Intelligent Analysis of Database Users Based on A Dynamic Model

Jieqing Ai, Jianyong Wang, Shouming Chen, Hao Guan [a]*, Chengdong Liang, Liang Chen

Information center of Guangdong

Power Grid Corporation, Guangzhou, China

[a] wudiguanhao001@126.com

*corresponding author

Keywords: Database Security; Behaviour Analysis; Anomaly Detection; Hidden Markov Model; User Profile;

Abstract: Database security issues play an important role in modern information systems. This paper aims to deal with one of the most difficult issues for database security. We propose a novel intelligent method for database user behaviour analysis. Specially, we use hidden Markov model (HMM) to model the user profile. Different from most of previous works which only focus on query syntax for anomaly analysis, our approach explores more temporal relationships between operations and can take advantage of big training data. Experimental result has shown that our model is quite effective and efficient.

1. Introduction

At present, more and more enterprises have built up information systems for handling business, such as billing, customer relationship management, business data analysis and electronic business. Despite the diversity of applications, most of the systems are based on database management. Therefore, it is necessary to analyze the database user behaviours and detect abnormal operations on the database. Anomaly detection for database refers to the process that a person or program monitors each event of the database and alert when anomalies have been detected [1,2]. After that, some measures should be taken to prevent further damages to the entire information system and minimize its damages.

This paper focuses on the intellegnent analysis of database user behaviour for abnomaly detection. All the operations on the database are based on SQL query statements. While most of previous works only focus on query syntax to construct user behaviour model, we pay more attention to temporal relationship between operations and takes into account the amount of some sensitive information a query result contains. For its popularity and effectiveness in the machine learning community, we use Hidden Markov Model (HMM) [3,4,5] which has shown good performance in sequence learning such as natural language processing (NLP) to model database user behavior. The data for model training and testing comes from the log files of a database system which belongs to a telecommunication service provider.

The trained model is then used to classify unknown users. When the classification score is lower than a threshold, an anomaly is detected. The main contributions of this paper are as follows. First, we propose a new database user hehaviour analysis and anomaly detection system. Our method takes full advantage of training data and can be independent of the database schema. Second, we take experiments on a real world information system of a telecommunication service provider to verify the effectiveness of the proposed mothod.

2. Related Work

User behaviour analysis and anomaly detection is an important topic in multiple areas. Specifically, anomaly detection refers to the process of finding the patterns in data that are not in accordance with expected behaviors [6]. Roughly, the common anomaly detection techniques can be categorized into nearest neighbor based method [7,8,9], clustering based method [10,11], statistical based methods [12,13], information theory based methods [14] and spectral based methods [15]. The authors in [16] propose to adopt the user profile to detect misuse of database. In [17], some machine learning methods have been introduced into anomaly detection. Despite the progress, most of the methods are based on query sytaxrelay or static features without full use of temporal information.

3. Proposed Method

3.1 System Overview

The overall flow chart of the proposed method is illustrated in Figure 1. The whole system can be divided into two stage, namely, the training stage and the detection stage.

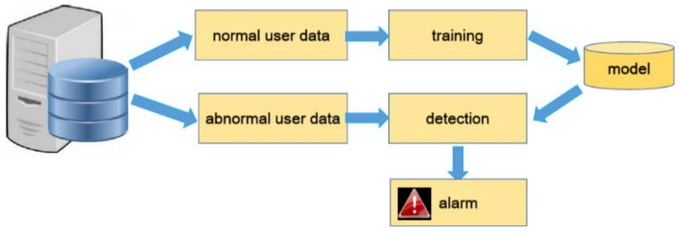

Fig.1 Main Flow Chart of the Proposed Anomalies Detection Method.

3.2 Basic Model

Different from traditional syntax-based methods, our primary object is to protect database by using a dynamic model which can exploit more temporal information in operations. We adopt Hidden Markov model (HMM) [3] as the basic model of our system. In practice, we take a temporal sequence of observable query statements as the behaviour model of a specific database user.

Generally, HMM is a probabilistic model that combines Markov chains and general stochastic processes using model parameters. It can be used to describe the statistical properties of stochastic processes. The transfer of the Markov chain is used to describe the state of the general stochastic process and the relationship between the state and the observation sequence. The state transition process of the HMM is not observable, but the model can be inferred from a series of observations.

An HMM is generally represented as an ordered tuple as:

$$\lambda = (A, B, \pi) \qquad (1)$$

Besides the three items, there are another two important parameters:

(1) Total number of states of the model. Use S to denote the set of model states, S = s_1, s_2; …, s_N. Use qt ∊ S to denote the model statement at time t.

(2) Total number of observations of the model. Use V to denote the set of model states, V = v_1, v_2, …, v_N.

The parameter A is the state transition probability matrix of state, A = $a_{i,j}$, where:

$$a_{i,j} = P(q_{t+1} = s_j \mid q_t = s_i), \qquad 1 \le i,j \le N \qquad (2)$$

The parameter B represents the probability distribution of the observation symbol and B = $b_i(j)$, where $b_i(j)$ represents the probability of the observed value in state i with its definition as:

$$b_i(j) = P(o_t = v_j \mid q_t = s_i), \qquad 1 \le i \le N, 1 \le j \le M \qquad (3)$$

The initial state probability distribution is:

$$\pi = \pi_1 = p(q_1 = s_i), 1 \le i \le N \qquad (4)$$

Let O be an observation sequence, then the sequence can be denoted as O = $O_1O_2 \ldots O_T$, where O_t represents the observation at time t.

3.3 The Proposed Analysis Model

The total number of states N and observations M are determined by the types of events in the log files, such as asking for a connection, logging on to the server, using the select query, etc. The observation sequence of the model is a series of actions in chronological order, and the maximum length of each observation sequence is T. The state of an event is unobservable behind the observation sequence, namely, the hidden variables.

In this paper, the Forward-Backward algorithm [3,4,5] is used to calculate the state transition probability matrix A, the probability distribution of the observation B as well as the probability of observing the sequence P(O|λ). The initial probability of each state is assumed to be uniformly distributed. The whole running flow of the proposed model can be illustrated by Fig.2.

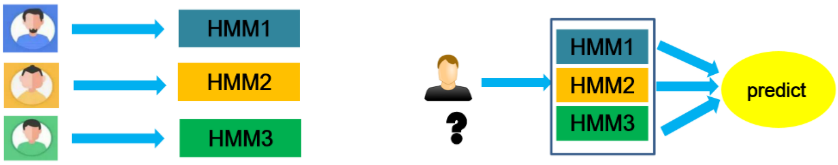

Fig.2 The Proposed Model for User Classification

As shown in Figure 2, we pre-define the number of roles in the system and we build a HMM for each role. The value of P(O|λ) is used as the classification score for each testing sequence. A sequence of an unknown user with the highest matching score is categorized into the corresponding role. If the matching score is lower than a threshold, the operation is viewed as an anomaly.

4 Experiment Results

4.1 Experiment Dataset

To evaluate the proposed method for user classification, we use a real world dataset from a telecommunication provider. All the data is from the operation logs of all the database users in one week.

The dataset is divided into two parts, one part is used to train the model, and the other is used for testing. The ground-truth labels for all the data are offered by the telecommunication provider. We mainly use the classification precision as the testing metric for our experiments.

4.2 Experiment Results with Comparisons

The experimental results are shown in table 1. From the results it can be observed that the proposed method is accuracy for user classification.

Table.1

	dataset	type	number	accuracy
1	A	training	300	96
2	B	testing	500	84

We also compare the proposed method with traditional sytax-based method and the comparison result is shown in Figure 3. Through the comparison, it can be observed that our method outperforms the traditional one. It can be explained by that the dynamic model which explores temporal relatonships can take more advantages of training data and be more robust and precious than traditional ones.

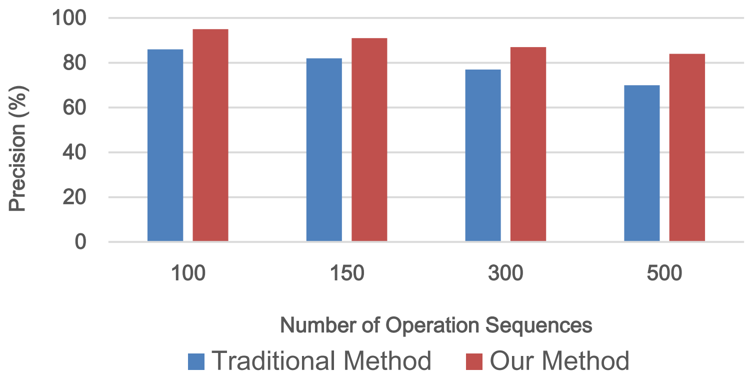

Fig.3 Comparision of The Proposed Method and Traditional Method

4.3 Insight Investigation of HMM States

We also test the accuracy of different number of states. The result is shown in Figure 4. We can see that with more number of states, the accuracy can be higher. Too many states, however, may bring down the accuracy which can be explained by the limitation of capacity of our training model.

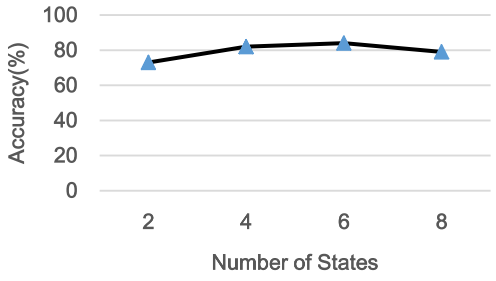

Fig.4 Accuracy on different number of states

5. Conclusion

In this paper, we focus on the problem of database user classification and analysis, which is import to protect information systems. We introduce the Hidden Markov Model to model user behaviour and classify different roles. Experiment results on a real world database of a telecommunication provider show that the proposed method can accurately classify different users according to their behaviour patterns.

Reference

[1].V. Chandola, A. Banerjee, and V. Kumar., “Anomaly detection: A survey,” ACM Computing Surveys, vol. 41, no. 3, 2009.

[2].M. Gupta, J. Gao, C. C. Aggarwal, and J. Han, “Outlier detection for temporal data: A survey,” IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 9, pp. 2250–2267, 2014.

[3].L. Rabiner and B. Juang, “An introduction to hidden markov models,” IEEE ASSP Magazine, vol. 3, no. 1, pp. 4–16, 1986.

[4].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.

[5].L. E. Baum, T. Petrie, G. Soules, and N. Weiss, “A maximization technique occuring in the statistical analysis of probabilistic functions of markov chains,” The annals of mathematical statistics, vol. 41, no. 1, pp. 164–171, 1970.

[6].M. Agyemang, K. Barker, and R. Alhajj., “A comprehensive survey of numeric and symbolic outlier mining techniques,” Intelligent Data Analysis, vol. 10, no. 6, pp. 521–538, 2006.

[7].S. Ramaswamy, R. Rastogi, and K. Shim, “Efficient algorithms for mining outliers from large data sets,” vol. 29, no. 2, pp. 427–438, 2000.

[8].M. Wu and C. Jermaine, “Outlier detection by sampling with accuracy guarantees,” in Proc. of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, 2006.

[9].P. Sun, S. Chawla, and B. Arunasalam, “Mining for outliers in sequential databases,” in Proc. of the 2006 SIAM International Conference on Data Mining, 2006.

[10]. R. Smith, A. Bivens, M. Embrechts, C. Palagiri, and B. Szymanski, “Clustering approaches for anomaly based intrusion detection,” Proceedings of intelligent engineering systems through artificial neural networks, pp. 579–584,2002.

[11]. Z. He, X. Xu, and S. Deng, “Discovering cluster-based local outliers,” Pattern Recognition Letters, vol. 24, no. 9, pp. 1641–1650, 2003.

[12]. E. Eskin, “Anomaly detection over noisy data using learned probability distributions,” in Proc. of the International Conference on Machine Learning, 2000.

[13]. K. Kadota, D. Tominaga, Y. Akiyama, and K. Takahashi, “Detecting outlying samples in microarray data: A critical assessment of the effect of outliers on sample classification,” Chem-Bio Informatics Journal, vol. 3, no. 1, pp. 30–45, 2003.

[14]. E. Keogh, S. Lonardi, and C. A. Ratanamahatana, “Towards parameter-free data mining,” in Proc. of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining, 2004.

[15]. A. Agovic, A. Banerjee, A. R. Ganguly, and V. Protopopescu, “Anomaly detection in transportation corridors using manifold embedding,” Knowledge Discovery from Sensor Data, pp. 81–105, 2008.

[16]. C. Y. Chung, M. Gertz, and K. Levitt, “Demids: A misuse detection system for database systems,” Integrity and Internal Control in Information Systems, pp. 159–178, 2000.

[17]. B. J. , V. M. Santos, R. J., “Approaches and challenges in database intrusion detection,” ACM SIGMOD Record, vol. 43, no. 3, pp. 36–47, 2014.