Information Security Risk Assessment Based on Information Measure and Fuzzy Clustering

Guo-hong Gao
School of Information Engineer, Henan Institute of Science and Technology, Henan Xinxiang, 453003, CHINA
Sanyuerj03@126.com

Xue-yong Li, Bao-jian Zhang and Wen-xian Xiao
School of Information Engineer, Henan Institute of Science and Technology , Henan Xinxiang, 453003, CHINA
emtlxy@gmail.com

***Abstract*—To address the problems of lack of training data and difficult to find optimal value in information security risk assessment, this paper applying a new information measure method and fuzzy clustering in information security risk assessment. The new method quantifies risk factors of all data and the dependence degree of safety with the mutual information computing. Then search optimal points in each degree of risk as original center points of K-means clustering algorithm, and use the K-means clustering algorithm for data classification. This method has less computation, and it can overcome the K-means's shortcoming of sensitive to initial value and problem of nonlinear and complexity of information security risk assessment. Experimental results show the effectiveness of our method.**

***Index Terms*—Information security, Risk assessment, Information Measure, Fuzzy Clustering**

I. INTRODUCTION

Risk assessment began to be used nuclear power plant safety assessment of Europe and America early in the 1960's, and subsequently developed and applied in aerospace engineering, chemical industry, environmental protection, health, transportation, promotion of the national economy, and many other fields. E-government information security is one of the prerequisites the economic and social information, and an important part of building the national information. How to ensure the reliability of e-government information security, many scholars on the issue at home and abroad have studied information security, from different aspects of information security and the different angles.

Today broadband networks and high-capacity electronic data storage technologies enable organizations and individuals to create, receive, store, access and publish information in quantities and at speeds and economies that remain impossible with physical forms of data. Organizations have embraced electronic forms of information for their ability to accelerate the pace of any information-based activity. Electronic forms of data have substantially replaced physical forms of data for most organizations.

Universal applications of Internet technology promote e-government and e-commerce booming, and gradually brought many security threats and security risks. Information security risk assessment has been attached great importance by government, military, business and scientific research institutions [1-4]. In the information age, information security can be seen as to protect information confidentiality, integrity, availability, authenticity, controllability, and defense and confront threat to national political, economic, and cultural security in the information domain, the process of adopting effective strategies. Information security is not only their information security, but also has great strategic value to the national security.

Organizations are increasingly relying on information systems to enhance business operations, facilitate management decision-making, and deploy businesses strategies. The dependence has increased in current business environments where a variety of transactions involving trading of goods and services are accomplished electronically [5,6]. Increasing organizational dependence on the IS has led to a corresponding increase in the impact of information systems security abuses. Therefore, the information systems security is a critical issue that has attracted much attention from both information systems researchers and practitioners.

At present, the main research achievements on information security risk assessment domestic include OCTAVE method[7], SP800-42 [8], PRA [9], SP80030 [10], etc. But these standards and methods have some shortcomings, some of them are simple qualitative analysis method, or having given a quantitative method, but it is too cumbersome to implement.

As information security risk assessment has some characteristics, such as nonlinearity, complexity of operation and subjectivity of issue, the characteristics caused it have some limitations to use the traditional model to conduct information security risk assessment. These traditional assessment methods have greater subjective arbitrary and vagueness, so they are more complex in operation.

This paper proposed a new information security risk assessment method based on combination the mutual information calculated with K-means clustering algorithm.

Manuscript received October 10, 2010; revised February 1, 2011; accepted March15, 2011.

Guo-hong Gao, Henan Institute of Science and Technology, Xinxiang, 453003, China (E-mail:Sanyuerj03@126.com)

doi:10.4304/jsw.6.11.2159-2166

In order to achieve assess effectively the level of information security risk factors, the new method quantifies risk factors of all data and the dependence degree of safety with the mutual information computing. Then search optimal points in each degree of risk as original center points of K-means clustering algorithm, and use the K-means clustering algorithm for data classification. This method is easy to implement, simple to calculate. The shortcoming of K-means is sensitive to original center points, our method can effectively avoid this problem, while overcome the nonlinear and complexity of information security risk assessment. Experimental results show the effectiveness of our method.

The remainder of the paper is organized as follows. Section 2 describes the information system security risk. Section 3 proposes risk assessment of information security based on mutual information and K-means. Then show the experimental results in section 4, and summarize out work in section 5.

II. INFORMATION SYSTEM SECURITY RISK

Information system security risk, defined as the product of the monetary losses associated with security incidents and the probability that they occur, is a suitable decision criterion when considering different information system architectures.

Security issues related to information technology continue to be a concern in today's society, and for decision makers in it. Security is a complex property, and several diverse factors need to be considered to assess the security of a system's architecture. To support decision makers a plethora of approaches, frameworks and methods has been proposed for analyzing and ranking security – all with some explicit or implicit definition of security.

The well-known information security standard, ISO/IEC 17799:2005, states in its introduction:

Information can exist in many forms. It can be printed or written on paper, stored electronically, transmitted by post or using electronic means, shown on films, or spoken in conversation. Whatever forms the information takes, or means by which it is shared or stored, it should always be appropriately protected.

Information security is the protection of information from a wide range of threats in order to ensure business continuity, minimize business risk, and maximize return on investments and business opportunities.

Information security refers to the components, procedures and data of information systems are not destroyed altered or leaked, due to accidental or malicious reasons, moreover the system run continuously and reliably, the service was not interrupted. Nature of the problem lies the resources of information systems is in the value and vulnerability, and easy to trigger risks. The vulnerability of the information system is the inherent reason which caused security problems; the risk faced by the information system is external reasons [11].

Information security risk assessment, from the perspective of risk management, analyzes systematically facing threats and existing vulnerabilities of network and information systems with scientific methods and means. Assess potential harm in case of threatening events happened and put forward targeted defense countermeasures against the threat and rectification measures, to prevent and resolve information security risks, or control the risks at an acceptable level, thus the protect of network and information security to the maximum extent. Information security risk assessment has three main factors [11]: threat identification, vulnerability identification, asset identification, as shown in Figure 1:

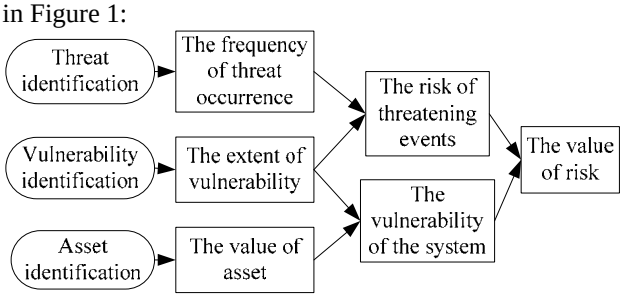


Figure 1 Elements of information security risk assessment

In the process of information security risk assessment, the risk calculation is the important stage to determine the risk level, the main process is following:

(1) Identify information assets, and set assignment assets values;

(2) Analyze threats, and set the likelihood of threats occurrence values;

(3) Identify the vulnerability of information assets, and set the extent of vulnerability values;

(4) Calculate the probability of safe event occurrence according to threats and vulnerabilities;

(5) Combined with the importance of information assets and the possibility of security incidents occurrence on these information assets, calculate the risk value of information assets.

We illustrate the risk computing with the following canonical form:

$$\text{riskvalue} = R(A,T,V) = R(L(T,V), F(C_a, V_a)) \quad (1)$$

R represents the computing function of security risk;

A represents asset;

T represents threat;

V represents vulnerability;

C_a represents the value of assets brought by security events

V_a represents the extent of vulnerability

L represents the possibility of threats led to the of security incidents by using vulnerability of assets;

F represents the losses caused by security events.

The determining of risk value is related to the results of risk assessment and the drawing up measures of risk control, so is the important and difficult stage in the risk assessment process. This is the core issue may researcher studied and made every effort to solve. Note that R is a complex function, not a linear relationship.

In order to prevent security breaches, businesses use controls (and various countermeasures) to safeguard their assets from various patterns of threats by identifying the information system assets that are vulnerable to threats. But, even in the presence of controls, the assets are often not fully protected from threats because of inherent control weaknesses. Thus, the risk assessment is a critical step for the information system security risk management.

In practice, the information system security risk assessment is quite complex and full of the uncertainty as well [12]. The uncertainty, existing in the process of assessment, has been the primary factor that influences the effectiveness of the information systems security risk assessment to a large extent. Therefore, in order to deal with the incompleteness and vagueness of information, the uncertainty must be taken into account in the information systems security risk assessment. However, most existing approaches applied to the ISS assessment have some drawbacks on handling uncertainty in the process of assessment.

III. Risk Assessment of Information Security Based on Mutual Information and K-means

In the information security risk assessment, risk elements main reflect in the assets of system, the existing threats and vulnerabilities, and risk analysis assess the level of risk factors from evaluation indexes, such as the frequency of threat occurrences, severity degree of the vulnerability, the value of asset, etc. The evaluation indexes have some properties of considerable ambiguity and uncertainty, so conventional methods are difficult to measure. Moreover, these evaluation indexes and risk of level of risk factors are nonlinear and dynamic relationship, so the conventional methods are difficult to process. This paper proposed a new information security risk assessment method based on mutual information calculation and K-means clustering algorithm, to address the problems of lack of training data and difficult to find optimal value. First, we used fuzzy evaluation method to quantify risk factors, and then calculated the mutual information value of risk factors to indicate the dependent degree of risk factors and risk levels. The data were classified by K-means with the optimal mutual information data as the initial cluster centers. This method is simple and less computation, and it can overcome the K-means's shortcoming of sensitive to initial value and can solve problems of difficult to find optimal value and defects of conclusion blurring.

A. K-means algorithm

Cluster analysis is a common unsupervised learning technique for statistical data analysis, which seeks to group objects of a similar kind into separate categories. It is widely used indifferent fields, including social sciences, database marketing and bioinformatics. Cluster analysis encompasses a number of heuristic and model-based methods, including the K-means algorithm and the normal mixture model (MM) method.

Data clustering is used frequently in a number of applications, such as vector quantization (VQ), pattern recognition, knowledge discovery, speaker recognition, fault detection, and web/data mining. Among clustering formulations that minimize a cost function, k-means clustering is perhaps the most widely used and studied [18]. The k-means clustering algorithm, which is also called the generalized Lloyd algorithm (GLA), is a special case of the generalized hard clustering scheme, when point representatives are adopted and the squared Euclidean distances are used to measure the distortion (dissimilarity) between a vector X and its cluster representative(cluster center) C. The k-means clustering algorithm performs iteratively the partition step and new cluster center generation step until convergence.

The K-means algorithm is a nonparametric approach that aims to classify objects into K mutually exclusive clusters by minimizing the expected squared distance of an object from its nearest center.

The K-means clustering algorithm partitions data into k clusters S_i ($i=1,2,\ldots,k$) and the cluster S_i is associated with a representative (cluster center) C_i. Denote the set of data points as $S=\{X_m\}$. $m=1,2,\ldots,N$, N is the number of data points in the set S. Let $d(X,Y)$ be the distortion between any two vectors X and Y.

Let C_{mm} be the nearest cluster center of X_m and $d_m = d(X_m, C_{mm})$. The goal of K-means clustering is to find a set of cluster centers $SC\{C_l\}$ such that the distortion J defined below is minimized, where $l=1$ to K and K is the number of clusters.

$$J = \sum_{m=1}^{N} d_m \tag{2}$$

The major process of GLA (K-means clustering) is mapping a given set of representative vectors into an improved one through partitioning data points. It begins with an initial set of cluster centers and repeats this mapping process until a stopping criterion is satisfied. The Lloyd iteration for the K-means clustering is given as follows:

(a)Given a set of cluster centers $SC_p=\{C_i\}$, find the partition of S; that is S is divided into K clusters S_j ,where $j=1,2,\ldots,k$ and $S_j=\{X \mid d(X,C_j) \le d(X,C_i)$ for all $i \ne j\}$.

(b)Compute the centroid for each cluster to obtain a new set of cluster representatives SC_{p+1}.

The *k*-means clustering algorithm is briefly described as follows:

(1) Begin with an initial set of cluster centers SC_0. Set $p=0$.

(2) Given the set of cluster centers SC_p, perform the Lloyd iteration to generate the improved set of cluster representatives SC_{p+1}.

(3) Compute the average distortion J for SC_{p+1}. If it is changed by a small enough amount since the last iteration, then stop. Otherwise set $p+1 \to p$ and go to step (2).

The nearest cluster center is determined by computing the distance between each cluster center and a data point. In this paper, distance function adopts Euclidean distance. The Euclidean distance between a data point $X = (x_1, x_2, \ldots, x_d)^t$ and a cluster center $C = (c_1, c_2, \ldots, c_d)^t$ is defined as

$$d(X, C) = \left[\sum_{i=1}^{d} |x_i - c_i|^2 \right]^{0.5} \quad (3)$$

To determine the best match (nearest cluster center) of a data point, K squared-error computations (distortion computations) are needed, where K is the number of clusters.

However, the K-means algorithm is a local search procedure and it is well known that suffers from the serious drawback that its performance heavily depends on the initial starting conditions [18]. In our method, we use mutual information computing to solve this problem.

B. Evaluation model

The basic idea of information security risk assessment model based on mutual information computing and K-means algorithm is following: First, quantify the risk assessment index with fuzzy evaluation approach, then search the optimal points in each risk level with mutual information computing after quantified and compatibility processing, which as the initial cluster centers of K-means clustering algorithm. The model structure is shown in Figure 2.

As shown in Figure 2, the degree of security level of information systems is divided into four classes respectively defined as L_1 , L_2, L_3 and L_4 . L_1 represents the minimum security level , L_4 represents the highest security level. There are 4 clusters in the K-means clustering results, so K=4.

Algorithm description:

(1) Information format of the original data obtained from the information system is not in line with our method, it must be processed to be vector form for our method requested. Because the index values of risk factors have property of uncertainty, in this paper, we use fuzzy evaluation approach to preprocess the indexes the information security risk factors;

(2) Data processing:

The data which have been input into our algorithm are divided into two classes: training data and test data. The main useful of training data is training the K-mean clustering algorithm. This is to say, K-means algorithm clusters with training data, to find original clusters and the cluster centres. this is the process of knowledge discovery. Then the K-means clustering algorithm classifies the test data with the original clusters and the cluster centres.

For training data,

a) Calculate mutual information value between the preprocessed the data and the given four optimal points set corresponding to the risk level L_1, L_2, L_3 and L_4 for each training data. The mutual information (MI) is defined as the reduction in uncertainty of one random variable due to knowing about another, or in other words, the amount of information one random variable contains about another.

Our model has a data set, storing the optimal points set corresponding to the risk level L_1, L_2, L_3 and L_4 which got through expert analyzed a large amount of data.

b) For each risk level, calculate the mutual information values between all data, and then we compute the sum of these mutual information values which got by the last step, so we can get the average value of all mutual information with the mutual information sum.

c) Get four average values of mutual information corresponding respectively to the for information security risk levels, then to search four optimal points which is nearest to the average values as the initial cluster centers of K-means clustering algorithm.

For the test data, compute the distance between each cluster center and each test data, to determine the nearest cluster center to join, cluster with K-means algorithm which already owned the initial cluster centers, to generate the new clustering results.

(3) Judge whether the clustering results change or not, If some data's cluster labels changed in the new cluster results, it indicates that the new cluster centers may be improper, the improper cluster centers may caused the higher false detect rate. In order to ensure and improve the detect accuracy; we re-calculate the mutual information value of all data to find the optimal points as the initial cluster centers of K-means algorithm.

C. Fuzzy preprocessing

In the information security risk assessment, risk factors assessment indexes have vague property and uncertainties, our method quantify information security risk factors with fuzzy evaluation approach. Concrete steps are as follows [11]:

(1) Correlation analysis on assets, vulnerabilities, threats and relationship of threat and vulnerability to identify information security risk factors.

(2)Based on fuzzy evaluation approach, create risk factors set $U = \{u_1, u_2, \ldots, u_n\}$

(3) Create evaluation set. Evaluate various risk factors from the confidentiality and integrity of assets, the degree of vulnerability, the technical content of threat and other aspects. Experts give the risk reviews for various risk factors, each risk reviews of the risk factor is divided into m grades, evaluation set is $V = \{v_1, v_2, \ldots, v_m\}$.

(4) Expert give risk review of each risk factor, create fuzzy mapping $f : U \to F(V)$, $F(V)$ is all fuzzy set on

V , $u_i \to f(u_i)$, where $u_i = (r_{i1}, r_{i2}, \ldots, r_{im}) \in F(V)$ mapping function f represents risk factor u_i 's support factor u_i 's attached vector to evaluation set is degree for each reviews in evaluation set.

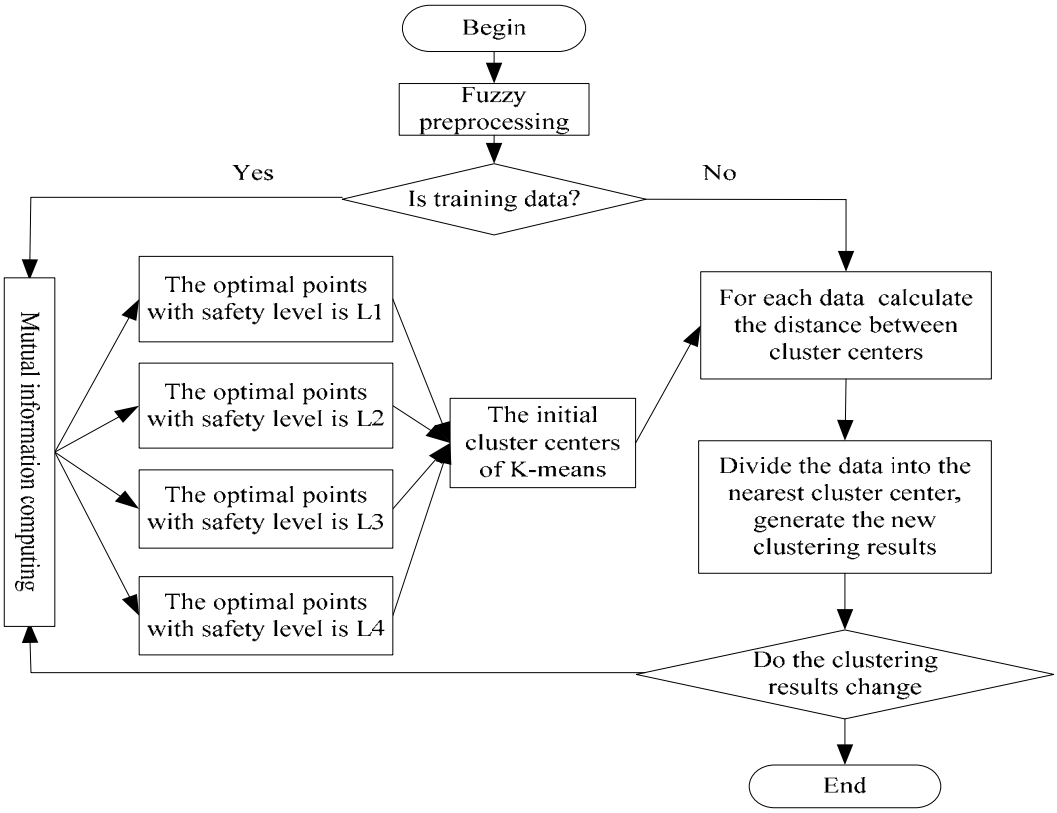


Figure 2 Structure of assessment model

Set risk $R_i = \{r_{i1}, r_i 2, \ldots, r_{im}\}, i = 1,2,\ldots,n$, then get the attached matrix R .

(5) The value of index in Evaluation set affect directly the degree of risk, so given the each evaluation index a weight. Set the weights distribution set is $A = (a_1, a_2, \cdots, a_n)$. By the fuzzy transformation operator, get

$$B = A \bullet R^T = (a_1, a_2, \ldots a_n) \begin{bmatrix} r_{11} & r_{21} & \cdots & r_{n1} \\ r_{12} & r_{22} & \cdots & r_{n2} \\ \vdots & \vdots & & \vdots \\ r_{1m} & r_{2m} & \cdots & r_{nm} \end{bmatrix}$$

$$= (b_1, b_2, \ldots, b_n) \quad (4)$$

B represents weights of all risk factors under an evaluation, it reflects the evaluation view of risk factors, the value is in $(0,1)$.

D. Mutual information computing

Mutual information is an important concept of information theory which measures the statistical dependence between two variables, i.e. the information that one variable carries about the other. It was first proposed as a registration measure in medical image registration in 1995, independently by Viola and Wells. It is an important concept in information theory field, which measures degree of interdependence between two messages, commonly used in text classification and feature's reduction and select [14-16]. Mutual information is also widely used to ascertain one of the most important parameters, time delay, in reconstructing phase space from nonlinear time series.

Mutual information is defined as the reduction in uncertainty of one random variable due to knowing about another, or in other words, the amount of information one random variable contains about another. In multi-word detection, mutual information can be defined as the amount of information provided by the occurrence of the word represented by Y about the occurrence of the word represented by X . Church and Hanks [17] proposed the association ratio for measuring word association based on the information on theoretic concept of mutual information.

We use mutual information measure statistical correlation degree between risk factor evaluation index and risk level, which is defined as:

$$I(a_k, d) = p(a_k \mid d) lb \frac{p(a_k, d)}{p(a_k) p(d)} \quad (5)$$

Where

$p(a_k) = |a_k| / N_1$, a_k is an attribute (specific value of an risk factor evaluation index), N_1 represents the number of different values of risk factor evaluation index in all sample data, $p(a_k)$ represents the probability of a_k appearance, namely the ratio of the number a_k appearance in all evaluation factors of the training data and the total number of evaluation factors values, the bigger $p(a_k)$, the a_k is more common and more probability of security.

$p(d) = |d| / N$ represents the probability of training data which belongs to level d (this paper divided the security degree into 4 levels, so the value of d may be one of L_1, L_2, L_3 and L_4).

$p(a_k \mid d) = |a_k \cap d| / |d|$ represents conditional probability of a_k, $|a_k \cap d|$ represents the number of sample data, whose attribute value is a_k and the risk level belongs to d, $|d|$ represents the total number of sample data, whose risk level belongs to d.

$p(a_k, d) = |a_k \cap d| / N$ represents the probability of a_k appearance while the data whose attribute is a_k belongs to risk level d

The value of $I(a_k, d)$ indicates the associated degree of a_k and risk level d, the bigger the value, the higher associated degree is. So a_k is greater contribution to level d.

Suppose evaluation of risk factors are took into account from the four factors, including the confidentiality and integrity of assets, the extent of vulnerability and the technical content of threat. Security level is divided into four grades, L_1 represents that the minimum security, L_4 represents the highest safety. We select eight data to fuzzy pre-progress, the results as shown in Table 1.

TABLE I SECURITY LEVEL

Data Number	b_1	b_2	b_3	b_4	Security Level
D1	0.1	0.3	0.3	0.2	L1
D2	0.5	0.3	0.4	0.3	L2
D3	0.5	0.3	0.4	0.3	L2
D4	0.3	0.7	0.5	0.3	L4
D5	0.3	0.3	0.2	0.3	L2
D6	0.3	0.1	0.3	0.3	L1
D7	0.5	0.3	0.3	0.5	L3
D8	0.3	0.3	0.1	0.1	L1

We first count the number of attribute values for each attribute, such as the number of 0.1 in b_1 is 1, So mutual information computing of the value 0.1in attribute b_1 corresponding risk level L_1 is:

$$I(0.1, L_1) = p(0.1 \mid L_1) lb \frac{p(0.3, L_1)}{p(0.3)\, p(L_1)} = 0.12 \quad (6)$$

E. Risk Assessment of Information Security Based on Mutual Information and K-means

K-means clustering algorithm is one of the main clustering analysis algorithms, which uses iteration method to update; the ultimate consequence is to obtain the minimum objective function and to achieve the optimum clustering effect. K-means algorithm works as follows: First, randomly choose k data objects from n data as initial clustering centers. For the rest objects, according to their similarity with the cluster center (distance), respectively assigned them to the most similar (represented by cluster center) clustering; then recalculate the cluster center for each a new acquired cluster; repeats this process until the standard measure of function began to convergence, generally using the average variance as the standard measure function.

The traditional K-means algorithm randomly select the initial cluster centers, with the different initial input value the clustering results easy to fluctuate, then cause the detection result has greatly deviation, moreover needs more iteration times. In our method, let mean mutual information value as the initial cluster centers, so at the beginning of clustering, cluster centers are optimal values, so our method can overcome the shortcomings of sensitivity to the initial value, also reduces the number of iterations. The distance function adopts Euclidean distance function.

For given training data, the data form pre-progressed is shown in Table 1, for data D_i, we can compute $I(b_i, L_i)$,which represents the contribution degree of the all attributes b_i to level L_i. L_i represents the level grade corresponding to D_i. We define SI as the n mutual information sum of n attributes of D_i.

$$SI(D_i) = \sum_{i=1}^{n} \alpha \cdot I(b_i, L_i) \quad (7)$$

Where α is a weight corresponding to each attribute, aims to adjust the attribute's ratio in $SI(D_i)$. The bigger $SI(D_i)$, indicates D_i has the bigger contribution to L_i. We define average mutual information value of all data, whose degree belongs to L_j:

$$AVSI(L_j) = \sum_{i=1}^{N_i} SI(D_i) / N_i, 1 \le j \le 4 \quad (8)$$

Where N_i represents the number of data whose level belong to L_j. $AVSI(L_j)$ represents the average

contribution value of data to L_j, whose level belong to L_j.

IV. EXPERIMENT

The experimental samples are from information systems of an e-commerce company. In our experiment, we divided the three main factors (threat identification, vulnerability identification, asset identification) which impact information security risk assessment into 10 specific factors [18]: information stolen, deleted or lost; network resources are destroyed; information abuse, false use or tampering; service disruption and prohibition; hardware defects; network vulnerabilities; data leaks; interference with communication; information and service restoration, interruption, delay, weaken. Please expert assess these risk factors, give weight to each factor, as input data. Divide assessment results into four categories: high-security, general security, suspected and dangerous, respectively as L_4, L_3, L_2 and L_1.

The data includes training data and test data. For training data, each training data can be expressed as a 1 × 10-dimensional vector, that is $R_i = [A_1, A_2, \ldots, A_{10}]^T$. In experiment, we select 200 training data shown in Table 2. The data in Table 2 which went through fuzzy preprocessing as training data, 10 security indexes of each sample have been calculated. And 100 test data as shown in Table 3, also list 10 security indexes. All data are divided into 10 groups; each group includes 20 training data and 10 text data to detect our algorithm.

TABLE II TRAINING DATA

Data Number	A_1	A_2	…	A_{10}	Security Level
D1	0.4	0.3	…	0.4	L3
D2	0.5	0.3	…	0.3	L2
D3	0.5	0.3	…	0.3	L2
D4	0.3	0.7	…	0.3	L4
D5	0.3	0.3	…	0.3	L2
D6	0.3	0.1	…	0.3	L1
D7	0.5	0.3	…	0.5	L3
⋮	⋮	⋮	⋮	⋮	⋮

TABLE III TEXT DATA

Data Number	A_1	A_2	…	A_{10}	Security Level
D1	0.6	0.3	…	0.4	L3
D2	0.3	0.3	…	0.3	L1
D3	0.5	0.3	…	0.4	L2
D4	0.4	0.7	…	0.6	L4
D5	0.4	0.4	…	0.3	L2
⋮	⋮	⋮	⋮	⋮	⋮

The detection accuracy rate of out method reached 98% for data in Table 3. Experimental results show that our method is effective, moreover has less computation than the method in [18], detective speed is fast. Our risk assessment method can more accurately assess the risk level of information systems, to more scientifically guard against the risk.

V. CONCLUSION

In order to solve the problem that information security risk assessment has less training data, difficult to solve the optimal value, this paper proposed a new information security risk assessment method based on mutual information computing and K-means clustering algorithm, to effectively assess information security risk factors level. Our method quantifies the dependency degree between risk factors and security level with mutual information computing. On each risk level, find the optimal points as the K-means algorithm initial cluster centers and then use the K-means clustering algorithm to class the data, and our method can dynamically adjust the cluster center according to the clustering results and mutual information computing value. This method is easy to achieve, and has less computation, avoid the K-means sensitive to initial value problem, and moreover overcome nonlinearity, complexity and other problems of information security risk assessment . Experimental results show that our method is excellent.

REFERENCES

[1] Feng Dengguo, Zhang Yang, Zhang Yuqing. Survey of information security risk assessment[J]. Journal of china Institute of communications, 2004,25(7):10-18

[2] Li Honglian, Wang ChunHua, Yuan Baozong, Zhu Zhanhui. A Learing Strategy of SVM Used to Large Traing Set [J]. Chinese Journal of Computers. 2004, (5) : 715-719.

[3] Zhang Ming, Yin Ping, Deng Zhihong, Yang Dongqing. SWM+BIHMM:A Hybrid Statistic Model for Metadata Extraction[J]. Journal of Software, 2008, 19 (2) :358-368.

[4] M. Karyda, E. Kiountouzis, S. Kokolakis, Information systems security policies: a contextual perspective, Computers and Security 24 (3) (2005) 246–260.

[5] A. Kankanhalli, H.H. Teo, B.C.Y. Tan, K.K. Wei, An integrative study of information systems security effectiveness, International Journal of Information Management 23 (2) (2003) 139–154.

[6] M. Karyda, E. Kiountouzis, S. Kokolakis, Information systems security policies: a contextual perspective, Computers and Security 24 (3) (2005) 246–260.

[7] Maiwald E. Netwok Secuity: A Begginner's[D]. The McGraw-Hill Companies, Inc，2001.

[8] ISO/IEC 17799. Information Technology-Code of practic for information security management [S]. .2000.

[9] MnSCU. Security Risk Assessment-Applied Risk Management[R]. Minnesota State Colleges & Universities, 2002, 7.

[10] Whitman M E, Herbert J. Principles of Information Security[M]. Canada:GEX Publishing Services,2003.

[11] Zhao Dongmei, Liu Jinxing, Ma Jianfeng. Risk assessment of Information Security Based on Improved

Wavelet Neural Network. Computer Science, 2010,37(2),90-93

[12] R.L. Winkler, Uncertainty in probabilistic risk assessment, Reliability Engineer- ing and System Safety 54 (2–3) (1996) 127–132.

[13] T. Kanungo, D. Mount, N. Netanyahu, C. Piatko, R. Silverman, A. Wu, An efficient k-means clustering algorithm :analysis and implementation, IEEE Trans. Pattern Anal. Mach. Intell. 24(7)(2002)881–892

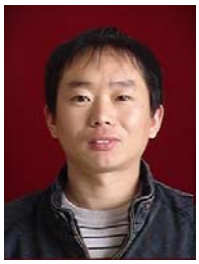
[14] J.A. Lozano, J.M. Pena, P. Larranaga, An empirical comparison of four initialization methods for the k-means algorithm, Pattern Recognition Lett. 20 (1999) 1027 – 1040.

[15] Zhang Wenliang, Huang Yalou, and Ni Weijian. Approach to Feature Selection of Spam Filtering Based on Contribution Defference. Computer Engineering 2007. 33(8): 08-82

[16] Wang Weiling, Liu Peiyu, Chu Jianchong. Improved feature selection algorithm with conditional mutual information [J].Computer Applications, 2007,27(2):433-435

[17] Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. Computational Linguistics, 16(1), 22 – 29.

[18] Dang Depeng, Meng Zhen. Assessment of information security risk by support vector machine[J]. Journal of Huazhong University of Science and Technology(Natural Science Edition), 2010, 3(38):46-49.



Gao Guohong (1975-), was born in Zhengzhou, China. He received his B.S degree in 2000 form Computer and Applications, Henan normal university in Xinxiang, his M.S degree in 2008 form School of Computer Technology, Huazhong University of Science and Technology, and enroll in Wuhan University of Technology in 2009, work hard at D.S degree. Currently he is a professor in the School of Information Engineer, Henan Institute of Science and Technology, Henan Xinxiang, 453003, China. The main publications include: Compute Operating System ;network and information security computer software.