
Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 6, pp. 5295–5301, DOI 10.12694/scpe.v25i6.3247

CONSTRUCTION METHOD OF INFORMATION SECURITY DETECTION BASED ON CLUSTERING ALGORITHM

SHAOBO CHEN*

Abstract. A method for K-prototype clustering, which can process mixed data types, is proposed first. An algorithm for information security evaluation of hybrid clusters based on K-prototypes was constructed. This method makes full use of the excellent global optimization performance of PSO and effectively solves the defect that the K-protection function quickly falls into local optimization. Simulation results show that the proposed method can effectively prevent local extreme values and improve overall performance.

Key words: Clustering; Information security; Safety evaluation; PSO algorithm; K-prototypes

1. Introduction. In enterprise information construction, various security devices will generate massive records. These log data are an essential source of threat information for information system security assessment. Therefore, it is a crucial issue in network security evaluation to quickly and effectively extract potential threats from extensive network data [1]. In recent years, more and more data mining methods have been applied to the security evaluation of information systems, and log-based cluster analysis has become a hot research direction. Cluster analysis is to classify something into one class according to the similarity of a particular class [2]. The similarity of the same class is high, and the similarity of different classes is low. Cluster analysis has been increasingly applied in data mining, pattern recognition, machine learning, image processing, etc. The existing clustering algorithms mainly include hierarchical clustering, partition clustering, density clustering and raster clustering.

K-means is a standard clustering algorithm based on blocks. The K-means algorithm transforms problems into multi-dimensional combinatorial optimal problems [3]. It groups a series of samples with several clusters and objective functions as constraints. So, you get the optimal solution. The classic K-Means method is fast and efficient. The disadvantage is that it only applies to numeric type data, is more sensitive to initial values, and only applies to spherical distribution data. Many improvements have been proposed to overcome some of K-Means' problems. Some scholars use genetic algorithms, particle swarm optimization, immune planning, ant colony, and other heuristic optimization algorithms to prevent algorithms from falling into the local optimal dilemma [4]. Algorithms such as K-modes and K-protection have been studied for the two types of mixed problems.

At present, the recorded data of security evaluation contains many characteristics of symbol type and number type, and some have strong characteristics of symbol type, such as network protocol type and network service type. They cannot be eliminated directly [5]. For this reason, the project intends to study a new way of clustering multi-source log data based on K-prototypes, combined with PSO's excellent global optimization performance, to solve the defects of the K-protection package algorithm that easily fall into local extreme values. Simulation results show that the proposed method can effectively prevent local extreme values, improve the overall convergence, and improve the accuracy and stability of the algorithm.

2. Clustering algorithm based on fuzzy K-prototypes. This paper introduces a fusion method of fuzzy K-means and fuzzy K-mode algorithms, which can effectively deal with data with different properties. Suppose $U = \{u_1, u_2, \cdots, u_n\}$ is a collection of data objects, and $u_i = \{E_{i1}, \cdots, E_{iq}, E_{i(q+1)}, \cdots, E_m\}$ represents that the data objects have m properties. The preceding q terms are continuous, and the $q+1$ through

*School of Mathematics and Computer Science, Shaanxi University of Technology. Hanzhong 723000, China (sxlgcsb@163.com)

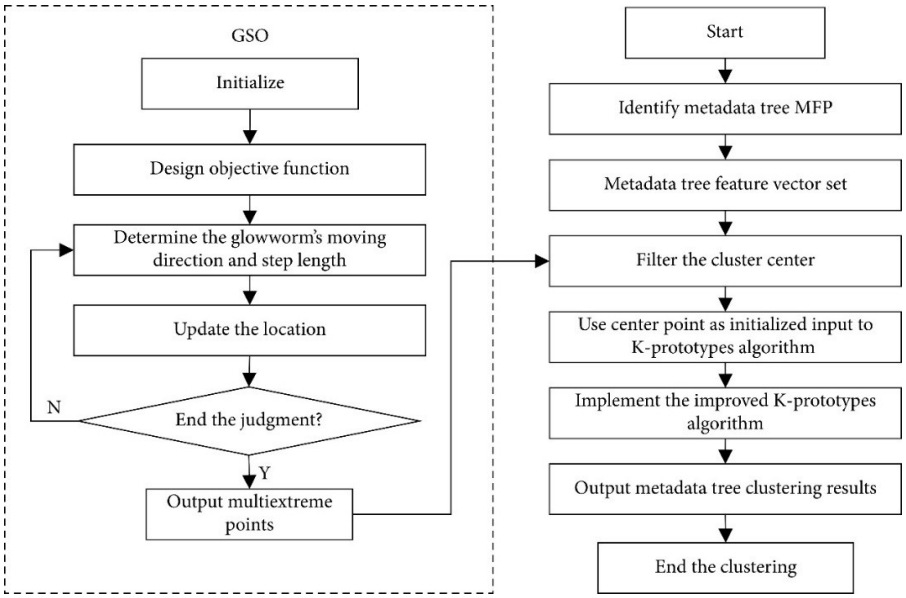


Fig. 2.1: *Flow of the clustering algorithm for fuzzy K-prototypes.*

m terms are categorical [6]. Therefore, the difference between data targets u_i and u_j can be obtained through calculation in formula (2.1). Figure 2.1 shows the process of the fuzzy K-prototypes clustering algorithm.

$$s\left(u_i, u_j\right) = \left[\sum_{h=1}^{q}\left(u_{ih} - u_{jh}\right)^2\right]^{1/2} + \mu \sum_{h=q+1}^{m} \gamma\left(u_{ih}, u_{jh}\right) \tag{2.1}$$

In (2.1), the former represents the difference in adjacent properties and the latter in class properties. μ is the weight used to adjust the degree of difference between two properties, called the "property ratio." The definition of $\gamma\left(u_{ih}, u_{jh}\right)$ is given in formula (2.2):

$$\gamma\left(u_{ih}, u_{jh}\right) = \begin{cases} 1, & u_{ih} \neq u_{jh} \\ 0, & u_{ih} = u_{jh} \end{cases} \tag{2.2}$$

The objective function of the fuzzy K-prototypes clustering algorithm is expressed in formula (2.3):

$$G(X, Y) = \sum_{k=1}^{z}\sum_{i=1}^{n}\left(x_{ki}\right)^{\partial} s\left(u_i, y_k\right) \tag{2.3}$$

$x_{ki} \in [0,1], \sum_{k=1}^{z} x_{ki} = 1, 0 < \sum_{i=1}^{n} x_{ki} < n; X$ is the dependency coefficient matrix of $n \times z.n$ is the number of data objects. Where z is the number of clusters. x_{ki} is the extent to which the i object belongs to the k cluster. Where Y is the cluster center of mass set, $Y = \{y_1, y_2, \cdots, y_z\}$. Where $\partial \in [1, \infty]$ is the fuzzy coefficient. In the fuzzy K-prototypes iterative method, membership x_{kj} is calculated in the following ways:

$$\underset{\substack{1 \leq k \leq z \\ 1 \leq i \leq n}}{\forall} x_{ki}\left(y_k, u_i\right) = \begin{cases} x_{ki} = 1, y_k = u_i \\ x_{ki} = 0, y_h = u_i, h \neq k \\ x_{ki} = \sum_{h=1}^{z}\left[\frac{s(y_k, u_i)}{s(y_h, u_i)}\right]^{-\frac{1}{(\partial - 1)}} \\ y_h \neq u_i, 1 \leq h \leq z \end{cases} \tag{2.4}$$

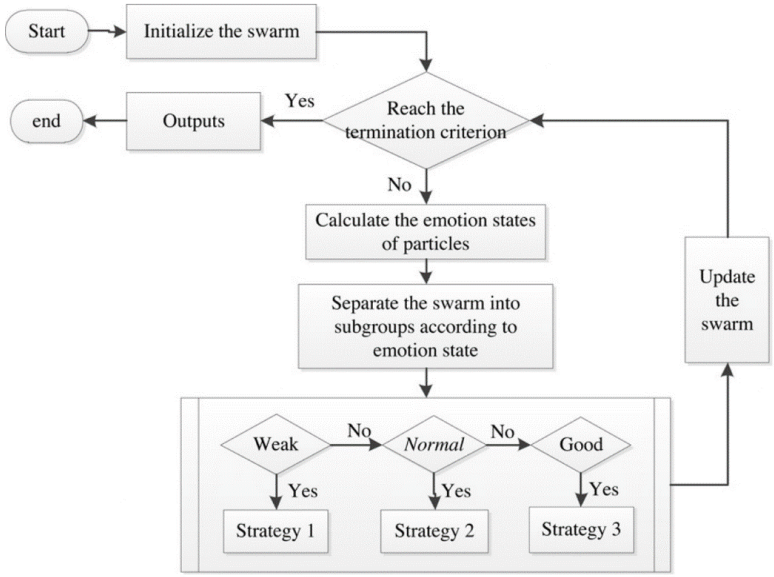


Fig. 3.1: *Flow of particle swarm optimization algorithm.*

In the iteration, $j(1 \leq j \leq q)$ adjacent properties E_{kj} of the cluster centroid y_k are computed by the following method:

$$E_{kj} = \frac{\sum_{i=1}^{n} (x_{ki})^{\partial} u_{ij}}{\sum_{i=1}^{n} (x_{ki})^{\partial}} \tag{2.5}$$

For feature $E_{kj} = e_j^{(c)} \in DOM(E_j)$ of category $j(q+1 \leq j \leq m), c$ must meet the following conditions.

$$\sum_{i=1}^{n} \left((x_{ki})^{\partial} \mid u_{ij} = e_j^{(c)} \geq \sum_{i=1}^{n} \left((x_{ki})^{\partial} \mid u_{ij} = e_j^{(t)} \right), 1 \leq c, t \leq n_j \right. \tag{2.6}$$

n_j is the number of numeric values of class property E_j.

3. Particle swarm optimization based on information security..

3.1. PSO algorithm. This project uses a binary particle swarm optimization algorithm to classify them. The group of particles is divided into N particles, and each particle is an active point in the discrete D-dimensional space [7]. At this time, the velocity of particle i at time t is $y_i(t), u_i(t), q_i(t)$, and the optimal position of the individual is $y_i(t)$, then the iterative formula of particle i velocity and position is as follows: Figure 3.1 shows the flow of the particle swarm optimization algorithm.

$$\begin{aligned} y_{is}(t+1) &= \varphi \cdot y_{is}(t) + z_1 c_1 \left(q_{is}(t) - u_{is}(t) \right) + z_2 c_2 \left(y_s(t) - u_{is}(t) \right) \\ u_{is}(t+1) &= \begin{cases} 1, & c_3 < \mathrm{Sig}\left(y_{is}(t+1) \right) \\ 0, & c_3 \geq \mathrm{Sig}\left(y_{is}(t+1) \right) \end{cases} \end{aligned} \tag{3.1}$$

$i = 1, \cdots, N$ (N is the size of the population, usually 20). $s = 1, \cdots, S$ (S stands for the dimensions of each component of the particle code given according to a particular problem). c_1, c_2, c_3 is any number in the interval

$(0,1)$. Where z_1 and z_2 are learning factors; It's generally thought that $z_1 = z_2 = 2; \varphi$ is not negative [8]. In this paper, it is called "inertia coefficient" or "inertia weighting". In the binary particle swarm optimization algorithm, $\varphi = 1$ is generally taken. $\mathrm{Sig}(\cdot)$ is the standard symbol, it's usually called $\mathrm{Sig}(u) = 1/(1+\exp(-u))$. 3.2 Fitness function. Label the category in the sample collection as $X = \{u_i \mid i = 1, 2, \cdots, N\}$. Where u_i is a vector in dimension S, then the class internalization of Xf_u is:

$$f_u = \sum_{i=1}^{N} \sum_{j=i}^{N} \|u_i, u_j\| \tag{3.2}$$

Cluster center X_z of X meets the following conditions:

$$X_z = \frac{1}{N} \sum_{i=1}^{N} u_i \tag{3.3}$$

Suppose the other categories of the sample collection represent $Y = \{y_i \mid i = 1, 2, \cdots, M\}$. Where y_i is a vector of S dimension, then the Euclidean distance between groups X and Y is their group interval $s(X, Y)$. The method synthesized the intra-group distance and interval, the intra-group aggregation degree and the inter-group dispersion degree and defined the adaptability value f. as the formula (3.5). The value of f' decreases with the decrease of the intra-group distance and the increase of the group distance [9]. The algorithm's convergence can be achieved by selecting the minimum fitness as the operating criterion.

$$f' = \frac{f_X + f_Y}{s(X, Y)} \tag{3.4}$$

4. Simulation experiment and result analysis.

4.1. Experimental data and parameter Settings. The experimental data used in this paper comes from the KDD99 intrusion detection system, which is highly similar to the real world and is widely used in intrusion detection systems. Each record has 42 characteristics, and an expert identifies the previous item as a regular link or an attack. Of the remaining forty-one attributes, nine are numbers, and thirty-two are numbers. Because many of the 41 experimental properties are useless, their appearance worsens the clustering result and the operation speed faster. Therefore, it is necessary to eliminate them when clustering them. This paper uses the attribute selection method proposed in the literature [10] to select 14 attributes, including four symbols and ten values. Because the data size of the complete knowledge discovery database is enormous, the paper selects some samples from the knowledge discovery database to carry out the algorithm test. In 10% of the training database, 55,555 samples were used as test data to ensure that the selected data matched the recorded data of the actual network [11]. There are 50,000 formally linked data; Of these, 5000 Dos attacks, 500 U2R attacks, 50 R2L attacks, and 5 Probe attacks. All test data are extracted from the 10% training library under the requirement of sampling quantity for each classification. First, the data should be normalized to prevent the impact of different dimensions on the data.

$$X_i = \frac{X_i - X_{\min}}{X_{\max} - X_{\min}} \tag{4.1}$$

X_i and X_i are the initial values and normalized values of each feature, respectively. $X_{\min}$ and $X_{\max}$ represent the maximum and minimum of this feature. The clustering results of this method are usually measured by three criteria: class target value, intra-cluster tightness and degree of separation between clusters. When the value of the cluster objective function is lower, the clustering results in the cluster are better [12]. With the decrease of the spacing between clusters, the clustering efficiency of clusters increases. As the distance between clusters increases, the clustering efficiency of clusters also increases. The number of particles in PSO-KP and the number of chromosomes in GA-KP were set to 20, and other regulation and control indicators were set in a general way. In the PSO-KP method, it is improved by taking $w_1 = 0.9, w_2 = 0.2$ as the initial value. $c_1 = c_2 = 2.0$ is used to express the acceleration coefficient. The maximum speed limit is denoted by $V_{\max} = 0.5$. The interaction possibility of the GA-KP method is expressed by $Pc = 0.6$, and the mutation possibility is expressed by $Pm = 0.05$.

Table 4.1: *Comparison of clustering results of various algorithms.*

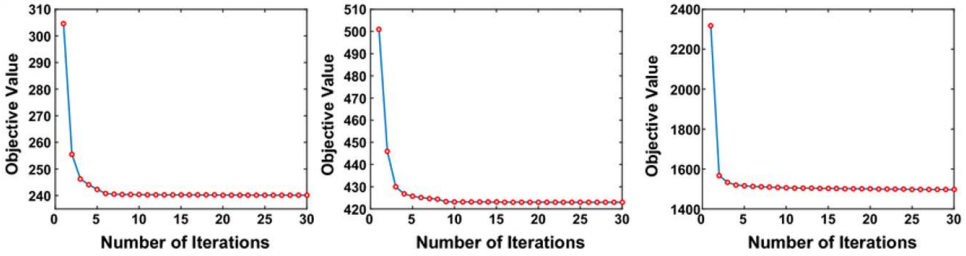
Clustering algorithm	Evaluation criterion 1: Objective function value			
	Mean value	Standard variance	Minimum value	Maximum value
K-prototypes	0.7290	0.0824	0.6031	0.8965
GA-KP	0.5675	0.0221	0.5272	0.6168
PSO-KP	0.5225	0.0179	0.4926	0.5628
Clustering algorithm	Evaluation criterion 2: In-class distance			
	Mean value	Standard variance	Minimum value	Maximum value
K-prototypes	40488	5929	33486	50917
GA-KP	39696	1529	34623	41753
PSO-KP	40042	1401	37632	42252
Clustering algorithm	Evaluation criterion 3: Distance between classes			
	Mean value	Standard variance	Minimum value	Maximum value
K-prototypes	16.3921	2.2467	9.4027	19.4889
GA-KP	17.0264	1.3935	13.5611	20.1120
PSO-KP	20.0553	0.5652	17.9706	21.0883

4.2. Result Analysis. In the test, three algorithms for K-prototypes, GA-KP and PSO-KP, were executed 30 times and 100 times for each cycle, respectively. Table 4.1 shows the objective function values, in-class distance, standard deviation, and minimum and maximum values of the results of 30 operations.

It can be seen from Table 4.1 that the mean value and standard deviation of functions obtained by the K-prototypes algorithm were the worst. The GA-KP method gives the second-best value. The PSO-KP method is the best [13]. There is little difference between the three methods in the distance between classes. K-prototypes have the advantages of high mean value and low intra-class cohesion. GA-KP and PSO-KP methods have lower mean and mean square error and higher degrees of intra-class aggregation. GA-KP is slightly better than the K-prototypes method regarding distances within clusters, but it is better than GA-KP. Therefore, the PSO-KP method can ensure small coupling between clusters. In addition, the K-prototypes algorithm may also have local extreme values, and PSO can solve the problem of K-prototype's local extreme values in a certain sense, but it can better prevent falling into local extreme values. Compared with the K-prototypes strategy PSO-KP, which organically integrates the advantages of particle swarm optimization with K-prototypes, the particle swarm optimization has a robust global search performance, thus avoiding the defect of K-prototypes function quickly falling into local optimization [14]. For the generation of the next generation population, the idea of K-prototypes is adopted for genetic optimization of the particles, combined with K-prototypes' powerful global optimization ability and fast convergence characteristics to accelerate the search speed. Figure 4.1 shows the convergence rate of each method's optimal clustering objective function in 30 cycles. The convergence of the K-prototypes algorithm is the best, but its convergence function is too large. It is easy to fall into local extreme values [15]. The GA-KP method has a lower convergence rate, but its final function is smaller than the K-prototypes function. Compared with the K-prototypes method, the convergence rate of the PSO-KP method is slightly inferior to the K-prototypes function, but its convergence rate is higher, it can reach the optimal adaptation point, and its stability is better than that of standard genetic algorithms.

Because the selected test samples have specific classification and recognition, the cluster analysis method can be used to classify them. Table 4.2 shows the accuracy rates of the clustering results of the three algorithms. The accuracy of the method is only slightly higher than that of the K-prototypes method, while that of the PSO-KP method is higher than that of the other two methods, and the accuracy rate for each test is [83.24%, 88.96%], almost unchanged. The experimental results also prove that the PSO-KP method improves the clustering effect and stability.

5. Conclusion. A new hybrid clustering algorithm for prototypes was constructed by combining PSO with K-prototype functions. The advantages of the PSO and K-prototypes algorithm were organically integrated. This method can effectively solve the problems existing in the K protection method. The selection of the original cluster is too particular, and the local extreme value can quickly occur to improve its clustering ability



(a) K-prototypes (b) GA-KP (c) PSO-KP

Fig. 4.1: *Convergence curves of optimal clustering objective functions for various algorithms.*

Table 4.2: *Clustering accuracy rate.*

Clustering algorithm	Accuracy rate of clustering results		
	Mean value	Minimum value	Maximum value
K-prototypes	51.90%	39.88%	71.45%
GA-KP	60.17%	50.49%	74.40%
PSO-KP	89.16%	86.40%	91.91%

and stability. Simulation results show the effectiveness of this method.

6. Acknowledgements. The work was supported by the Reform of the Curriculum System of University Computer based on WPS Advanced Office System (Industry-University Co-operation Project of the Ministry of Education) (No. 220801701293406).

REFERENCES

[1] Pu, G., Wang, L., Shen, J., & Dong, F. (2020). A hybrid unsupervised clustering-based anomaly detection method. Tsinghua Science and Technology, 26(2), 146-153.
[2] Lv, Z., Chen, D., Lou, R., & Song, H. (2020). Industrial security solution for virtual reality. IEEE Internet of Things Journal, 8(8), 6273-6281.
[3] Subburayalu, G., Duraivelu, H., Raveendran, A. P., Arunachalam, R., Kongara, D., & Thangavel, C. (2023). Cluster based malicious node detection system for mobile ad-hoc network using ANFIS classifier. Journal of Applied Security Research, 18(3), 402-420.
[4] Benzaïd, C., & Taleb, T. (2020). AI for beyond 5G networks: a cyber-security defense or offense enabler. IEEE network, 34(6), 140-147.
[5] Mahela, O. P., Khan, B., Alhelou, H. H., & Siano, P. (2020). Power quality assessment and event detection in distribution network with wind energy penetration using stockwell transform and fuzzy clustering. IEEE Transactions on Industrial Informatics, 16(11), 6922-6932.
[6] Zhao, Z., Qi, H., Qi, Y., Zhang, K., Zhai, Y., & Zhao, W. (2020). Detection method based on automatic visual shape clustering for pin-missing defect in transmission lines. IEEE Transactions on Instrumentation and Measurement, 69(9), 6080-6091.
[7] Zhong, W., Yu, N., & Ai, C. (2020). Applying big data based deep learning system to intrusion detection. Big Data Mining and Analytics, 3(3), 181-195.
[8] Jia, B., Zhang, X., Liu, J., Zhang, Y., Huang, K., & Liang, Y. (2021). Blockchain-enabled federated learning data protection aggregation scheme with differential privacy and homomorphic encryption in IIoT. IEEE Transactions on Industrial Informatics, 18(6), 4049-4058.
[9] Asif, M., Abbas, S., Khan, M. A., Fatima, A., Khan, M. A., & Lee, S. W. (2022). MapReduce based intelligent model for intrusion detection using machine learning technique. Journal of King Saud University-Computer and Information Sciences, 34(10), 9723-9731.
[10] Zhang, H., Li, Y., Lv, Z., Sangaiah, A. K., & Huang, T. (2020). A real-time and ubiquitous network attack detection based on deep belief network and support vector machine. IEEE/CAA Journal of Automatica Sinica, 7(3), 790-799.
[11] Hazman, C., Guezzaz, A., Benkirane, S., & Azrour, M. (2023). lIDS-SIoEL: intrusion detection framework for IoT-based smart environments security using ensemble learning. Cluster Computing, 26(6), 4069-4083.

[12] Wen, J., Yang, J., Jiang, B., Song, H., & Wang, H. (2020). Big data driven marine environment information forecasting: a time series prediction network. IEEE Transactions on Fuzzy Systems, 29(1), 4-18.
[13] Aamir, M., & Zaidi, S. M. A. (2021). Clustering based semi-supervised machine learning for DDoS attack classification. Journal of King Saud University-Computer and Information Sciences, 33(4), 436-446.
[14] Zhang, P., Liu, X., Xiong, J., Zhou, S., Zhao, W., Zhu, E., & Cai, Z. (2020). Consensus one-step multi-view subspace clustering. IEEE Transactions on Knowledge and Data Engineering, 34(10), 4676-4689.
[15] Baraneetharan, D. E. (2020). Role of machine learning algorithms intrusion detection in WSNs: a survey. Journal of Information Technology and Digital World, 2(3), 161-173.

Edited by: Hailong Li
Special issue on: Deep Learning in Healthcare
Received: Jan 29, 2024
Accepted: Mar 18, 2024