User abnormal behaviour sequence detection method based on Markov chain and SVDD

User abnormal behaviour sequence detection method based on Markov chain and SVDD

Shengyuan Zou*, Chaowen Chang and Peisheng Han

Information Engineering University, Zhengzhou, Henan, 450001, China

*Corresponding author's e-mail: zou_sheng_yuan@126.com

Abstract. In order to solve the problem of insufficient use of sequence information and low detection efficiency of traditional anomaly detection methods, this paper introduces Markov chain into user behaviour sequence detection, and proposes a description based on Markov chain and support vector data field (SVDD) User Behaviour Sequence Detection Method (ASDMS), which first uses the Markov chain to accurately quantify the user behaviour sequence, then constructs the user's normal behaviour sequence model based on the support vector data field description model, and identifies the user anomaly behaviour. The experimental results show that the ASDMS method has better performance and timeliness than the traditional abnormal behaviour detection method.

1. Introduction

In an increasingly complex network security environment, how to deal with diversified and complex attacks and internal threats becomes the top priority of network security[1]. With the continuous development of cyber-attack technology, traditional rule base matching, information entropy, and statistical analysis methods are difficult to deal with complex attack behaviours such as precision-designed APT. User behaviour analysis accurately describes the behaviour of users inside and outside the system. It can be used to discover abnormal behaviours of normal user operations caused by user accounts being cracked, illegally exploited, malicious operations, virus infections, etc. Compared with detection methods based on signature matching, user abnormal behaviour analysis is focused on normal users. Behavioural modelling, which can detect unknown attack types and sudden internal threats that do not conform to the current behaviour patterns, has attracted the attention of more and more enterprises and scholars.

2. Related work

Currently user abnormal behaviour analysis and detection methods use the training data set to construct the user's normal behaviour model to identify the abnormal behaviour of the model. In the existing research results, the mainstream user abnormal behaviour detection methods are divided into sequential pattern mining and massive data classification.

The sequential pattern mining method is a method combining the association rules and the time dimension. It analyses the normal user behaviour sequence, discovers the relationship between the behaviour sequences, and constructs the user behaviour model by combining the user state information. The main methods are: literature Using the data merging method is used to describe the user behaviour, constructing a part of the regularized symbiotic matrix model to determine whether the user behaviour is legal[2], and the stepwise merging method used in [2] is based on the shell frequency statistics

without the relatively scientific user service sequence description; An effective one-off mining algorithm with wildcards is designed to improve the validity and completeness of pattern mining[3]. This method can describe user behaviour more accurately, but the time performance cannot meet the application of large data volume background requirements; The sliding window method is studied[4], combined with the sequential pattern mining method for abnormal behaviour detection, but the modelling process of this method is complex, and the time consuming is poor in the application scenarios of big data; Fully use of the structural information of the entity is fully used in [5] for pattern mining, so that the description of the user's normal behaviour is more readable, but the data dimension rapidly expands with the increase of the entity attribute, which needs to balance the timeliness and precision of the algorithm; The Markov method is used to discretize the original data to form the system state sequence[6], The improved Jensen-Shannon divergence is used to describe the user behaviour sequence trend, and the test data bias is evaluated. The sequence description method has certain reference significance for this paper, but the description process is too complicated; The data is collected through smartphone sensors, an association spanning tree is designed to solve the redundancy problem in the process of generating association rules [7], and a normal user behaviour model is established. However, this method focuses on association analysis and ignores the information contained in the user sequence. Accuracy is low; Combining the graph theory method is combined with the anomaly detection method [8,9], the graph theory concept is used to describe the network assets to find abnormal behaviour, but the graph-based method makes the data scale explode and consumes a lot of computing resources.

Massive data classification method identifies user anomalous behaviour from the perspective of similarity analysis among a large number of users. The main methods are: The hybrid perturbation generation method is used to construct the differential member classifier[10], the collaborative learning of the sample subsets improves the accuracy of the classifier, but it is difficult to find suitable model parameters; The Youden index and information gain are used to comprehensively evaluate the selected data attributes, and classify the data through the improved FCM clustering algorithm to identify the user's abnormal behaviour[11], but its recognition of local dense anomaly clusters is not strong.

The above research has certain guiding significance for the user's abnormal behaviour detection method, but they are not sensitive enough, the use of data sequence relationship is not sufficient. For this reason, this paper introduces the Markov chain steady state process and proposes an anomaly sequence detection method based on the Markov chain and SVDD method(ASDMS) based on the ability of strong description of sequence data and the strong representation ability of Support Vector Domain Description (SVDD) for data clusters. ASDMS improves the accuracy and timeliness of user behaviour detection.

3. User abnormal behaviour sequence detection method based on Markov chain and SVDD

The ASDMS method is mainly divided into two stages of training and detection. In the training stage, the sequence classification coding method based on Markov chain steady-state vector is firstly used to abstract the user behaviour sequence, which can not only accurately describe the relationship between user behaviour sequences, but also have obvious data dimension reduction effect, then use SVDD method to analyse the steady-state vector set of normal user behaviour, depict the boundary model of normal user behaviour; in the detection phase, bring the current user behaviour sequence into the boundary model to identify user behaviour abnormal.

3.1. Markov chain steady state vector

User behaviour is a series of related sequence information. Markov chain is a powerful mathematical tool for describing sequence information, which can accurately describe the transfer characteristics of user behaviour. Firstly behaviour sequence S is divided into k subsequence $A=\{S_1,\cdots,S_n\}(|S_i|\le k)$, the description process of S is transformed into finding the common mathematical features of the

elements of $\mathbf{A}$, then each subsequence S_i constructs Markov chain symbol transition probability matrix $\mathbf{H}_i$ which is regard as the basic mathematical description form of S_i, in order to eliminate the inaccuracy caused by the length of, this article introduces $\mathbf{SV_i}$ (Steady State Vector) for each $\mathbf{H}_i$ as the final mathematical description form of S_i.

For each $\mathbf{H}_i$,if the matrix is traversable, a unique stationary probability distribution characterizing the user behaviour jump, i.e. SV_i, can be obtained. The matrix ergodicity can be considered that the random process can be transformed in the finite step jump in any state. For any other state, the states of the corresponding Markov chain are fully connected, ie all jumps have a non-zero probability.

But for the subsequence S_i, some state or state transitions may not occur in finite-length sequences, the ergodicity of H_i cannot be guaranteed, in order to solve this problem, we transfer $\mathbf{H}_i$ into Google Matrices[12] $\mathbf{G}$ which is a random traversal matrix originally used by Google's PageRank algorithm to process large sparse matrices that represent link jumps between web pages. Google Matrices $\mathbf{G}$ in this article is calculated as follows:

$$\mathbf{G} = d\mathbf{H} + (1-d)\frac{1}{k}\mathbf{e} \tag{1}$$

$\mathbf{H}$ is the transition probability matrix of the subsequence; d is the damping coefficient between 0 and 1 which represents the influence $\mathbf{H}_i$ on $\mathbf{G}$; k represent the length of S_i, the long-length $\mathbf{H}_i$ is more likely to be a fully connected matrix, then there is smaller proportion for the latter term in Equation 1; $\mathbf{e}$ represents a square matrix in which each element is 1. Due to $\mathbf{G}$ is traversed, each $\mathbf{SV}_i$ represents the behaviour pattern of the user within the subsequence.

The choice of d depends on the continuity of the user behaviour. In the user behaviour sequence, in addition to the sequence of behaviours that are mostly executed sequentially within a single businesses, there is also the user behaviour of random jumps between businesses, the more high jumps user behaviour, the lower valuable $\mathbf{H}_i$ is on G_i, if there are more zero or near zero elements in H_i, we choose lower d .Since the state of the user behaviour jump map is small in the context of big data, d_i cannot be specified in each S_i. This paper determines $d = 0.95$ through the statistical analysis and experimental analysis of user behaviour sequences which is one of the parameter bases of Section 4.

In order to eliminate the randomness effect of the user sequence jump represented by the matrix $\mathbf{G}$, we use $\mathbf{G}$ corresponding steady state vector $\mathbf{SV}$ as the final mathematical description of the user subsequence behaviour pattern, the conversion from $\mathbf{G}$ to $\mathbf{SV}$ uses a power level function:

Definition the power level function of $\mathbf{G}$:

$$\mathbf{E}_i = \prod_{1}^{n} \mathbf{G}_i \tag{2}$$

When the power level function is satisfied $\mathbf{E}_i * \mathbf{G}_i = \mathbf{E}_i$,the matrix row elements of $\mathbf{E}_i$ are consistent, so $\mathbf{SV}_i = \mathbf{E}_i * \mathbf{v}$,among them v is a unit vector whose element is 1, the mathematical description $\mathbf{SV}_i$ of S_i is a multidimensional vector that describes the jump characteristics of a sequence of user behaviour. For the convenience of subsequent research, the subsequence steady state vector set is defined as $B = \{\mathbf{SV}_1, \cdots, \mathbf{SV}_n\}$ in the corresponding user behaviour at each stage in $\{S_1, \cdots, S_n\}$.

3.2 Support vector data field description model

Subsequence steady state vector set B is an accurate description of the sequence of user behaviour, the ASDMS method is digitally aggregated the boundary of B as range of the sequence of user behaviour. This paper uses the SVDD single classification model to transform the user behaviour

sequence boundary characterization problem into the problem of recognizing the minimum boundary of B, the goal can be transformed into an optimal evaluation function $f: X \to Y$, for $x_i \in X$ there is a corresponding $y_i \in Y$ where X represents subsequence steady state vector set B, Y represents the minimum boundary of B.Based on the principle of empirical risk minimization, the optimization problem can be described as follows:

$$f^* = \arg\min_f \theta(f) + \frac{\eta}{n}\sum_{i=1}^{n} L(f(x_i)) \tag{3}$$

L, θ, η represent Respectively the loss function, the regularization and reconciliation parameters of f.The vector set B maps to the new feature space and is used to train acts normal behaviour model ς. After training, the high-dimensional minimum hypersphere model of SVDD is obtained where c is regard as a centre of the hypersphere of the radius R covering all normal behavioural data, The distance between the mapping position of the SV vector in the feature space and the origin is used as an indicator to test whether the user behaviour sequence is abnormal which is as shown in equation (4):

$$f(x) = \left\|\phi(x) - c\right\|^2 - R^2 \tag{4}$$

The user behaviour evaluation function is shown in equation (5):

$$h(\mathbf{M}) = \bigcup_{i=1}^{t} f(\mathbf{SV}_i) < 0 \tag{5}$$

M presents the steady-state vectors set corresponding to the sequence of user behaviours to be tested, and if the instances to be tested all fall within the hypersphere, the evaluation function value $f(\mathbf{M})$=*True*, the sequence instance is determined to be a normal user behaviour, and if there is data falling outside the hypersphere, the evaluation function value $f(\mathbf{M}) = False$, the sequence instance is determined to be an abnormal behaviour of the user. At the same time, considering the sample description bias and the abnormal point interference, the slack variable is introduced here $\xi > 0$ requiring the sample to meet the constraints

$$\left\|\phi(x_i) - c\right\|^2 \le R^2 + \xi, \xi \ge 0 \tag{6}$$

At this point, the objective function is transformed into equation (7), and the constraint of equation (6) needs to be satisfied.

$$\min_{R,c,\xi} R^2 + C\sum_{i=1}^{n} \xi_i \tag{7}$$

C is the parameter balancing R、ξ, the parameters C can be seen as a description of the positive correlation of the data set impureness: $0 \le C < 1$.This problem is transformed into a convex optimization problem, which can be solved by the Lagrangian multiplier method. The Lagrangian function of the constrained optimization function is

$$L_p = R^2 - \sum_{i=1}^{n} \beta_i \xi_i + C\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i (R^2 + \xi_i - \left\|\phi(x_i) - c\right\|^2), \alpha_i \ge 0, \beta_i \ge 0 \tag{8}$$

Seek the partial guidance respectively for R、c with ξ_i

$$\sum_{i=1}^{n} \alpha_i = 1 \tag{9}$$

$$c = \sum_{i=1}^{n} \alpha_i \phi(x_i) \tag{10}$$

$$\alpha_i = C - \beta_i \tag{11}$$

Bring them into equation (8)

$$\max L_D = \sum_{i=1}^{n} \alpha_i K(x_i, x_i) - \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j K(x_i, x_j) \quad s.t.\ 0 \le \alpha_i \le C, \sum_{i=1}^{n} \alpha_i = 1 \tag{12}$$

The final problem is transformed into an extremum problem with a parameter function. According to the parameters α_i, the value of the sample can be divided into the three categories: if $\alpha_i = 0$, the corresponding sample is located in the sphere, and the sample point is a normal subsequence; if $0 < \alpha_i < C$, the sample is on the sphere, and the corresponding sample point is a support vector sample; if $\alpha_i = C$, the corresponding sample point falls outside the sphere, and the sample point is an abnormal subsequence. As shown in Equation 5, the distribution characteristics of the sample points corresponding to all subsequence in the test instance can be analysed, and the norm of the current user behaviour sequence can be analysed.

4. Analysis of results

In order to verify the accuracy and timeliness of ASDMS, this paper uses the public experimental data of Purdue University[13] to test the performance and effectiveness of the method and compared it with other related algorithms. The experimental data contains the activity records of eight Unix users within two years. The user's data file filters out the host name, URL, etc., and only retains the name and parameters of the shell command. The commands in the command stream follow the order in the shell session. The order is arranged, and different shell sessions are connected in time series. In the experiment, data of four users (user1, user2, user3, user4) was selected for experimentation, user1 was set as a legal user, and user2, user3, and user4 were set as illegal users. Each user has 15,000 commands, and the first 10000 commands of user1 are used as the training data of the normal user model. The 500 commands of user2, user3, and user4 are respectively dispersed into the normal behaviour sequence of user1 in the session. The 6500 commands is used as test data.

The detection method of this paper is to use the static transit vector of the first-order homogeneous Markov chain to describe the normal behaviour profile of the legitimate user. The state of the Markov chain corresponds to the type of shell command executed by the legitimate user. Shell commands can be divided into five categories: system management, network management, software and tools, file directory management, hardware and kernel. In the experiment, Python scripts are used to complete the mapping between shell commands and categories.

The comparison algorithms include the collaborative learning method (SCL)[9], the pattern mining method (PM)[5], the co-occurrence matrix method (CM)[3], and the graph segmentation method (I-MCL)[10]. The experimental environment is 64-bit Windows10 OS, Intel(R) Core(TM) i5-7300HQ CPU@2.5GHz processor, 12G memory, Python 3.6.7, using lukasruff code exposed on GitHub[14] to implement the SVDD.

4.1. Experimental evaluation index

In the detection of user behaviour anomaly, the data set has unbalanced characteristics, most of which are normal data, but the abnormal data can provide a larger amount of information, and the misclassification of a small amount of abnormal data has a huge misleading negative impact on the analysis system. The experimental method uses three indicators of accuracy, recall and detection accuracy to compare the performance and accuracy of different methods.

Table 1. Evaluation indicators

Actual type	Test results	
	normal	abnormal
normal	TP	FN
abnormal	FP	TN

The accuracy rate represents the proportion of instances that are correctly detected by the model, i.e. $\frac{(TP+TN)}{(TP+FP+FN+TN)}$, reflecting model detection performance; recall rate $\frac{TN}{(FP+TN)}$, reflecting

the false negative rate of the model; the detection accuracy is $\frac{TN}{(FN+TN)}$ reflects the false positive rate of the model.

The following data sets are used to compare the accuracy, recall rate and detection accuracy of five methods for analysing user anomalous behaviour from different angles. The experimental results are shown in Table 2:

Table 2. Confusion matrix indicator test results

Detection method	Test results		
	Accuracy	Recall rate	Detection accuracy
ASDMS	90.52%	87.25%	94.87%
I-MCL	89.26%	79.56%	89.77%
CM	83.54%	85.22%	84.56%
PM	85.22%	81.26%	87.54%
SCL	82.44%	70.24%	80.26%

It is seen from the experimental results that since ASDMS and PM are superior to other algorithms in sequence description methods, the detection accuracy is relatively high. It can be seen that the user behaviour sequence description method is the key factor for analysing user behaviour.

To further illustrate the effectiveness of the ASDMS user anomaly behaviour detection method proposed in this paper, the ROC (Receiver Operating Characteristic Curve) is used to analyse the performance of the algorithms. The ROC curve is a visual representation of the trade-off between the true rate of the classifier and the false positive rate. In the ROC curve, the axis x is a false positive rate (a normal sample classified as an abnormal sample), the axis y is the true rate (the abnormal sample that is correctly classified), and the values of ROC is in $(0,1]$. $(0,0)$ represents a model that each instance is predicted to be normal. $(1,1)$ represents a model that each instance is predicted to be abnormal. A good recognition algorithm will be near the upper left corner of the graph. The ROC curves of various algorithms are shown in Figure 1. It can be seen that the ASDMS algorithm can achieve ideal detection at low error rates.

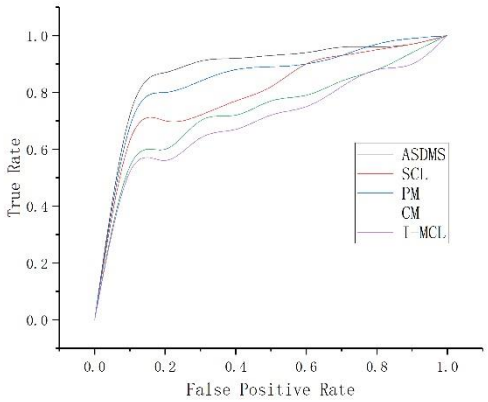

Figure 1. ROC curve of each algorithm

4.2. Algorithm timeliness analysis

In some application scenarios with high real-time requirements, algorithm time-consuming is also an important indicator of performance. The experiment evaluates the timeliness of the algorithm under different data scales. The results are shown in Figure 2:

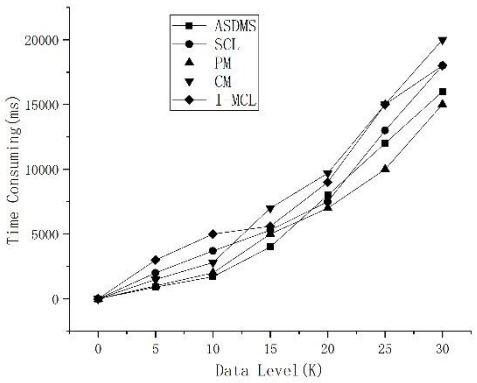

Figure 2. Analysis of timeliness of the algorithm

On the whole, under the 10,000-level data scale, the timeliness of each method is not much different. The ASDMS method for the abnormal behaviour detection method in this paper has good timeliness performance when the data size is small, and with data size increases, the time consumption has dropped significantly.

It can be seen from the above experimental results that the Markov chain coding method for the Shell sequence can significantly reduce the data description dimension, and the support vector data domain description can quickly process low-dimensional data, which has a good time effect, but the increased data scale increases the time consumption of the Markov chain coding process, and the time performance of the method is slightly reduced. The next step is to implement the method and optimize the Markov chain coding process in the distributed system.

Finally, the parameters of the algorithm are discussed. The method is divided into two parts: the encoding process of Markov chain of user behaviour and the description of the support vector data field to identify the abnormal subsequence. During user behaviour coding, subsequence length k too short will cause the behaviour description state space to be too scattered, and it cannot accurately describe the user behaviour, on the other hand, if the subsequence is too long, the anomaly fragment is submerged in the normal behaviour sequence, the sensitivity of the model to the abnormal behaviour is reduced; in the support vector data domain description model, the equilibrium parameter C as a balance R with ξ evaluate the generalization of the model's boundaries to normal user behaviour data, and its settings have an important impact on the accuracy of the model. Therefore k and C as the two important parameters of the ASDMS algorithm, have a perturbation relationship, and some experimental parameters are selected for the experiment. The experimental results are shown in Figure 3:

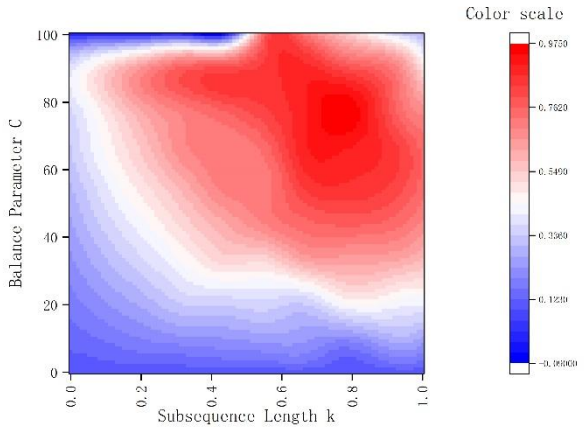

Figure 3. ASDMS method parameter mixed disturbance evaluation

We analysis the experimental results, subsequence length k have a great impact on the effect of the detection. Tested by trial and error， $k=9, C=0.83$ is the best parameter combination when the evaluation parameter AUC reaches a peak value of 0.91, the scheme has a good detection effect.

5. Conclusion

This paper proposes a Markov chain based on the sequence behaviour coding method and the user anomaly behaviour recognition scheme described in the support vector data domain. The classification coding and detection method can synthesize the similarity between the user behaviour sequence information and the user behaviour, and maximize the amount of information in the user behaviour sequence data. It can be seen from experiments that the scheme has the ability to detect abnormality of user behaviour sequences caused by unknown attacks or internal threats, and has better accuracy and timeliness performance. The scheme of this paper is applicable to the analysis and detection of user behaviour data in big data environment. It can be widely used in the detection of malicious behaviour of users in e-commerce, e-government, and important system intranets, and improve the ability of important information systems to prevent covert attacks.

References

[1] Atzori L, Iera A, Morabito G. The internet of things: A survey. Computer networks, 2010, 54(15): 2787-2805.
[2] Li, C., Tian, X.G., Xiao X. (2012) User behaviour anomaly detection method based on shell command and symbiotic matrix. Machine Research and Development, 49(9):1982-1990.
[3] Wu, X.D., Xie, F., Huang, Y.M. (2013) Sequential pattern mining with wildcards and one-off conditions. Journal of Software, (8):1804-1815.
[4] Song, H.T., Wei D.W., Tang G.M. User behavior anomaly detection algorithm based on pattern mining. (2016) Chinese Computer Systems., 37(2):221-226.
[5] Du, X.K., Li, G.H., Wang, J.K. (2015) Pattern matching method based on information element. Journal of Software, 26(10): 2596-2613
[6] Kang, Y.H., Zadorozhny, V. (2016) Process monitoring using maximum sequence divergence. Knowledge and information systems, 48(1): 81-109.
[7] Sarker, I.H., Salim, F.D. (2018) Mining user behavioral rules from smartphone data through association analysis. In Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer. Cham. 450-461.

[8] Gamachchi, A., Li, S., Boztas, S. (2018). A graph based framework for malicious insider threat detection. https://arxiv.org/abs/1809.00141.
[9] Yang, L.Q., Wen, J.Y., Liu S.F. Application of an improved graph segmentation algorithm in user behavior anomaly detection. Information network security. (6): 35-40.
[10] Lu, Y., Li W., Luo J.Z. (2014). A method for detecting abnormal behavior of network users based on selective collaborative learning. Chinese journal of computers. 37(1):28-40.
[11] Tang, C.H., Liu, P.C., Tang, S.S. Anomaly intrusion behavior detection based on fuzzy clustering and features selection. Journal of computer research and development. 52(3):718-728.
[12] Langville, A.N., Meyer, C.D. (2011) Google's pagerank and beyond. Mathematical intelligencer, 30(1):68-69.
[13] Lane, T.D. (2000) Machine learning techniques for the computer security domain of anomaly detection. [Ph.D. dissertation], The Purdue University:
[14] Ruff, L., Vandermeulen, R., Goernitz, N. Proceedings of the 35th International Conference on Machine Learning, PMLR 80:4393-4402(2018).