# Research on data mining methods of potential risk for large cloud computing network

Wang Chen

Shandong Institute Of Commerce And Technology, 250103

**Keywords:** cloud computing; potential risk; data mining;

**Abstract:** During the study process of mining methods for data of potential risk in large cloud computing network, when mining data of risk with traditional methods, principal element features and context features of data have significant fluctuations, which results ininaccurate data mining results. For this, a new large-scale cloud computing networks mining methods for data of potential risk is proposed. The algorithm gives a general definition to abnormal data from the view of geometry, reducing the dimension of vectors characteristics of data of potential risk under cloud computing, feature extraction algorithm is adopted to preprocess data of potential risk, and enlarge the distance between data of potential risk in dense distribution regions of samples, while, shorten the distance between samples in sparse distribution regions of samples, so as to prompt uniform overall distribution of sample library under cloud computing, and achieve accurate mining for data of potential risk in large-scale cloud computing networks. The simulation proved that the proposed mining method for data of potential risk in cloud computing network has higher mining efficiency and accuracy.

## 1 Introduction

Cloud computing is an Internet-based Supercomputing, integrates large-scale data and processor resources in the computer products, through relevant analysis to ensure that companies convert resources to a valuable application direction [1.2.3]. However, with the rapid development of network technology, criminal activity of potential risk in network is gradually increasing, results in the improving amount of data of potential risk under cloud computing environment [4.5.6]. Seeking effective data mining methods for ensuring the safety of related systems under the cloud computing environment is important, and attracts attention of many experts and scholars [7.8.9]. Since the mining method for data of potential risk under cloud computing network have broad space for development, it has become the research focus of industry, and been concerned widespread, there have been a lot of methods developed [10].

The most commonly used large-scale cloud computing networks mining methods for data of potential risk, including association rules algorithm, fuzzy rule algorithms and neural network algorithms and so on. However, when using the above method for data mining, principal element features and correlation characteristics of data still have significant fluctuations, which make it hard to obtain accurate data mining results.

For the above problem, a new large-scale cloud computing networks mining methods for data of potential risk based on multi-index abnormal data detection is proposed. The algorithm gives a general definition to abnormal data from the view of geometry, with non-linear manifolds learning algorithm to reduce the dimension of vectors characteristics of data of potential risk under cloud computing, feature extraction algorithm is adopted to preprocess data of potential risk, integrate improved classical manifold learning algorithm as well, and enlarge the distance between data of potential risk in dense distribution regions of samples, while, shorten the distance between samples in sparse distribution regions of samples, so as to prompt uniform overall distribution of sample library under cloud computing, and achieve accurate mining for data of potential risk in large-scale cloud computing networks. The simulation proved that the proposed mining method for data of potential risk in cloud computing network has higher mining efficiency and accuracy.

 

## 2 Mining principle of cloud computing network mining for data of potential risk

Through cloud computing, network data of risk digs principal elements characteristics information of samples, to establish mining models for data of potential risk and subdivide state of clouding environment data of risk, so as to obtain the same root section properties. The hierarchystructure is adopted to build identification model, each level of the model represents feature set of data of network risk, according to information entropy corresponding to each level, to mine data features of risk under corresponding cloud computing environment. Concrete steps are detailed below:

If the current number of nodes under the cloud computing environment represented by $k$, the number of sites identification represented by $p_i(i=1,2,\ldots,k)$, $B$ is the node attribute under current cloud computing environment, this property can be divided into $u$ sub-attributes $u_1,u_2\cdots u_u$, these sub-attributes can divide data of potential risk into $c$ subsets, and the root node of identification constituted by attribute $B$, probability of same root node possessing bit nodes of correlation represented by $H_i$ is $p_i$,or $\frac{|H_{i,D}|}{|D|}$. At this time, the data information entropy of sites are described below:

$$J(Z)=Pi\log 2(p) \qquad (1)$$

The target information described below:

$$w(B)=\frac{|Z_r|}{|D|}J_{(z)} \qquad (2)$$

Assuming $P,l$ indicates the number of nodes of potential risk under cloud computing, the amount of information on the division of identification information described as follows:

$$J=(P,l)=-\frac{p}{p+l}\log_2\frac{p}{p+l}-\frac{l}{p+l}\log_2\frac{l}{p+l} \qquad (3)$$

$$z(B)=\frac{P_i+l_i}{p+l}J(P_i,l_i) \qquad (4)$$

Information gain of risk attribute corresponding to nodes of risk under cloud computing is:

$$win(B)=J(P,l)-Z(B) \qquad (5)$$

In summary, miningprinciple of data of potential risk under cloud computing network is clarified, but with the traditional mining algorithms for data of risk, principal element features and associated features of data have significant fluctuation, which makes it difficult to obtain accurate data mining results.

## 3 Proposal and implementation of mining optimization method for data of potential risk under large scale cloud computing network

For the problem happened when the traditional mining algorithms for data of risk is adopted, principal element features and associated features of data have significant fluctuation, which makes it difficult to obtain accurate data mining results, a new mining method for data of potential risk under large-scale cloud computing network based on abnormal data algorithm is proposed.

3.1 quantification definition based on based on abnormal data

A abnormal data refers to more average features away from this data in a data set, at least possess $p*100\%$, from the opposite view, the data close to this data are less, at most occupy

$(1-p)*100\%$ . Therefore, from the essence, the so-called abnormal data refers to the data of relative isolation, whichis data have lessdata in neighborhood.

In the cloud computing network, each record of the basic form is called the original point, and each record is processed and transformed into numerical data, when the original point is called quantization points. Assuming, utility point is $D_i = \left(x_{i1}, x_{i2}, \cdots x_{im}\right)$, $i = 1, 2, \ldots, n$. Collection of all utility points called the utility point set, denoted as $uD$.

The distance between two points $D_P$ and $D_q$ is defined as follows:

$$d_k\left(D_P, D_q\right) = \left(\sum_{i=1}^{m} \left|xpi - xqi\right|^k\right)^{1/k} \quad (6)$$

Introducing the following definitions:

Definition 1: For arbitrary point $D_P$ in $uD$, a relatively small positive number $\delta \geq 0$ is given, if arbitrary point $D_q$ in $uD$ meet the requirements, $d_k\left(D_P, D_q\right) \leq \delta$ , then $D_q$ is $\delta -$ proximal point of $D_P$ ,and all $\delta -$, set of all adjacent points are called $\delta -$ neighborhood of point $D_P$.

Definition 2: For arbitrary point $D_P$ in $uD$, the critical value $n_0$ of experience is selected, as the case may be, assuming the number of points in $\delta -$ neighborhood of point $D_P$ is $N_P$, if $N_P \leq N_0$, this point $D_P$ is called outlier (or isolated points) of $uD$, also known as outlierof centered original points $D\left(\delta N_0\right)$.

3.2 Implementation of mining optimization method for data of potential risk under large-scale cloud computing network

The traditional method is suitable for studying in not closedrestriction linearity, it must be sufficiently smooth for sampling, and also more sensitive to noise, where the number of neighbors $k$ as an important parameter, which ranges is particularly important. Typically, premise of algorithm is uniformly dispersed and continuous extraction sample library, due to instability of network data characteristics,it is possiblysmaller or larger in a certain period of time, and overall distribution is uneven. Within the scope of sparse distribution of samples, the local neighborhood consisting of $k$ neighbors, is much larger than the local neighborhood constituted in the area of dense distribution of samples. Thusadvantages of first two methods are selected to propose a new method of selecting the number of neighbors $k$.

The number of neighbors is constituted by $k_{base}$ and $k_{add}$, namely:

$$k = k_{base} + k_{add} \quad (7)$$

Where, $k_{base}$ is a basic value, $k_{add}$ is an additional value, setting the value of $k_{base}$ based on the experience, and then calculate the distance of the sample point like $x_i$ and $k_{base}$ neighbor, and multiplied by a constant $h$. The following calculation contains reconstruction error $\varepsilon$ of additional neighbors, defined as:

$$\varepsilon\left(k\right) = \sum_{I=1}^{N} \left| x_i - \sum_{i=1}^{N} W_{ij} x_j \right|^2 \quad (8)$$

Since $k_{add}$ is different for data samples of potential risk under cloud computing, $\varepsilon$ can be used as a function of $k$, improving the number of optimal neighbors able to be selected in manifold learning algorithms is to achieve the minimum value $k$ of reconstruction function. The best value of $k_{add}$ is 6, the constant value is 1.1.

Under the conditions of choosing to optimize the number of neighbors, increasing distance between data samples of potential risk under cloud computing in dense distribution area of samples,

shorten the distance between samples in sparse distribution region, and prompting even overall distribution of data sample library, so as to reduce interference of the value of $k$ to data mining results under cloud computing.

By the following formula to calculate weight $w_{ij}$ between data points $x_i$ of potential risk and it's neighbors under cloud computing, namely minimization:

$$\varepsilon(w)=\sum_{i=1}^{n}\left|x_i-\sum_{j=1}^{n}w_{ij}x_j\right|^2 \tag{9}$$

By calculating the weight $w_{ij}$ between samples $x_i$ of high-dimensional space and it's neighbor $x_j$ to obtain data of potential risk in low-dimensional embedding space.

Assuming $d+1$is eigenvector corresponding the smallest eigenvalue in $M$ , the corresponding smallest eigenvector is abandoned, the remaining $d$ eigenvectorscompose matrix, and ultimately achieve comprehensive mining for data of potential risk under cloud computing network.

To sum up,the mining for data of potential risk under large-scale cloud computing network have high accuracy and applicability.

## 4 Experiments and simulations

In order to verify the effectiveness of data mining methods proposed in the paper, there is need for a implementation, the implementation of simulation is conducted through kdd-cup99 detection data sources.

Experiments were performed with improved algorithm and traditional algorithms, dimensionality reduction is conducted for five training subsets, and five classifiers were trained, the results are described in Table 1.

Table 1 Comparison of training time

| | Methods | ALL | DOS | PROBE | R2L | U2R |
|---|---|---|---|---|---|---|
| Training time (S) | Improved Method | 5 | 6.5 | 8 | 11 | 7 |
| Training time (S) | Traditional method | 25 | 36 | 83 | 20 | 72 |
| Mining Time (S) | Improved Method | 2.8 | 1.0 | 2.6 | 2 | 1.4 |
| Mining Time (S) | Traditional method | 7 | 4 | 15 | 6 | 9 |

It can be seen from Table 1, for the attacks of potential risk typeunder different cloud computing environment, the training time and mining time of proposed method is shorter than traditional methods, indicating that the method is effective.

In order to verify the improved method has a stronger performance than the traditional methods, a test library with new attack and without new attacks were utilized to compare two methods, the experiment regarded all attack as one type of attack, 7 experiments were carried out to obtain excavation accuracy curve of the two methods, as shown in Figure 1

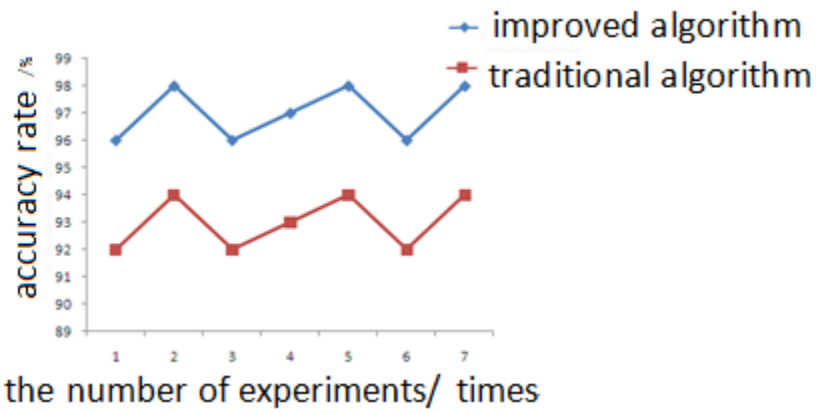


Figure 1. Comparison of the mining accuracy with two different methods

In general, the experimental data shows that improved algorithm ensure efficiency and accuracy of mining for data of potential risk under cloud computing network, and high value of application.

## 5 Conclusion

In this paper, for the problem of inaccurate data mining results caused by significant fluctuations of PCA features and context features of data, when mining data of risk with traditional methods, a new large-scale cloud computing networks mining methods for data of potential risk is proposed. The algorithm gives a general definition to abnormal data from the view of geometry, reducing the dimension of vectors characteristics of data of potential risk under cloud computing, feature extraction algorithm is adopted to preprocess data of potential risk, and enlarge the distance between data of potential risk in dense distribution regions of samples, while, shorten the distance between samples in sparse distribution regions of samples, so as to prompt uniform overall distribution of sample library under cloud computing, and achieve accurate mining for data of potential risk in large-scale cloud computing networks. The simulation proved that the proposed mining method for data of potential risk in cloud computing network has higher mining efficiency and accuracy.

## References

[1] Li Shan, Yu Xiaoyong, Gao Like, et al. Design of data detection platform based on CIM/XML power system model standard. [J] Guangxi electric power. 2014. 1: 9-12.

[2] Li Chengbing, Yao Chen. Study of recognizing discrepant traffic data and its validation [J]. Computer engineering and applications, 2013.20:244-246.

[3] Yang Donglin, Zhou Zhigang. Design and implementation of MAC protocol of mobile ad-hoc network based on multipoint data detection [J]. Information technology, 2014, 4:8-11.

[4] Liu Ruiqin, Liu Xuejun. Abnormal Data Stream Detection Based on Accelerating DTW in WSN [J]. Chinese Journal of Sensors and Actuators, 2013.6:887-893.

[5] Wang Bin, Wang Chao, Li Jin. A Big Difference for Network Anomaly Data Feature Detection Algorithm of the Simulation Analysis [J]. Computer simulation. 2013, 8:277-280.

[6] Lei Chenxi, Tang Xianghong, Li Shaobo. Algorithm for online data outlier detection of circuit breaker [J]. Application Research of Computers. 2014.6： 1706-1709.

[7] Xu Yuanbin, Zhong Xiaoqiang, Wang Dan, et al. abnormal detection of the power energy data parallelization based on MapReduce model [J]. Information research, 2014, 8:74-78

[8] Zhang Huaying. Abnormal Information Detection of Big Web Database Based on Chaotic Feature Analysis [J]. Bulletin of Science and Technology.2014.2:215-217.

[9] Cao Xu, Cao Ruitong. The method for abnormal network detection based on big data analysis. Telecommunication science. 2014.6:152-156.

[10] Qu Guangqing, Bian Zhendong, Li Hong. Research on the data analysis for detecting abnormal energy consumption [J]. China plant engineering, 2014, 1:51-53.