View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Elsevier - Publisher Connector

SciVerse ScienceDirect

Physics Procedia 24 (2012) 1186 – 1191

Physics Procedia

2012 International Conference on Applied Physics and Industrial Engineering

# Application of Fuzzy C-Means Clustering Algorithm Based on Particle Swarm Optimization in Computer Forensics

Deguang Wang, Baochang Han, Ming Huang

*Software Technology Institute Dalian Jiaotong University*
*Dalian, China*

**Abstract**

Computer forensics is the technology of applying computer technology to access, investigate and analysis the evidence of computer crime. It mainly include the process of determine and obtain digital evidence, analyze and take data, file and submit result. And the data analysis is the key link of computer forensics. As the complexity of real data and the characteristics of fuzzy, evidence analysis has been difficult to obtain the desired results. This paper applies fuzzy c-means clustering algorithm based on particle swarm optimization (FCMP) in computer forensics, and it can be more satisfactory results.

© 2011 Published by Elsevier B.V. Selection and/or peer-review under responsibility of ICAPIE Organization Committee. Open access under CC BY-NC-ND license.

*Keywords: particle swarm optimization, computer forensics, fuzzy c-means clustering*

## 1. Introduction

With the rapid development of computer networks, almost the daily work of all trades is more and more inseparable from computer. High-tech crimes, commercial fraud and other phenomena involving computers and the Internet occur more frequently. The traditional means of criminal investigation can't solve such computer crimes, so people pay more and more attention to computer forensics. What is computer forensics? Judd Robbins defined: "Computer forensics is simply the application of computer investigation and analysis techniques in the interests of determining potential legal evidence". Hilton Chan gave a modern definition: "a scientific and systematic methodology for identifying, searching, retrieving, recovering and analyzing digital evidence from computers, computer storage media & electronic devices and presenting the findings which meets the standard required by a court of law".

The digital evidence is also called electronic evidence. The concept covers all the electronic data that relate to the evidence. It includes the data of computer system itself, such as document files, image files, graph files, audio and video files, system logs, and so on. It also includes the data from the network, like

1875-3892 © 2011 Published by Elsevier B.V. Selection and/or peer-review under responsibility of ICAPIE Organization Committee.
Open access under CC BY-NC-ND license. doi:10.1016/j.phpro.2012.02.177

firewall logs, IDS logs and switch or router memory data, and so on. Along with the development of IT technology, digital evidence not only refers to the data stored in a computer, but also includes the data that can provide help for the court case in PDA, mobile phone, iPod, digital camera, mp3 digital devices. The digital evidence is in the nature of high technology, vulnerability, concealment and multimedia

According to occurrence times of electronic evidence, computer forensics will be divided into static and dynamic forensics. Static forensics refers to the original data of computer hardware, preservation, examination and analysis, and then find out the digital evidence relevant to the case. And produces a detection report has legal effect to prove the fact that there is criminal. Dynamic forensics refers to the memory data, network activity data, system operating conditions and other relevant data of the computer is switched on or networked computers and related equipment (including switches, routers, etc.) for real-time monitoring, analysis and preservation. And produces a detection report has legal effect to prove the fact that there is criminal. Whether static or dynamic forensics, the processes include six steps what are protection of suspicious computers and related peripherals, data access, data transfer, data preservation, data analysis and data submit. The results of data analysis will directly affect the progress of the case, and also affect the reliability and the validity of the evidence data that is ultimately submitted to the court. So data analysis is one of the most critical steps in computer forensics.

It is to obtain useful information from visible and known data in traditional forensic analysis. The computer forensics analysis is to obtain useful information from the mass of the various types of electronic data. The process is not very good manual completed, so we need the help of computer systems to screen out the data relevant to the computer crime. The data mining is precisely extracting the information of people interested from the large scale data. Cluster analysis is an important aspect of the study of the data mining. It is process of division physical and abstracted collection into several classes composed of similar objects. Cluster analysis is the learning process of a free guide. It can be divided criminal acts into a number of classes or cluster. In same cluster crime acts enjoy high similarity. On the contrary, in different clusters they have large difference [1]. This article is fuzzy c-means clustering algorithm based on particle swarm optimization used in computer forensics, and effectively clusters digital evidence in order to analysis.

## 2. Particle Swarm Optimization

Particle swarm optimization (PSO) is an evolutionary computation technique of global search strategy (a search method based on a natural system) developed by Kennedy and Eberhart [2] [3]. Through the cooperation and competition of the particles in population, it generates swarm intelligence to guide the optimization search. Advantage of the unique memory function so that it can dynamically track the current search conditions to adjust search strategy.

There are $M$ particles in D-dimensional search space. $V_i = (v_{i1}, v_{i1}, \cdots, v_{iD})$ , $X_i = (x_{i1}, x_{i2}, \cdots, x_{iD})$ represent the velocity and position of the $i^{th}$ particle, respectively. The particles have memory and each particle keeps track of its previous best position ( $P_{best}$ ) and its corresponding fitness. There exist a number of $P_{best}$ for the respective particles in the swarm and the particle with greatest fitness is called the global best ( $G_{best}$ ) of the swarm. The basic concept of the PSO technique lies in accelerating each particle towards its $P_{best}$ and $G_{best}$ locations, with a random weighted acceleration at each time step [4]. Change the velocity and position of the particle according to (1) and (2) respectively.

$$v_{id}(t+1) = \omega v_{id}(t) + c_1 r_1 (p_{id} - x_{id}(t)) + c_1 r_1 (p_{gd} - x_{id}(t)) \quad (1)$$

$$x_{id}(t+1) = x_{id}(t) + v_{id}(t+1) \quad (2)$$

Where: $d \in [1, D]$ and $i \in [1, M]$. The constants $c_1$ and $c_2$ are learning factor, $r_1$ and $r_2$ are quasi-random numbers selected from a uniform distribution in $[0.0, 1.0]$. $v_{id} \in [-v_{\max}, v_{\max}]$, $v_{\max}$ is the maximum allowable velocity for the particles, $\omega$ is inertia weight, and a large inertia coefficient encourages global exploration while small one promotes local exploration [5]. With the increase of iteration usually it is made to linear decrease from 0.9 to 0.4.

## 3. Fuzzy C-Means Clustering Algorithm

Traditional cluster analysis requires every point of the data set to be assigned into a cluster precisely, so we call it hard clustering. But in fact, most things exist ambiguity in the attribute, there are no explicit boundaries among the things, and no the nature of either-or. So the theory of the fuzzy clustering is more suitable for the nature of things, and it can more objective reflect the reality. Currently, the fuzzy c-means clustering (FCM) algorithm is the most widely used.

FCM partitions set of $n$ objects $X = (x_1, x_2, \cdots, x_n)$ into $K$ fuzzy clusters with $C = (c_1, c_2, \cdots, c_k)$ cluster centers. In fuzzy matrix $U = (u_{ij})$, $u_{ij}$ is the membership degree of the $i^{th}$ object with the $j^{th}$ cluster [6] .The characters of $u_{ij}$ are as follows :

$$u_{ij} \in [0,1] \quad \forall i \in 1,2,\cdots,n; j \in 1,2,\cdots,K$$

$$\sum_{j=1}^{K} u_{ij} = 1 \qquad \forall i \in 1,2,\cdots,n$$

$$0 < \sum_{i=1}^{n} u_{ij} < n \qquad j \in 1,2,\cdots,K$$

Update $u_{ij}$ according to (3):

$$u_{ij} = \begin{cases} [\sum_{k=1}^{K} \frac{d_{ij}^{\frac{2}{b-1}}}{d_{ik}^{\frac{2}{b-1}}}]^{-1}, & d_{ik} \neq 0 \\ 0, & d_{ik} = 0 (k = j) \\ 1, & d_{ik} = 0 (k \neq 0) \end{cases} \quad (3)$$

Where: $b > 1$ is fuzziness exponent, $c_j (j \in [1, K])$ is the clustering center. $d_{ij} = \|x_i - c_j\|$. Update $c_j$ according to (4):

$$c_j = \frac{\sum_{i=1}^{n} u_{ij}^b x_i}{\sum_{i=1}^{n} u_{ij}^b} \tag{4}$$

The objective function is the equation (5):

$$J(U, C) = \sum_{i=1}^{n} \sum_{j=1}^{K} u_{ij}^b d_{ij}^2 \tag{5}$$

FCM is to minimize the objective function when $u_{ij}$ meet conditions. The $U$ obtained from the algorithm is a fuzzy partition matrix, and it corresponds to the fuzzy partition of $X$. We apply the method of maximum subjection principle to get the certainty of partition:

In the $j^{th}$ column of $U$, if $u_{ij_0} = \max_{1<j<K} (u_{ij})$, $x_i$ merges into $j_0$. It means $x_i$ has the maximum membership degree to the cluster $j_0$, so we merge it into $j_0$.

## 4. FCMP and the application in computer forensics

The nature of FCM algorithm is to apply the gradient descent method to find out optimal solution, so there is a local optimization problem. And the algorithm convergence speed is greatly influenced by the initial value, especially in the case of large number of clusters. PSO is an effective global optimization algorithm. This paper applies fuzzy c-means clustering algorithm based on particle swarm optimization (FCMP) in computer forensics, the combinations of FCM and PSO can get better clustering results.

The central theme of FCMP: In macroscopic view, we apply PSO to search cluster center with the guidance of fitness function, and get the clustering center that trend to ideal for clustering. In microscopic view, we apply FCM to adjust algorithm process, and to cluster the sample space. Each of the two algorithms performs its own functions. At the same time they are interdependence and interaction, mutual benefit and mutual complementarities to improve quality of clustering.

In FCMP, the core of FCM algorithm is to determine the clustering center, so PSO codes the clustering center. Each particle represents a collection of clustering center. There are $n$ criminal records $X = (x_1, x_2, \cdots, x_n)$, and $K$ types of crime $C = (c_1, c_2, \cdots, c_k)$. Each of the criminals has $D$ behaviors. The position of any particle $x_i$ is defined by $K$ clustering centers, so the position is $K \times D$-dimensional vector. And each particle has velocity and fitness, the velocity is also $K \times D$-dimensional vector. So the particle can be coded as follows:

$$c_{11} c_{12} \cdots c_{1D} c_{21} c_{12} \cdots c_{2d} \cdots c_{K1} c_{K2} \cdots c_{KD}$$

| $v_1 v_2 \cdots v_{KD}$ | $F(x_i)$ |
|---|---|

If the scale of the particle swarm is $N$, there are $K$ types of clustering. We define the fitness function of each particle:

$$F(x_i) = \frac{1}{J(U,C)+1} \tag{6}$$

The better the clustering result is, the smaller $J(U,C)$ is, and then the larger individual fitness $F(x_i)$ is.

The procedures of FCMP are as follows:

Step1: Initialize the particle swarm whose scale is $N$: Initialize a particle on randomly selected $K$ criminal records from the $n$ criminal records, and iterate $N$ times.

Step2: Compute membership degree $u_{ij}$ according to (3).

Step3: Compute the fitness of each particle according to (6).

Step4: Compare each particle's fitness with its $P_{best}$. If the current value is better than $P_{best}$, then set $P_{best}$ equal to the current value

Step5: Compare $P_{best}$ of particles with each other then update $G_{best}$.

Step6: Change the velocity and position of the particle according to (1) and (2) respectively.

Step7: Repeat Step2 to Step6 until to reach the end condition.

Step8: Determine each cluster according to the method of maximum subjection principle.

Where: $u_{ij}$ is the membership degree of the $i^{th}$ criminal record with the $j^{th}$ criminal, $d_{ij}$ is the dissimilar degree between the $i^{th}$ criminal records with the $j^{th}$ criminal, both of which are all determined by the behaviors among the crimes.

## 5. Experimental Analysis

We carry 50 suspicious logs on classification of using the statistical analysis of the log. It carries on the scanning match to related field of each diary record according to the regular of the invasion behavior regular library. Then get a classification of logs, and sequence the logs of each classification $Log = \{l_1, l_2, \cdots, l_{50}\}$.We use FCMP to cluster the 50 logs, $N = 30, K = 5$ and take 50 of iterations, make $\omega$ to linear decrease from 0.9 to 0.4. The result as shown below:

$$C_1 = \{l_{39}, l_{40}, l_{41}, l_{42}, l_{43}, l_{44}, l_{45}, l_{46}, l_{47}, l_{48}, l_{49}\}$$
$$C_2 = \{l_1, l_2, l_3, l_{4,} l_5, l_6, l_7, l_9, l_{10}, 1_{11}, l_{12}\}$$
$$C_3 = \{l_{21}, l_{22}, l_{23}, l_{24}, l_{25}, l_{26}, l_{27}, l_{28}, l_{29}, l_{31}, l_{35}\}$$
$$C_4 = \{l_{13}, l_{14}, l_{15}, l_{16}, l_{17}, l_{18}, l_{19}, l_{20}\}$$
$$C_5 = \{l_{30}, l_{32}, l_{33}, l_{34}, l_{36}, l_{37}, l_{38}, l_{50}\}$$

Obviously, $l_{35}$ is inaccurate in $C_3$, $l_8$ is inaccurate in $C_4$. In fact, $l_{50}$ in $C_5$ is the noisy data which is a normal data. The inaccurate data are all on both ends of classify, it is because that the membership degree is not accurate. In this experiment, the correct rate of FCMP is 97%, and the result meets our requirement basically.

## 6. Conclusion

This article proposed one kind of conception applies the FCMP in the aspect of the computer forensics collection the analytical evidence. Massive suspicious data which collects facing the data collection process, obtains the approximate optimal solution as far as possible, by facilitates the gain useful information. But the FCMP itself has some insufficiencies:

- PSO has the problems of early convergence and poor performance of global convergence and so on. FCM is easy to fall into local optimum and the algorithm convergence speed is greatly influenced by the initial value. Although FCMP has some improvements on these insufficiencies to some extent, the problems remain.
- The algorithm is still in the experimental stage. When process large-scale data, the performance has a declining trend. The setting of some parameters is to be studied.

Therefore, in future work, we should continue to improve the algorithm to make it more suitable for forensic analysis. As far as possible has its own contribution in computer forensics.

## References

[1]Deguang WANG and Lili HAO, "Application of Ant Colony Clustering in Computer Forensics" 2009 Second International Conference on Information and Computing Science,pp87-90.

[2]Y. Shi and R. Eberhart, "Parameter Selection in Particle Swarm Optimization" Proc. Seventh Annual Conf. on Evolutionary Programming, pp. 591-601, 1998

[3]R. Eberhart and Y Shi, "Particle swarm optimization: developments, applications and resources" Proceedings of the 2001 Congress on Evolutionary Computation, Vol. 1, pp. 81 -86, 2001.

[4]Sharaf, A.M. and El-Gammal, A.A.A, "A Discrete Particle Swarm Optimization Technique (DPSO) For Power Filter Design " Design and Test Workshop (IDT), 2009 4th International, pp.1-6

[5]A. Kashefi Kaviani, S.H. Fathi, N. Farokhnia, A. Jahanbani Ardakani, "PSO, an Effective Tool for Harmonics Elimination and Optimization in Multi-level Inverters" Proc Industrial Electronics and Applications, pp. 2902 – 2907, 2009

[6]Hesam Izakian, Ajith Abraham, Václav Snášel. "Fuzzy Clustering Using Hybrid Fuzzy c-means and Fuzzy Particle Swarm Optimization" 2009 World Congress on Nature & Biologically Inspired Computing (NaBIC 2009), pp.1690-1694,2009.