Real-time fault detection approach of software under big data environment

Jian xianrui

Chengdu vocational and technical college, Chengdu, Sichuan, 610041

Keywords: K-means clustering algorithm; feature extraction; real-time software fault detection

Abstract. For large data environment, the traditional K-means clustering algorithm flaw in the software in real-time fault detection process, we propose a K-means clustering software intelligent real-time fault detection method is improved. The K-means clustering algorithm and particle swarm optimization combined during the iterative process, the combination of K-means to optimize the upcoming PSO algorithm offspring individuals use K-means clustering to obtain local optimal solution calculates and uses these individuals continue to participate in an iterative process, so that the algorithm can improve the convergence speed, avoid falling into local optimal solution, to obtain accurate software fault signal characteristics. Experimental results show that the use of K-means tilt feature extraction software, intelligent real-time fault detection algorithm can effectively improve the accuracy of fault detection, and achieved satisfactory results. On the condition of big data, traditional K-mean cluster algorithm showed some flaws of software fault detection in real-time way. In this research, an improved k-mean cluster method was proposed for the purpose of testing software fault in intellectual and real-time way. This improved method combined k-mean cluster algorithm with Particle cluster algorithm. K-mean cluster were used in the optimizing procedure among the iterative process. Individuals from Particle cluster algorithm were computed by using K-mean algorithm and obtained the locally optimal solution, and then these individuals continue to participate in the iterative processing. This improved converging rate of the algorithm, avoided falling into the locally optimal solution and finally got the accurate features of software fault signal. Results showed that this testing method with using the tilt feature of k-mean algorithm improved detection accuracy effectively.

Introduction

Software fault threaten the security of the network seriously [1,2], therefore, it is necessary to test the software accurately in real-time way, take safeguard measures [3] and avoid the expansion of software failures in the network. This problem has been a key issue which scholars and experts focus on [4]. At this stage, software which used commonly under real-time fault detection in big data environments includes artificial immune algorithm [5], neural network algorithm [6] and K-means clustering algorithm [7-9]. In the big data environment, the accuracy of traditional K-means clustering algorithm was reduced due to the subjective choice of the K value, which tends to be easy to get local optimal solution, rather than the global optimal solution in real-time fault detection [10]. In this research, we proposed an improved intellectual software fault detection method on basis of K-means clustering.

The K-means clustering algorithm and particle swarm optimization were combined during the iterative process. The offspring individuals of PSO algorithm were computed by using K-means clustering to obtain local optimal solution, which subsequently participate in the iterative process. Thus, convergence speed of the algorithm were improves, with avoiding falling into local optimal solution and obtaining accurate software fault signal characteristics.

Real-time fault detection process for software

Real-time fault detection process of software includes computer run data acquisition, data preprocessing, fault type extraction, fault detection, fault data analysis and fault testing result

expression. Fault type extraction is the core part of fault detection. The flowchart of fault detection is shown in Figure 1.

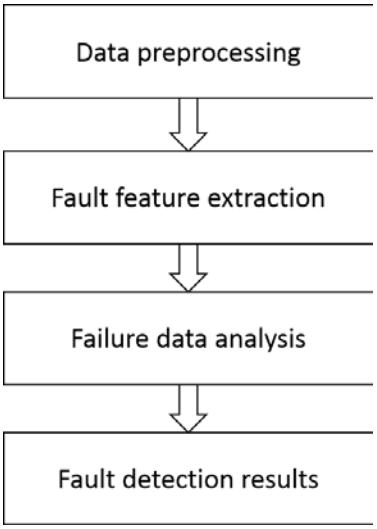


Figure1 The flowchart of software fault detection

K-means clustering algorithm

Principle of K-means clustering algorithm is shown as follows.

Collecting a data set and set a cluster score of K. The value of K is determined by the actual characteristics of the data. In data mining process, the data will be introduced into cluster center according to the duplication of distance function by K-means clustering algorithm. Firstly, K cluster centers are selected randomly, and then the distances between each individuals and the cluster center were calculated. Every individual will be assigned to the nearest cluster center. After all individuals are assigned successfully, each cluster center will be recalculated based on the recent number of individuals. This process will be recirculated until achieve the preset condition of computational termination.

Data mining process based on K-means clustering algorithm is described as below.

(1)K value is determined according to the data characteristics, that is, determining the number of cluster center. Assumed that the data set is m, and $A_j(L)$ is set as the cluster center, where $\mathrm{j}=1,2,...,\mathrm{k}$, and then the distance between the data and the cluster center is calculated.

(2) Initialize the cluster centers $\mathrm{F}(\mathrm{x}_i, \mathrm{A}_j(L))$, $i=1,2,...,m$, $\mathrm{j}=1,2,...,\mathrm{k}$.

(3) Euler method is used to allocate the remaining data to the nearest cluster center data, assuming it match the following conditions:

$$\mathrm{D}(\mathrm{x}_i, A_j(L)) = \min\left\{D(\mathrm{x}_i, A_j(L))\right\} \qquad (1)$$

Then $x_i \in \omega_k$.

(4)Mean value of the data clustering center is set as the mean of new clustering center

$$C(l) = \sum_{j=1}^{k}\sum_{k=1}^{n_j}(\left\|x_k^j - A_j(L)\right\|)^2 \qquad (2)$$

Assuming average value of the cluster centers were the same as that of the last iteration process, $\left\|C(l)-C(l-1)\right\| < \xi$, then the operation is stopped, otherwise returns to step (3), setting the $l = l+1$, new data cluster centers is then calculated, the formula is as follows:

$$A_j(L+1) = 1/\mathrm{n}_j \sum_{i=1}^{k} X_i^j, \quad \mathrm{j}=1,2,...,\mathrm{k} \qquad (3)$$

Computational termination condition is that no data can be reassigned to different cluster center, or cluster centers remain unchanged. Also minimum value of quadratic sum of the local error can be used as a condition to stop computation.

Improvment of the K-mean algorithm with Particle cluster algorithm

Here, we presents a data mining method based on improved K-means clustering algorithm, in which the K-means clustering algorithm and particle swarm optimization were combined. During the iterative process, K-mean value was used in the optimizing process. The offspring individuals of PSO algorithm were computed by using K-means clustering to obtain local optimal solution, which subsequently participate in the iterative process. Thus, convergence speed of the algorithm were improves and avoid falling into local optimal solution.

Basic operation process of PSO algorithm is divided into two stages and described as follows:

(1)PSO relevant parameters are initialized. Setting the category of machine data as c, that is setting c clustering centers. The initial position and velocity of the particle were produced according to cluster centers, among which the particle position $X = \{v_1, v_2, \cdots, v_c\}$ was used to describe the data clustering centers.

(2) Cluster centers for all of the data were calculated by using the K-mean clustering algorithm:

$$d_{ij} = \left\| x_i - v_j \right\| \qquad i = 1, 2, \cdots, n,\ j = 1, 2, \cdots, c. \tag{4}$$

(3)The following formula was used to calculate the fitness value of particles:

$$f(x_i) = \frac{1}{J(U,V)+1} \tag{5}$$

(4) Using the following formula to update the position of the particle. The historical best position would be substituted by the current position of the individual particles, if the current position was assumed as the best position individual particles.

$$P_i^{t+1} = \begin{cases} X_i^{t+1}, & \mathrm{f}(\mathrm{x}_i^{t+1}) \geq \mathrm{f}(\mathrm{x}_i^{t}) \\ P_i^{t}, & otherwise \end{cases} \tag{6}$$

P_i^{t+1} stands for the historical best position of individual particles.

(5) Global optimum position of Particle Swarm is updated by using the following formula. Assuming the current position of a particle is the best among the whole group, the historical global optimum position will be replaced by it.

$$P_g^{t+1} = \begin{cases} p_i^{t+1}, & \mathrm{f}(\mathrm{p}_i^{t+1}) \geq \mathrm{f}(\mathrm{p}_g^{t}) \\ P_g^{t}, & otherwise \end{cases} \tag{7}$$

P_g^{t+1} stands for the historical best position of particles swarm.

(6) The moving speed and position of the particles are updated by using the following formula:

$$v_{id}(i+1) = \omega \times v_{id}(i) + c_1 \times rand() \times (P_{best} - x_{id}(i) + c_2 \times rand() \times (g_{best} - x_{id}(i)) \tag{8}$$

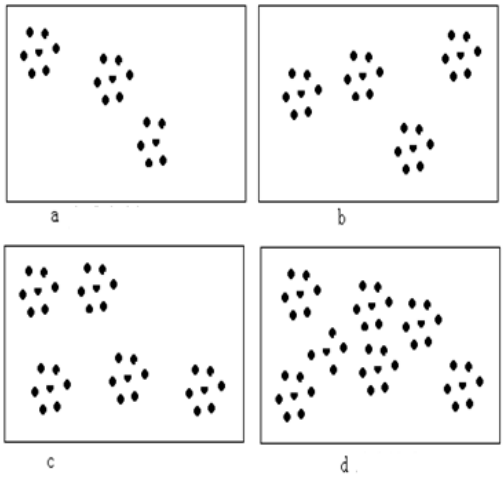
$$x_{id}(i+1) = x_{id}(i) + v_{id}(i+1) \tag{9}$$

(7) Repeating the computing process of steps (2) to (7), until met the iterative conditions. If the assumptions met the condition of termination processing, then output the value of data feature detection. Feature data categories were determined by using Euclidean distance method. Thus achieved the particular data mining.

Application results analysis of the improved algorithm

Analysis of mechanical failure cluster results. In order to verify the effectiveness of the proposed method, it is necessary to do an experiment. Computer hardware for experimental platform is Intel Core i5 4570 quad-core 3.2GHZ CPU, 8G DDR3 1600 memory, windows7 Ultimate operating

system, with using VC ++ algorithm to write the procedures of the algorithm. To test the performance of different algorithms in clustering the software fault in real-time way under big data environments, traditional clustering algorithm and improved algorithm were used to test failure data cluster respectively. In the course of the experiment, four types of software fault data sets were clustered and the clustering results of simulation experiments were shown in Figure 2.



a:Principal component analysis algorithm;b
:Bayesian algorithm; c:K means clustering algorithm d:the proposed algorithm
Figure2 cluster effectiveness of different algorithm

Figure2 indicates that the proposed algorithm is effective for a variety of software failure data clustering, which provided accurate data support for the following feature detection of software fault.

Above-described four types of software fault data set were used in the clustering experiments on basis of different algorithms. Clustering error was taken as criteria for evaluating the performance of different algorithm. Mean cluster error of 100 experiments were obtained as shown in Table 1.

Table 1 mean cluster error of different algorithms

Software fault data set	Traditional algorithm	Improved algorithm
1	0.161	0.098
2	0.132	0.074
3	0.283	0.178
4	0.241	0.169

Data in Table 1 indicated that the accuracy of the improved algorithm is much higher than the traditional one, which reflects the advantages of this algorithm.

Analysis of real time fault data for software. For the selected software in this research, 100 sets of real-time fault sample data were collected and used for the real-time fault detection. Signal characteristics of different software failure data were described in Figure 3.

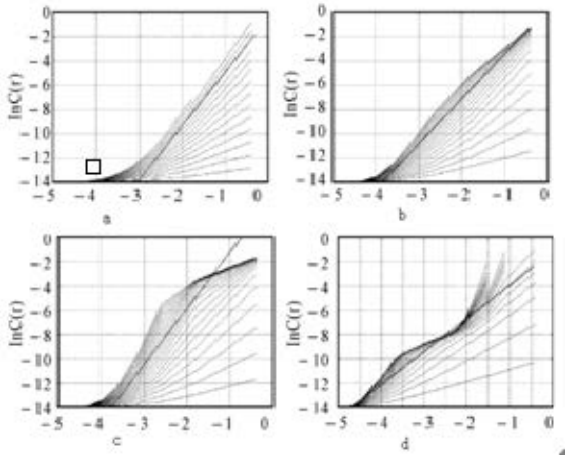


Figure3 the real-time characteristic quantity of different software

Time-domain spectroscopy of real-time fault feature data under big data environment were examined by using the traditional algorithm and improved algorithm. Results were shown in figure 4.

a:traditional K means algorithm; b:the proposed algorithm

Figure4 comparison of test result between different algorithms

Results shown above indicate that the performance of this improved algorithm is better than that of the traditional one in fault signal detection, fault clustering detection, with improving the 17.21% of accuracy. This is because tilt factor is applied in clustering fault feature data, which avoid interference of some local optimal solution and improve convergence speed of the PSO algorithm and also avoid falling into local optimal solution prematurely.

Conclusions

Traditional K-means clustering algorithm possesses flaws in the software real-time fault detection process under big data condition. Therefore, we proposed an improved K-means clustering algorithm for software intelligent fault detection in real-time way. The PSO algorithm was combined with K-means clustering algorithm, in which tilt factor was introduced to avoid local optima caused by small category of fault. Experimental results show that our algorithm improved

the detection accuracy effectively and reduced the fault classification error, thus provided accurate theoretical basis for the real-time diagnosis of software failures in big data environment.

References

[1] Liu Fucheng, Gao Shang. Personal credit scoring based on hybrid support vector machines with cluster analysis. Information Technology,2013,(2):42-44.

[2] Zhang Wuqiang, Mu Ruihui, Zhang Hang. Routing Strategy for Sensor Network with Load Balance Based on Fuzzy K-Means and Node Position. Science Technology and Engineering. 2013,(4):912-916.

[3] Ke Zunhai,Liu Yong, Xu Yichun, Lei Bangjun. An approach based on modified K-means for moving pbjects detection. J of China Three Gorges Univ(Natural Sciences), 2012,24(6):98-102.

[4] Zhang xiaofang. Research on application of clustering analysis algorithms in distance education system. Bulletin of science and technology, 2013,29(4):106-108.

[5] Zhuang Xia, Dai Min, He Yuanqing. Fault Diagnosis of Sensor Node Based on Artificial Immune and Fuzzy K-Means. Computer Measurement & Control, 2013,21(3):611-613.