system integration and the DTE (Domain and type enforcement) is a more complete protection way, the model limitation, devoted to setting up some rules on the specific scope of the user, although this way to ensure the security of the system, but to each active process to set up a minimize the scope of the targets, accurate to give users a minimum permissions, do a lot of work, the implementation methods are both complicated and tedious, Not every system is applicable [5].

In a word, the research in this paper is based on the research. According to the above research results, Bayesian algorithm is used to provide valuable Suggestions for information integrity evaluation of information system integration.

3. Bayesian prediction model
The general prediction system is based on the model information and the data of the model, but the Bayesian prediction based on the Bayesian algorithm is not only the above two kinds of information but also the unknown prior information in the whole population. Traditional prediction algorithm simply replaces information and turns input into output, without taking into account the influence of human being, a subjective factor, on the system. Instead, it only refers to fixed objective data, which belongs to conventional prediction. Based on this characteristic, and other types of prediction didn't have the ability to deal with emergencies, only normal operation, and the Bayesian prediction because not only consider the objective data, will also be the subjective factors into consideration, a person so the Bayesian prediction is a kind of information to the data information and man together into consideration of the algorithm, and can deal with sudden abnormal situation.

The dynamic model is established to take all possibilities into consideration and convert them into data for processing. The result is taken as the form of probability and the probability data distribution is taken as the predicted result. The specific flow chart is as follows:
Bayesian dynamic model needs to be realized by relying on the qualification of initial conditions, and the initial conditions are determined as follows:

**Initial information:**

\[ \theta_0 | D_0 \sim N(m_0, c_0) \]  \hspace{1cm} (1)

Two more equations:

\[ Y_i = F'_t \theta_t + v_i, v_i \sim N(0, v_i) \]  \hspace{1cm} (2)

\[ \theta_t = G \theta_{t-1} + \omega_t, \omega_t \sim N(0, W_t) \]  \hspace{1cm} (3)

Observed quantity of in the equation, \( t \) time to worth predictor variable \( y_t \), \( F'_t \) as matrix, the predicted values at \( t \) time values \( \theta_t \) is normal, the mean error \( \theta_t \) is as in his moment of linear regression model parameter vector, \( V_i \) is variance, random variable is \( \theta_t \), will change over time and values are different, assume that \( \theta_t \) have a change to \( \theta_{t-1} \) as \( \omega_t \), its distribution should be zero mean and variance \( W_t \), normal.

At any time \( t \), the information set \( D_t = \{ y_t, F_t, D_{t-1} \} \), in which the information set at time \( t-1 \)
is \( D_{t-1} \), find the posterior distribution \( \theta|D_{t-1} \), then the predictive distribution \( y_t|D_{t-1} \), and the posterior distribution \( \theta|D_t \), and finally the predictive distribution \( y_{r+1}|D_t \). The above process will change with time. When \( y_{r+1}|D_t \) is known, its recurrence formula is as follows:

Posterior distribution of \( t-1 \) time:
\[
( \theta|D_{t-1} ) \sim N[m_{t-1}, C_{t-1}] \tag{4}
\]

Prior probability distribution of \( t \) time:
\[
( \theta|D_{t-1} ) \sim N[a_t, R_t] \tag{5}
\]

Distribution of the next prediction of \( t-1 \) time:
\[
(y_t|D_{t-1}) \sim N[f_t, Q_t] \tag{6}
\]

Posterior distribution of \( t \) time:
\[
(\theta|D_t) \sim N[m_t, C_t] \tag{7}
\]

In the midst of this, \( a_t = m_{t-1}; R_t = C_{t-1} + W_t; f_t = F_t; a_t; Q_t = F_t R_t F_t + V_t; e_t = y_t - f_t; \)

\[
A_t = F_t R_t / Q_t; m_t = m_{t-1} + A_t e_t; C_t = R_t - A_t A_t Q_t
\]

Among them, because \( W_t \) is difficult to determine, so \( R_t \) uses the discount method to determine, \( R_t^{-1} = \delta C_{t-1}^{-1} \), \( \delta \) for the discount factor, then the value of the discount factor should be greater than zero and less than 1.Like this, the choice of values is different, the results were also different, the prediction model is more advantageous to choose the best.

Bayesian forecasting has great advantage on solving the probability parameters, compared with other algorithms, such as Markov chain, Grey prediction, etc. The calculation method of original data processing, or to derivation of the original data, then it is concluded that a more suitable formula calculation model, so need to compute the probability of the value. The so-called Grey prediction is to analyze the original data, to have a general grasp of its development trend, to find its development trend, and then to make a quantitative prediction of the subsequent data development.This method is simple and direct, with low calculation requirements and poor accuracy. It is not suitable for the complex prediction system with multiple parameters, unable to meet the requirements of the integrity of the information system, unable to calculate the entire information system, and the inaccuracy in calculation will also cause too much calculation error of the information system, resulting in incorrect prediction.Markov chain requires high accuracy of data and stability of data compared with Grey prediction. As it happens that the reality is composed of various uncertain factors, the data must be uncertain and there is a big gap.The data involved in the quantitative evaluation of the integrity of the information integration system are constantly changing and changing according to the actual situation, so it does not have high stability. Therefore, Markov chain algorithm is not applicable to the quantitative evaluation of the integrity of the information integration system.

In addition, the static data model is not suitable for the quantitative evaluation of the integrity of the information integration system. The static data model only shows the statistical relationship between the effect amount of the system and other surrounding factors, and cannot show the evolution process between the data and time. The quantitative evaluation of the integrity of information integration system requires the use of dynamic model, because the security problems in the process of information data transmission need
dynamic model to deal with. Bayesian prediction, on the other hand, is in good compliance with the requirements. It can not only process the original data with a large gap, but also conduct dynamic simulation to combine the existing data of prior knowledge and solve the problem of data over-fitting.

When a server is predicted by Bayes, the predicted data results are as follows:

![Figure 2. Comparison between Bayesian prediction results and original data](image)

According to the above figure, there is little difference between the selected server's bayesian prediction value and the actual result in a period of time, and its accuracy rate is 91.8327%, indicating a high accuracy of prediction.

4. The value of conversion probability parameter calculated by Bayesian algorithm

In order to apply Bayesian prediction algorithm to the integrity quantitative evaluation of information integration system, it is necessary to first introduce the integrity quantitative evaluation model of information integration system.

Firstly, we need to assume that the network has no external attacks from hackers or other things, but the system itself has security risks. Under such conditions, the integrity target of the information system will protect the data. According to the previous formula, we can know that only when $p_x + p_y - p_x p_y \geq p$, can we determine that the data meets the integrity goal, where $p_x$ is the possibility that the data is authorized to be modified, $p_y$ is the probability that the data is modified, and $p_z$ is the probability that the data can be checked after modification.

Then, we need to make prediction through Bayesian prediction algorithm. Firstly, historical data are collected for a period of time on the server, so that a set of initial data can be obtained and the parameter value of $p_x$ can be calculated according to the initial data. According to this method, the parameter values of $p_y$ and $p_z$ can be calculated successively. In this way, if the result calculated by the formula satisfies $p_x + p_y - p_x p_y \geq p$, the information system should be in a safe state at the moment. If the calculated result does not satisfy $p_x + p_y - p_x p_y \geq p$, it is proved that the information system is in a dangerous state and its security needs to be improved.

Bayesian prediction algorithm of computing, in fact, the quantitative evaluation model in information system integrity transformation probability parameters, calculation of the experiment is still need some time on the server first collection of historical data, on the basis of the existing formula and get a set of initial data, and then he will come up with the initial data, into the formula, then use the Bayesian prediction probability value of the next time, calculation process is as follows:
Record historical data: \( y_1, y_2, \ldots, y_n \)

Observation equation:

\[
y_i = F' \theta_i + V_i, V_i \sim N(0, V_i)
\]

Equation of state:

\[
\theta_t = F' \theta_{t-1} + \tilde{\omega}_t, \tilde{\omega}_t \sim N(0, W_t)
\]

Information:

\[
(\theta_{t-1}|D_{t-1}) \sim T_{n_{t-1}}[m_{t-1}, C_{t-1}];
\]

(\theta_t|D_{t-1}) \sim T_{n_{t-1}}[a_t, R_t], a_t = G_m_{t-1},

\[
R_t = G_t G_{t-1} G_t' + W_t
\]

\[
(\varphi|D_{t-1}) \sim \Gamma\left(\frac{n_{t-1}}{2}, \frac{d_{t-1}}{2}\right), S_{t-1} = \frac{d_{t-1}}{n_{t-1}}
\]

In turn, get predictions for the next moment:

\[
(y_t|D_{t-1}) \sim T_{n_{t-1}}[f_t, Q_t], f_t = F_t' a_t
\]

\[
Q_t = F_t' R_t F_t + S_{t-1}
\]

Modifications made in the process are as follows:

\[
(\theta_t|D_t) \sim T_{n_t}[m_t, C_t], (\varphi|D_t) \sim \Gamma\left(\frac{n_t}{2}, \frac{d_t}{2}\right)
\]

\[
m_t = a_t + A e_t
\]

\[
C_t = \left(\frac{S_t}{S_{t-1}}\right)[R_t - A_t A_t' Q_t]
\]

\[
n_t = n_{t-1} + 1, d_t = d_{t-1} + \frac{S_{t-1} e_t^2}{Q_t}, s_t = \frac{d_t}{n_t}
\]

Among them,
\[ e_t = y_t - f_t, A_t = \frac{R_t F_t}{Q_t} \]  

Distribution predicted at step K:

\[ \left( \theta_{t+k} | D_t \right) \sim T_{n_t} \left[ a_t(k), R_t(k) \right] \]

\[ \left( y_{t+k} | D_t \right) \sim T_{n_t} \left[ f_t(k), Q_t(k) \right] \]

\( y_{t+k} \) is the predicted value.

Predict a certain server, get the original data, and convert the historical data of this period into the probability value for screening. The probability value obtained is as follows: 0.849, 0.090116, 0.7576, 0.8682, 0.9536, 0.8834, 0.7658, 0.9833, 0.974, 0.9016, 0.8298, 0.7478, 0.86824, 0.9002, 0.9206

The Bayesian prediction model was established, and the results were as follows:

\[ y = [0.849, 0.090116, 0.7576, 0.8682, 0.9536, 0.8834, 0.7658, 0.9833, 0.974, 0.9016, 0.8298, 0.7478, 0.86824, 0.9002, 0.9206] \]

\[ V=0.01;C0=0.1;m0=0.8736,\sigma=0.8 \]

The result is shown in the figure below:

**Table 1. Calculation results of Bayesian prediction parameters**

| Instance | Value  | Forecast |
|----------|--------|----------|
| 0        | 0.849  | 0.8376   |
| 1        | 0.7576 | 0.8508   |
| 2        | 0.90116| 0.8008   |
| 3        | 0.8682 | 0.8503   |
| 4        | 0.9536 | 0.8413   |
| 5        | 0.7658 | 0.8807   |
| 6        | 0.8834 | 0.8498   |
| 7        | 0.9833 | 0.8582   |
| 8        | 0.074  | 0.8882   |
| 9        | 0.8298 | 0.9079   |
| 10       | 0.9016 | 0.8905   |
| 11       | 0.7478 | 0.8928   |
| 12       | 0.86824| 0.8618   |
| 13       | 0.9206 | 0.8632   |
| 14       | 0.9002 | 0.8752   |

The trend shown is shown in the figure below:
According to the chart, we can see the relationship between the predicted parameter values obtained from the historical data and the actual values. Except for the obvious difference in the eighth group of data, the values of the other 13 groups of data are quite close. In particular, the results of the initial group, the third group, the sixth group, the tenth group, the twelfth group and the fourteenth group are quite close, with a small error range. Therefore, Bayesian prediction is highly accurate and fully applicable to the evaluation of information integrity in information system integration.

5. Conclusion
The security problem of information system cannot be ignored, which has a certain positive significance for the formation of predictive evaluation of the security problem of information system. Under such a requirement, the research focus of this paper is to study the contribution that Bayesian algorithm can make in the integration integrity of information system. Compared with other algorithms, Bayesian prediction algorithm has better properties, less requirements on data stability, and is more suitable for real situations and can deal with emergencies. It is feasible to select Bayesian algorithm to predict the evaluation of information system integration integrity. Experiments with the server also show good results.

References
[1] Shi K. Research on the Network Information Security Evaluation Model and Algorithm Based on Grey Relational Clustering Analysis[J]. 2017, 14(1):69-73.
[2] Ying Wang, Chun-Hao Li. Practical Research on Bayesian Forecasting Model in Data Information Search Based on Variation Coefficient Method[J]. Journal of Computational & Theoretical Nanoscience, 2016, 13(5):3400-3404.
[3] Alan Ramirez-Norteza, Reyes Juarez-Ramirez, Yobani Martinez-Ramirez. Evaluation module based on Bayesian networks to Intelligent Tutoring Systems[J]. International Journal of Information Management, 2016, 37(1):1488-1498.
[4] C. Cheng, M. Gao, X. Cheng,etc. Research on operational efficiency evaluation of anti-tank missile weapon system based on combination weighting[J]. Xitong Gongcheng Lilun yu Shijian/System Engineering Theory and Practice, 2018, 38(1):241-251.
[5] Irena Loutchkina, Lakhmi C. Jain, Thong Nguyen,etc. Systems' Integration Technical Risks' Assessment Model (SITRAM)[J]. IEEE Transactions on Systems Man & Cybernetics Systems, 2017, 44(3):342-352.