Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/100132

An Effective Network Security Log Mining Algorithm based on Fuzzy Clustering

*Peng Wang, Xikun Ma and Jingjie Yu**

Nanjing General Hospital of Nanjing Military Region, Jiangsu, Nanjing, 210000, China

Received: 2 Jul. 2015, Revised: 29 Aug. 2015, Accepted: 1 Sep. 2015
Published online: 1 Jan. 2016

Abstract: In this paper, we concentrate on the network security log mining problem, and proposed a novel fuzzy clustering algorithm to solve it. The architecture of network security log mining system is discussed at first, and three main modules are included in this system, such as data pre-processing, pattern mining and pattern analyzing. The main work of network security log mining is to find the frequent attack sequences from log files, several properties related to network security are considered in this paper, that is, start time of attacking, attacking type, end time of attacking, source IP of attackers, route path of attacking, attackers' target IP, attackers' port number, network protocol, and so on. To solve the problems in the traditional methods, we proposed a new modified fuzzy clustering algorithm by introducing the concept of interval weights, and attribute weights are denoted as interval constrained variables in our proposed algorithm. Afterwards, the final clustering results are obtained by minimizing an objective function through collaboratively optimizing the attribute weight. To testify the effectiveness of the proposed algorithm, we conduct experiments on a collection of network access log files under the real environment. Experimental results demonstrate that the proposed algorithm can achieve high quality clustering results with high time efficiency.

Keywords: Network security log, Data mining, Fuzzy clustering, Attribute weight, Membership degree

1 Introduction

In recent years, the large-scale information on the Internet increase with the development of web technologies. People aim to find useful information from the massive data sources. All the above problems generate a new research topic, which is named Web mining [1,2]. Particularly, currently Web log mining is of great importance in Web mining research [3]. Web usage mining refers to the process of utilizing the traditional data mining technology to search interesting information on Web log files [4,5]. Web log data are quite different from the conventional log data, hence, it brings many difficulties in Web log mining.

As an important branch of Web log mining [6], network security log mining has attracted more and more attentions. With the rapid development of modern network technology, network security is of great importance in network information management [7]. Although network technology provides conveniences for people's daily life, there are more and more network security threats at the same time. Currently, there are many types of network threats, such as 1) Computer virus [8], 2) Web Spam information [9], 3) Hacker attacking [10], and 4) Harmful information propagating [11] and so on. Managers of network security should solve the large-scale network security logs, that are belonged to a kind of information to save different network behaviors. In particular, the data dimension of network security log feature is quite high, and there are a large-scale network security logs to be solved. Hence, how to effectively mine the network security log becomes an important topic for researchers.

To mine the network security log effectively, fuzzy clustering technology is utilized in this paper. As is well known that Clustering refers to divide a dataset into several groups with the rule that similar samples are belonged to a cluster, at the same time, dissimilar samples are allocated to different categories. Particularly, fuzzy clustering creatively integrates the concept of membership with data partition process. Considering membership can indicate the degree to which an object belongs to the clusters definitely, fuzzy clustering can greatly enhance the quality of data grouping [12,13].

* Corresponding author e-mail: Wp116301@sina.com

This paper illustrates a novel network security log mining algorithm based on fuzzy clustering. The rest of the paper is structured as follows. Section 2 illustrates the related works of this paper. In section 3, overview of the network security log mining system is given. Section 4 proposes the fuzzy clustering based network security log mining algorithm. To demonstrate the effectiveness of the proposed algorithm, experiments are conducted in section 5. Finally, the conclusions are drawn in section 6.

2 Related works

In this section, we provide the related works about network security log mining algorithm based on fuzzy clustering, which has been a hot topic in recent years. Currently, this research field will attract more and more attentions. Firstly, some works about data mining for log files are listed as follows.

Wang et al. considered the relevance of URLs for a query, and then utilized the Open Directory Project categories to disambiguate queries and URLs. Particularly, the authors exploited various features and clustering algorithms for intent clustering, and then detected critical actions from each intent cluster to form a search script. Afterwards, a nature language description is generated for each action, and then a topic for each search script is summarized [14].

Zamora et al. concentrated on the problem of automatic detection of query intent in Web search engines. Firstly, they studied on features which are powerful in the former researches, and then some features are extracted from the click-through data. Furthermore, four text-based classifiers are proposed to testify the effectiveness of the above text-based features. The proposed classifiers can effectively detect query intent in more than 90% of the evaluation samples [15].

Wu et al. propose a new method which can utilize query log in MR to solve the drawbacks. Particularly, the correlation between each pair of database images is obtained from the query log, and then the affinity matrix of semantic structure can be modified. Furthermore, the relevance score of each given image to the specific user's query can be achieved from the query log [16].

Khairudin et al. aimed to study on the effect of temporal attribute in relational rule mining for Web log data. Particularly, the authors integrated the features of time in the rule mining process and then studied on the influence of different temporal parameters. The main innovations of this paper lie in that the rules obtained from temporal relational rule mining are compared with rules which are got from traditional methods [17].

Guerbas et al. presented a refined time-out based heuristic for session identification, and designed a specific density based algorithm for navigational pattern detecting. Based on the above works, a novel approach of online prediction is given [18].

Francisco et al. analyzed results on query contextualization through the association of tags to queries, which is also named query folksonomies. Particularly, results of this paper highly depend on the analysis of large query log induced graphs, that is also means click induced graphs [19].

As the anonymization of query logs is of great importance, Navarro et al. proposed the anonymization of query logs exploiting micro-aggregation. Furthermore, this method can ensure the k-anonymity of the users in the query log [20].

To promote the quality of network security log mining, in this paper, we utilize the fuzzy clustering technology. In fuzzy clustering, the fuzzy logic is combined with the classification theory to construct the definition of the fuzzy partition, in which each input vector can be belonged to more than one classes with different membership scores. As far as we know, fuzzy clustering has been widely utilized in intelligent computing as follows.

Sanchez et al. proposed a novel approach to seek fuzzy information granules from multivariate data through a gravitational inspired clustering algorithm. In this paper, the authors incorporated the theory of granular computing to the context of the given data. Furthermore, the Fuzzy Granular Gravitational Clustering Algorithm is compared with classification accuracy and clustering validity indices [21].

In paper [22], the authors presented a fuzzy clustering method to predict situations in complex process industries. The proposed algorithm integrates a static measurement with historical process data, and then a modified estimation algorithm using Markov's theory is designed as well.

Szilagyi et al. proposed some generalized formulations of the suppression rule, and then used them to an infinite number of novel clustering algorithms. The authors identified the close relation between s-FCM clustering models with generalized improved partition. Meanwhile, the constraints under which the generalized s-FCM clustering models minimize the objective function of GIFP-FCM are provided [23].

Singh et al. analyzed and compared the lifetime of the network with three different fuzzy-based methods, and three important parameters, e.g. energy centrality and node density are utilized for cluster head choosing. Particularly, node density and centrality are used through a fuzzy system to choose cluster heads [24].

Different from the above works, in this paper, we introduce the attribute weight into the fuzzy clustering algorithm, and utilize it in the network security log mining problem. Particularly, this idea of network security log mining has not been tried yet.

3 Overview of the network security log mining system

To describe the network security log mining algorithm, the system architecture should be illustrated in advance, because the system architecture is the supporting platform for the network security log mining algorithm. Architecture of the World Wide Web is given in Fig.1.

As is shown in Fig.1, the user requirements and the server response can be solved by this system, and the structure of WWW utilizes the client/server model. Particularly, WWW structure is independent from platform and servers are transparent to Web site users. In Fig.1, CP, CW and PW represent communications between clients and proxy servers, communications between clients and Web servers, proxy servers and Web servers respectively. Afterwards, we will describe the basic structure of Web log files (shown in Table.1) The

Table 1: Basic structure of Web log files

Field name	Description
Date	Date of user requesting pages
Time	Time of user requesting pages
C_IP	IP address of client host
C_N	User name of client host
S_N	Server name
S_IP	IP address of Server
S_Port	Server port
Cs_method	Request methods of users
URL_S	User requesting pages
URL_Q	User queries
P-stat	Status Identifier returned by HTTP
C_host	Host OS
User_agent	Service provider

main task of network security log mining is to seek the frequent attack sequences from log files. Supposing that $DB = \{X_1, X_2, \cdots, X_m\}$ refer to a set of attack sequences which is memorized in the database of network log files. Therefore, each element S_i, $i \in [1,m]$ contains several properties, for example, 1) Start time of attacking, 2) Attacking type, 3) End time of attacking, 4) Source IP of attackers, 5) Route path of the attacking behaviors, 6) Target IP of attackers, 7) Port number of attackers, 8) Protocol of used in the attacking behaviors, and so on.

Considering the physical topology and user access model are different for different websites, and it is difficult to make sure users, session, transactions from log files. Thus, the flow chart of network security log mining is given in Fig.2, in which three main modules are included, such as 1) data pre-processing, 2) pattern mining and 3) pattern analyzing.

As is shown in Fig.2, to enhance the effectiveness of data mining process, some pre-processing works should be done in advance, including: 1) data cleaning, 2) user identification, 3) session identification, and 4) path supplement. After executing the data pre-processing, pattern mining can be done by several technologies, such as clustering, path analyzing, association rules, sequence pattern, and classification. Afterwards, the pattern mining results can be represented and analyzed by query mechanism, OLAP, and visualization technology.

In a word, the proposed network security system is mainly constructed by four steps, that is 1) Collecting log files, 2) Data pre-processing, 3) Obtaining frequent attack sequences, and 4) Clustering frequent attack sequences. For the third step, Apriori algorithm [25] is utilized to find the frequent itemset form a set of log files. The main task of this paper lies that we propose a novel fuzzy clustering algorithm to cluster the frequent attack sequences to obtain the network security log mining results.

4 Fuzzy clustering based network security log mining algorithm

In this paper, we introduce the fuzzy clustering technology to the network security log mining problem. Fuzzy c-means clustering (FCM) is commonly used in intelligent computation. Assuming that $X = \{x_1, x_2, \cdots, x_n\} \subset R^S$ represents a dataset with n points, the Fuzzy c-means clustering algorithm aims to divide the dataset X to several clusters. Particularly, the objective function of Fuzzy c-means clustering is illustrated as follows.

$$J_m = \sum_{k=1}^{c} \sum_{i=1}^{n} u_{ki}^m \|x_i - v_k\|^2 \tag{1}$$

Subject to:

$$\sum_{k=1}^{c} u_{ki} = 1,\ 0 \le u_{ki} \le 1,\ 0 \le \sum_{i=1}^{n} u_{ki} \le n \tag{2}$$

In equation 1, $\|\cdot\|$ represents the Euclidean norm, and the parameter m means the weighting exponent of the membership function which can compute the number of fuzziness of the clusters. Furthermore, v_k represents the centroids of the k^{th} cluster, and the symbol u_{ki} is the value of the membership degree of the element x_i belonged to the k^{th} cluster. To minimize the value of J_m, the Lagrange multiplier approach is utilized. Thus, u_{ki} and v_k can be updated as follows.

$$u_{ki} = 1 \Big/ \sum_{l=1}^{c} \left(\frac{\|x_i - v_k\|^2}{\|x_i - v_l\|^2} \right)^{\frac{1}{m-1}} \tag{3}$$

$$v_k = \frac{\sum_{i=1}^{n} u_{ki}^m \cdot x_i}{\sum_{i=1}^{n} u_{ki}^m} \tag{4}$$

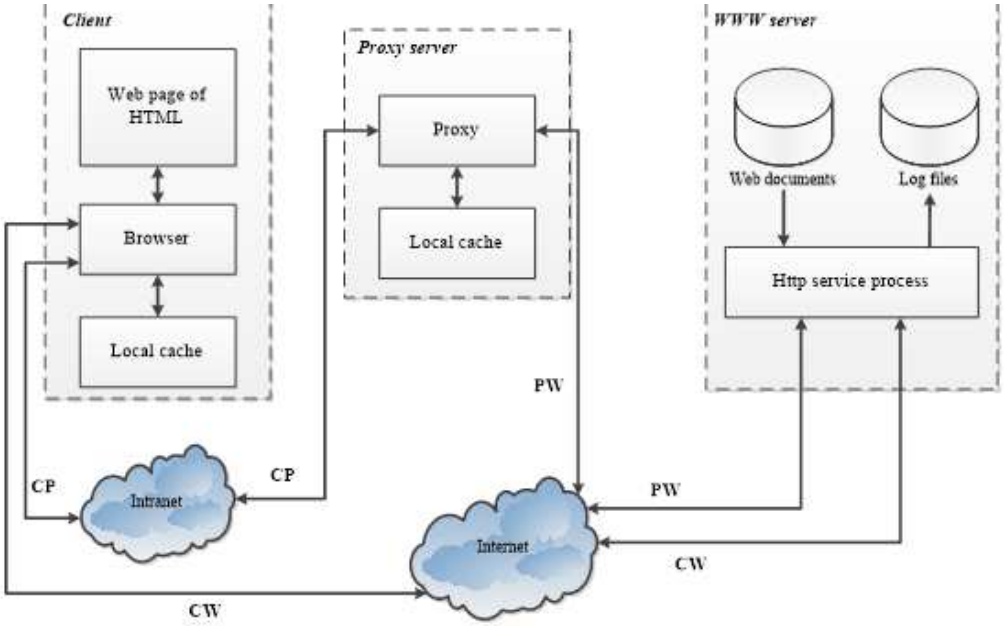

Fig. 1: Architecture of the World Wide Web

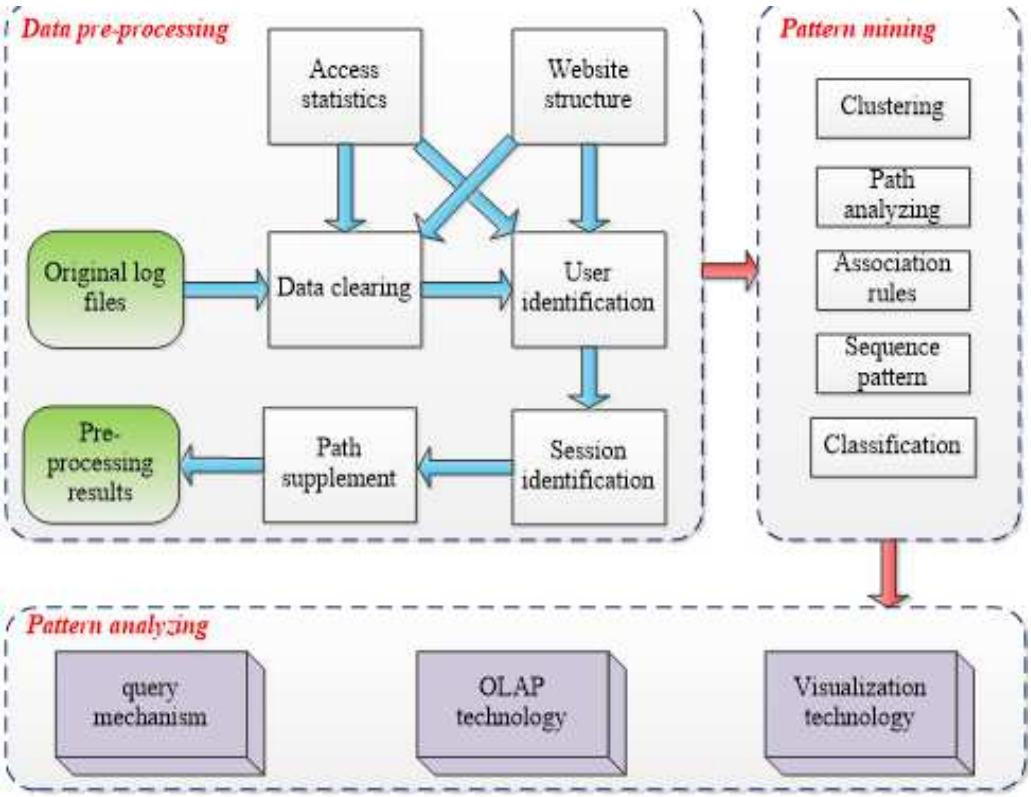

Fig. 2: Flow chart of network security log mining

However, the convergence speed of fuzzy c-means clustering is not satisfied. Hence, in this paper, we proposed a modified fuzzy c-means clustering exploiting interval weights, in which attribute weights are regarded as interval constrained variables. Our modified fuzzy c-means clustering algorithm is illustrated as follows.

Supposing that there is a set of samples $\{x_1, x_2, \cdots, x_n\} \subset \Re^s$, and the attribute weight vector is represented as $W = [w_1, w_2, \cdots, w_s]^T$. At the same time, two conditions should be satisfied, that is, $\forall j,\ wj \in [\alpha_j, \beta_j]$, $0 \le \alpha_j \le 1, 0 \le \beta_j \le 1$ and $\sum_{j=1}^{s} w_j = 1$. Based on the above constraints, the fuzzy c-means clustering can be solved by the minimizing the following function:

$$J(U,V,W) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m} \|X_k - V_i\|_W^2 \tag{5}$$

If the there is no w_j is limited to interval, the Lagrange multiplier could be exploited to solve the above problem, and the Lagrange function is defined as follows.

$$\overline{J}(U,V,W) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m} \|X_k - V_i\|_W^2 + \sum_{k=1}^{n} \theta_k \cdot \left(\sum_{i=1}^{c} u_{ik} - 1\right) + \upsilon\left(\sum_{j=1}^{z} w_j - 1\right) \tag{6}$$

In Eq.6, $\theta = [\theta_1, \theta_2, \cdots, \theta_n]^T$ and υ denotes the Lagrange multipliers. Afterwards, the modified fuzzy clustering algorithm is described as the following steps:

(1) Computing the interval endpoints α_j and β_j for the weight of attribute w_j

(2) Supposing the size of genetic population is represented as S, the max number of generations is G, the crossover probability is CP, and the mutation probability is MP. Thus, the genetic population GP^1 can be computed as follows.

$$GP_1^1 = \left[\frac{\alpha_1 + \beta_1}{2}, \frac{\alpha_2 + \beta_2}{2}, \cdots, \frac{\alpha_q + \beta_q}{2}\right] \tag{7}$$

$$GP_h^1 = [rand(\alpha_1, \beta_1), rand(\alpha_2, \beta_2), \cdots, rand(\alpha_q, \beta_q)] \tag{8}$$

(3) If the genetic generation index is equal to l, $l \in \{1, 2, \cdots, G\}$, then for each chromosome GP_p^t, the value of $V_p^{(l)}$ and $U_p^{(l)}$ can be obtained by a alternating optimization process.

(4) Computing fitness value of each GP_p^t by the following equation:

$$Fit\left(GP_p^t\right) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m} \|Xk - Vi\|_w^2 + R \cdot \left|\sum_{j=1}^{s} w_{pj} - 1\right| \tag{9}$$

Where R refers to the penalty gain. After obtaining the fitness value, we rank all the individuals in ascending order according to its fitness value.

(5) Executing the roulette wheel selecting using the selecting probability, executing the whole arithmetic crossover using the crossover probability CP, and executing uniform mutation by the mutation probability MP.

(6) If the index of genetic generation l is equal to G, the final attribute weights and the final clustering results can be obtained, and the algorithm is ended.

(7) Otherwise, let $l = l + 1$ and go to step three.

5 Experiment

In order to testify the performance of our proposed algorithm, a series of experiments are conducted to make performance evaluation. The network security log files utilized in this experiment are collected from Dell PowerEdge 2900 System, and the processors used in this system is Intel Xeon with the frequency 2.00GHz. The Dell PowerEdge 2900 system is designed to provide high performance in a tower chassis or a 5U rackable option with next generation dual core Intel Xeonprocessors. Moreover, this system can support 12 memory slots for the memory with 48GB for memory-intense workloads and other applications as well.

Particularly, one year network access log files are collected to make the dataset, in which 935 users and 13829 sessions are included. In the data cleaning process, the sessions with the assess number smaller than 0.0005 or higher than 0.9 are pruned from the dataset. Furthermore, the support of the Apriori algorithm is set to be 0.05, and the slipping windows depth is set to 2 in this algorithm. Details of the dataset used in this experiment is listed in Table.2

To make performance comparison, three typical clustering methods are chosen: 1) hard c-means (HCM) [26], 2) fuzzy c-means (FCM) [27], and 3) suppressed fuzzy c-means clustering (s-FCM)[28]. Hard c-means clustering is belonged to unsupervised classification approaches which classify a set of input vectors into a pre-defined groups. FCM utilizes a probabilistic constraint to define the fuzzy membership functions, and it has been widely used. s-FCM is designed to make a step from FCM method towards HCM algorithm, through operating with the fuzzy membership functions calculated in each iteration of the fuzzy c-means clustering's alternating optimization approach.

Before conducting the experiment, the evaluation metrics used in this paper is illustrated in advance. To evaluation the intra-cluster homogeneity and the inter-cluster separation of the final clustering results. Overall cluster quality (OCQ) is utilized. The definition of OCQ is defined as follows.

$$OCQ(\alpha) = \alpha \cdot CMP + (1 - \alpha) \cdot SEP \tag{10}$$

Table 2: Details of the dataset

Name of attribute	Number
Total number of access entries	1,784,863
Clean assess entries	245,976
Users	17,357
Total number of accessed Web pages	1037
Number of Web pages assessed more than ten times	875
Total identified sessions	31,745
Number of identified sessions which are requested than two times	16,314

Where CMP and SEP refer to the cluster compactness and cluster separation respectively, and the parameter α $(0 \leq \alpha \leq 1)$ denotes the weight which balance the importance of CMP and SEP. CMP and SEP are defined as follows.

$$CMP = \frac{1}{NC} \cdot \sum_{i=1}^{NC} \frac{v(c_i)}{v(D)} \tag{11}$$

Where NC refers to the number of clusters extracted from the dataset D. $v(c_i)$ and $v(D)$ means the deviation of the cluster c_i and the deviation of dataset D respectively.

$$SEP = \frac{1}{NC \cdot (NC-1)} \cdot \sum_{i=1}^{NC} \sum_{i \neq j, j=1}^{NC} \exp\left(-\frac{d^2(x_{ci}, x_{cj})}{2\sigma^2}\right) \tag{12}$$

Afterwards, to evaluate the quality of clustering results, Precision, Coverage, F-measure, and Sensitivity are utilized as well. Supposing that k represents a set of frequent attack sequence, p means the depth of the sliding window, and T denotes the clustering results. Precision (denoted as P), Coverage(denoted as C), F-measure (denoted as F) and Sensitivity (denoted as S) are defined as follows.

$$P(T,k) = \frac{|T \cap (k-p)|}{|T|} \tag{13}$$

$$C(T,k) = \frac{|T \cap (k-p)|}{|k-p|} \tag{14}$$

$$F(T,k) = \frac{2 \times P \times C}{P+C} \tag{15}$$

$$S = \frac{TP}{TP+FN} \tag{16}$$

Where TP and TN refer to the number of true positive samples and true negative samples respectively. We found that , in this experiment, if the number of clusters is larger than 24, some groups of the clustering results may include sessions less than 3% of the total sessions. Furthermore, we discover that the navigation patterns which these clusters present are not representative patterns in the total sessions. Therefore, we set the number of clusters in the range of 2 to 24. In the following parts, experimental results using the above performance evaluation metrics are given in Fig.3-Fig.7.

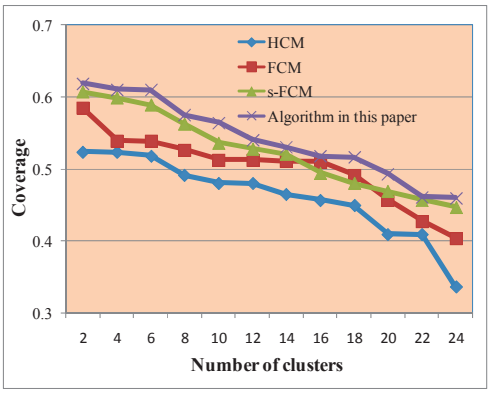

Fig. 3: Comparison of coverage for different methods

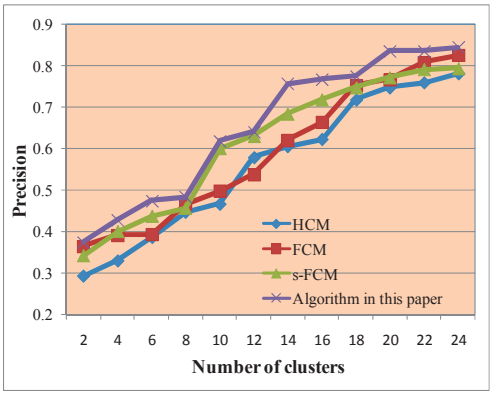

Fig. 4: Comparison of precision for different methods

From Fig.3 to Fig.7, it is clear that our proposed fuzzy clustering algorithm can detect and obtain the clusters from network security log, and performs better than other approaches. Moreover, the proposed algorithm can keep the effectiveness when the number of clusters increasing.

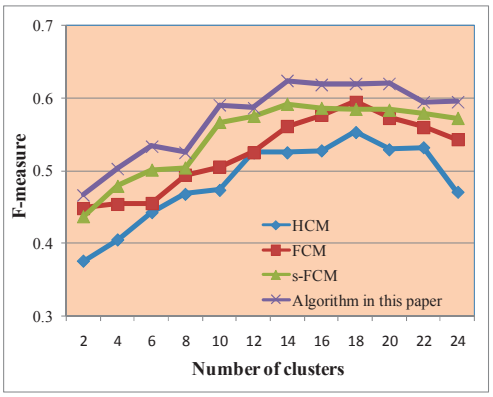

Fig. 5: Comparison of F-measure for different methods

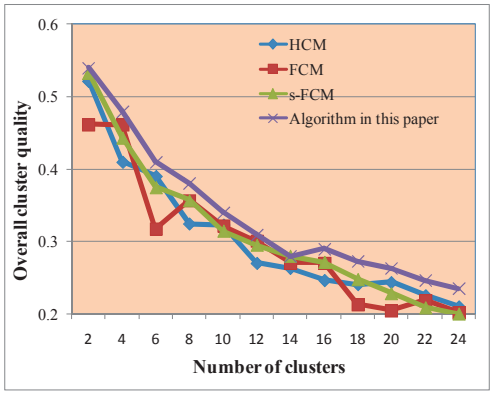

Fig. 6: Comparison of overall cluster quality for different methods

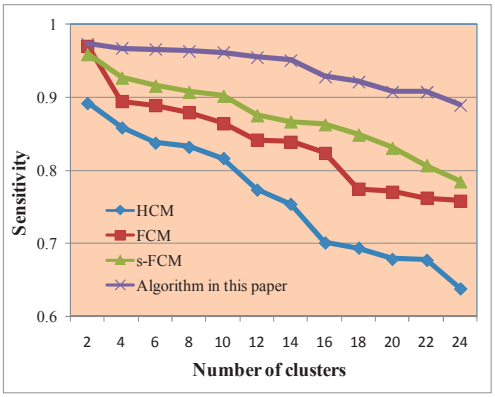

Fig. 7: Comparison of sensitivity for different methods

Furthermore, the conclusions can be drawn that our proposed algorithm is more effective than others especially in cases with high degree of complexity. In addition, the membership degrees calculated by the proposed algorithm are more suitable than HCM, FCM, and s-FCM, and our algorithm can allocate suitable membership degrees to the samples. The reasons lie in the following aspects:

(1) In this paper, the network security log mining system is made up of data pre-processing, pattern mining and pattern analyzing, in which cluster is a very important function. High quality clustering results obtained by the proposed algorithm can effectively enhance the system performance.

(2) In our proposed algorithm paper, interval number is utilized attribute weighting for the fuzzy clustering. Particularly, the interval-represented attribute weights can obviously promote the quality of attribute weight vector.

(3) HCM exploits the bivalent logic to describe partitions, and its converging speed is quite fast. However, HCM is sensitive to initialization, and it may frequently reach stuck in local minima when mediocre partitioning.

(4) FCM can create high quality partitions than HCM, and it has a reduced but still observable sensitivity to initial cluster prototypes. However, the converging of FCM is quite slower than HCM. Although there are some drawbacks in FCM, it is still belonged to one of the most powerful clustering methods.

(5) s-FCM makes a step from FCM towards HCM, and the main innovations of s-FCM lie in that it implement the clustering using the fuzzy membership functions which is calculated in each iteration of the alternating optimization approach. However, this method does not consider the attribute weight in vectors, and this drawback restricts the clustering quality.

Although our proposed algorithm can provide high quality clustering results, and then network security log can be effectively mined. The computation cost is another important problem, hence, the running time of each method is tested (shown in Fig.8).

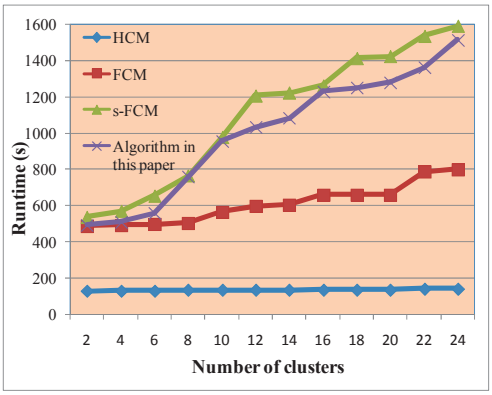

Fig. 8: Comparison of runtime for different methods

As is shown in Fig.8, we can see that the running time of our proposed is only lower than s-FCM, and is much less than HCM and FCM. However, clustering quality of HCM and FCM is not satisfied. Therefore, comprehensively considering the clustering quality and computation cost, the proposed algorithm is superior to other three methods.

6 Conclusion

We presented a novel network security log mining algorithm based on a modified fuzzy clustering method. Firstly, we give the framework network security log mining system, which is made up three parts: 1) data pre-processing, 2) pattern mining and 3) pattern analyzing. In this system, the main task of network security log mining is to seek the frequent attack sequences from log files. To tackle the problem in traditional Web log mining, we designed a modified fuzzy c-means clustering algorithm utilizing interval weights. Particularly, attribute weights are represented as interval constrained variables as well. Based on the above works, clustering results can be got though solving an objective function based on a collaboratively optimizing process. In the future studies, we will test if the proposed algorithm is suitable for the large-scale network security log mining.

References

[1] Xu Guandong, Yu Jeffrey, Lee Wookey, social networks and social Web mining, World Wide Web-internet And Web Information Systems, 2013, 16(5-6): 541-544.
[2] Jiang Daxin, Pei Jian, Li Hang, Mining Search and Browse Logs for Web Search: A Survey, ACM Transactions on Intelligent Systems and Technology, 2013, 4(4), Article No. 57.
[3] Velasquez Juan D., Web mining and privacy concerns: Some important legal issues to be consider before applying any data and information extraction technique in web-based environments, Expert Systems with Applications, 2013, 40(13): 5228-5239.
[4] Thorleuchter Dirk, Van den Poel Dirk, Weak signal identification with semantic web mining, Expert Systems with Applications, 2013, 40(12): 4978-4985.
[5] Matthews Stephen G., Gongora Mario A., Hopgood Adrian A., Web usage mining with evolutionary extraction of temporal fuzzy association rules, Knowledge-based Systems, 2013, 54: 66-72.
[6] Guerbas Abdelghani, Addam Omar, Zaarour Omar, Effective web log mining and online navigational pattern prediction, Knowledge-based Systems, 2013, 49: 50-62.
[7] Jianhua Che, Weimin Lin, Yong Yu, Wei Yao, Optimized Hypergraph Clustering-based Network Security Log Mining, Physics Procedia, 2012, 24: 762-768.
[8] Gan Chenquan, Yang Xiaofan, Liu Wanping, Propagation of computer virus both across the Internet and external computers: A complex-network approach, Communications in Nonlinear Science And Numerical Simulation, 2014, 19(8): 2785-2792.
[9] Wang De, Irani Danesh, Pu Calton, A Perspective of Evolution After Five Years: A Large-Scale Study of Web Spam Evolution, International Journal of Cooperative Information Systems, 2014, 23(2).
[10] Garber Lee, Hackers Could Attack Networked Traffic Control Equipment and Cause Gridlock, Computer, 2014, 47(6): 15-16.
[11] Javier Ortega, F., Troyano Jose A., Cruz Fermin L., Polarityspam: Propagating Content-based Information Through A Web-graph To Detect Web-spam, International Journal of Innovative Computing Information and Control, 2012, 8(4): 2915-2928.
[12] Huang Hsin-Chien, Chuang Yung-Yu, Chen Chu-Song, Multiple Kernel Fuzzy Clustering, IEEE Transactions on Fuzzy Systems, 2012, 20(1): 120-134.
[13] Linda Ondrej, Manic Milos, General Type-2 Fuzzy C-Means Algorithm for Uncertain Fuzzy Clustering, IEEE Transactions on Fuzzy Systems, 2012, 20(5): 883-897.
[14] Wang Chieh-Jen, Chen Hsin-His, Intent mining in search query logs for automatic search script generation, Knowledge and Information Systems, 2014, 39(3): 513-542.
[15] Zamora Juan, Mendoza Marcelo, Allende Hector, Query Intent Detection Based on Query Log Mining, Journal of Web Engineering, 2014, 13(1-2): 24-52.
[16] Wu Jun, Shen Hong, Xiao Zhi-Bo, Boosting Manifold Ranking for Image Retrieval by Mining Query Log Repeatedly, Journal of Internet Technology, 2014, 15(1): 135-143.
[17] Khairudin Nazli Mohd, Mustapha Aida, Ahmad Mohd Hanif, Effect of Temporal Relationships in Associative Rule Mining for Web Log Data, Scientific World Journal, 2014, Article No. 813983.
[18] Guerbas Abdelghani, Addam Omar, Zaarour Omar, Effective web log mining and online navigational pattern prediction, Knowledge-based Systems, 2013, 49: 50-62.
[19] Francisco Alexandre P., Baeza-Yates Ricardo, Oliveira Arlindo L., Mining query log graphs towards a query folksonomy, Concurrency and Computation-practice & Experience, 2012, 24(17): 2179-2192.
[20] Navarro-Arribas Guillermo, Torra Vicenc, Erola, Arnau User k-anonymity for privacy preserving data mining of query logs, Information Processing & Management, 2012, 48(3): 476-487.
[21] Sanchez Mauricio A., Castillo Oscar, Castro Juan R., Fuzzy granular gravitational clustering algorithm for multivariate data, Information Sciences, 2014, 279: 498-511.
[22] Isaza Claudia V., Sarmiento, Henry O., Kempowsky-Hamon, Tatiana, Situation prediction based on fuzzy clustering for industrial complex processes, Information Sciences, 2014, 279: 785-804.
[23] Szilagyi Laszlo, Szilagyi Sandor M., Generalization rules for the suppressed fuzzy c-means clustering algorithm, NEUROCOMPUTING, 2014, 139: 298-309.
[24] Singh Ashutosh Kumar, Purohit Neetesh, An optimised fuzzy clustering for wireless sensor networks, International Journal of Electronics, 2014, 101(8): 1027-1041.
[25] Xu Rui, Wunsch Donald, Survey of clustering algorithms, IEEE Transactions on Neural Networks, 2005, 16(3):645-678.

[26] S.McQueen,Some methods for classification and analysis of multivariate observations, in:The Fifth Berkeley Symposiumon Mathematical Statistics and Probability,1967,pp.281-297.
[27] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum, New York, NY, 1981.
[28] Feng Zhao, Jiulun Fan, Hanqiang Liu, Optimal-selection-based suppressed fuzzy c-means clustering algorithm with self-tuning non local spatial information for image segmentation, Expert Systems with Applications, 2014, 41:4083-4093.

Wang Peng is currently working at Nanjing General Hospital of Nanjing Military Region.His research interests wireless network and network security.

Ma Xi-kun researches on medical informatization at Nanjing General Hospital of Nanjing Military Region.

YU JING-JIE researches on medical informatization at Nanjing General Hospital of Nanjing Military Region