International Journal of Security and Its Applications
Vol.9, No.10 (2015), pp.177-186
http://dx.doi.org/10.14257/ijsia.2015.9.10.16

An Optimized k-means Algorithm for Selecting Initial Clustering Centers

Jianhui Song, Xuefei Li and Yanju Liu

School of Automation and Electrical Engineering, Shenyang ligong University, Shenyang, Liaoning, 110159, P.R. China
hitsong@126.com

Abstract

Selecting the initial clustering centers randomly will cause an instability final result, and make it easy to fall into local minimum. To improve the shortcoming of the existing k-means clustering center selection algorithm, an optimized k-means algorithm for selecting initial clustering centers is proposed in this paper. When the number of the sample's maximum density parameter value is not unique, the distance between the plurality samples with maximum density parameter values is calculated and compared with the average distance of the whole sample sets. The k optimized initial clustering centers are selected by combing the algorithm proposed in this paper with maximum distance means. The algorithm proposed in this paper is tested through the UCI dataset. The experimental results show the superiority of the proposed algorithm.

Keywords: *k-means; clustering center; density parameter; maximum distance*

1. Introduction

As we all know, clustering algorithm plays an important role in machine learning and data mining. Besides, clustering algorithm has a wide range of applications: economic filed, image classification, target recognition, etc. Clustering algorithm belongs to unsupervised classification which is opposite to supervised classification. The algorithm takes similar objects into a cluster and maximizes the distance between different clusters. The clustering result depends on the similarity of the objects of the related cluster. Clustering algorithm can be divided into five categories: partitioning methods, hierarchical methods, density-based method, grid-based methods and model-based methods. However, it's difficult to give clear boundaries to the five categories since there may exist overlap among the five classes. Each category has few standard algorithms. In consideration of that the paper is mainly based on partitioning method, which will be introduced little more.

Most portioning methods are carried out based on distance. For a given dataset S which contains n samples, the dataset is divided into k parts which is less than n, and the initial parts are obtained. The object is moved from one group to another one by iterative method of relocation. The whole process needs to be carried out on the condition that the distance between any two objects of the same cluster is the shorter the better, and the distance between any two objects in different clusters is the farther the better. Algorithms like k-means, k-medoids and clarans all follow this principle. All of these algorithms are suited to small and medium-sized datasets, and they need to be extended when used for large-scale datasets.

Being a kind of unsupervised clustering algorithm, k-means is reliable in theory and simple. It has a wide application in fields of image processing, pattern recognition, data mining, etc. However, there exist some problems that cannot be ignored.

ISSN: 1738-9976 IJSIA

Firstly, the algorithm requires specific pre-given number k of clusters which needs to be estimated through the different specific application. For example, in the application of bow (bag of words) model, the model varies with k which will affect the final recognition rate. Secondly, the choice of initial clustering centers is important. If there are no good initial clustering centers, the final clustering result will fall into local minimum and cause the final result invalid probably. The traditional method of k-means algorithm on choosing initial clustering center usually adopts the random method which can reduce the possibility of falling into local minimum to some degree and get efficient result. However, when facing large amount of data, it will be a great waste of time. The k-means algorithm based on division need to calculate the distance between samples and clustering center and update the present clustering center continuously. Thus, it costs a great amount of time during the process of clustering.

To solve the first problem, Zhou Shibing [1] takes Silhouette as efficient criterion to judge the number of clustering centers k. Zhou puts forward a new kind of algorithm for determining the number of clustering centers by combining with algorithm of max-min distance and AP (Affinity Propagation clustering). And the experimental comparison proves that Zhou's algorithm is superior to traditional algorithm. To solve the second problem, Tong Xuejiao [3] obtains the number of clustering centers according to the idea of greedy algorithm and makes the final clustering result have features of higher accuracy rate and stability. Han Lingbo [4] proposes a kind of algorithm which is based on maximum intensity parameter by analyzing the intensity parameter of sample data and Han gains improvement on stability and accuracy rate. Huang Min [5] has improved algorithm which is based on maximum intensity parameter on the basis of analyzing Han's shortages and her result is better than Han's. Xing Changzheng [6] makes some processes on isolated points on the basis of references [5], and puts forward a kind of algorithm for choosing initial clustering centers based on average density. Xing has reduced the algorithm's sensitivity to isolated points. To solve the problem of facing massive data, Zhou Lijuan [4] discloses the main design approach and strategy of parallel k-means algorithm, and puts forward a kind of parallel clustering algorithm based on programming framework of MapReduce. Zhou's algorithm can handle the massive data conveniently.

The algorithm based on distribution of maximum density proposed by Han Lingbo has obtained a better result than traditional algorithm; however, it does not solve the problem properly when the number of related samples that have the maximum density parameter value is not single. Huang Min has solved this problem and obtained the advanced algorithm based on distribution of maximum density, but she ignored the possibility that the related samples with same maximum density parameter may belong to different clusters. Therefore this paper put forward an advanced algorithm based on Huang Min's. It takes into consideration of the relationship among the samples that have the same maximum density parameters.

The algorithm proposed in this paper improves the accuracy of clustering and the integrity of k-means algorithm. Figure 1 shows the detailed flowchart of the proposed algorithm. The related explanations of the expressions in figure 1 can be seen in chapter 2.

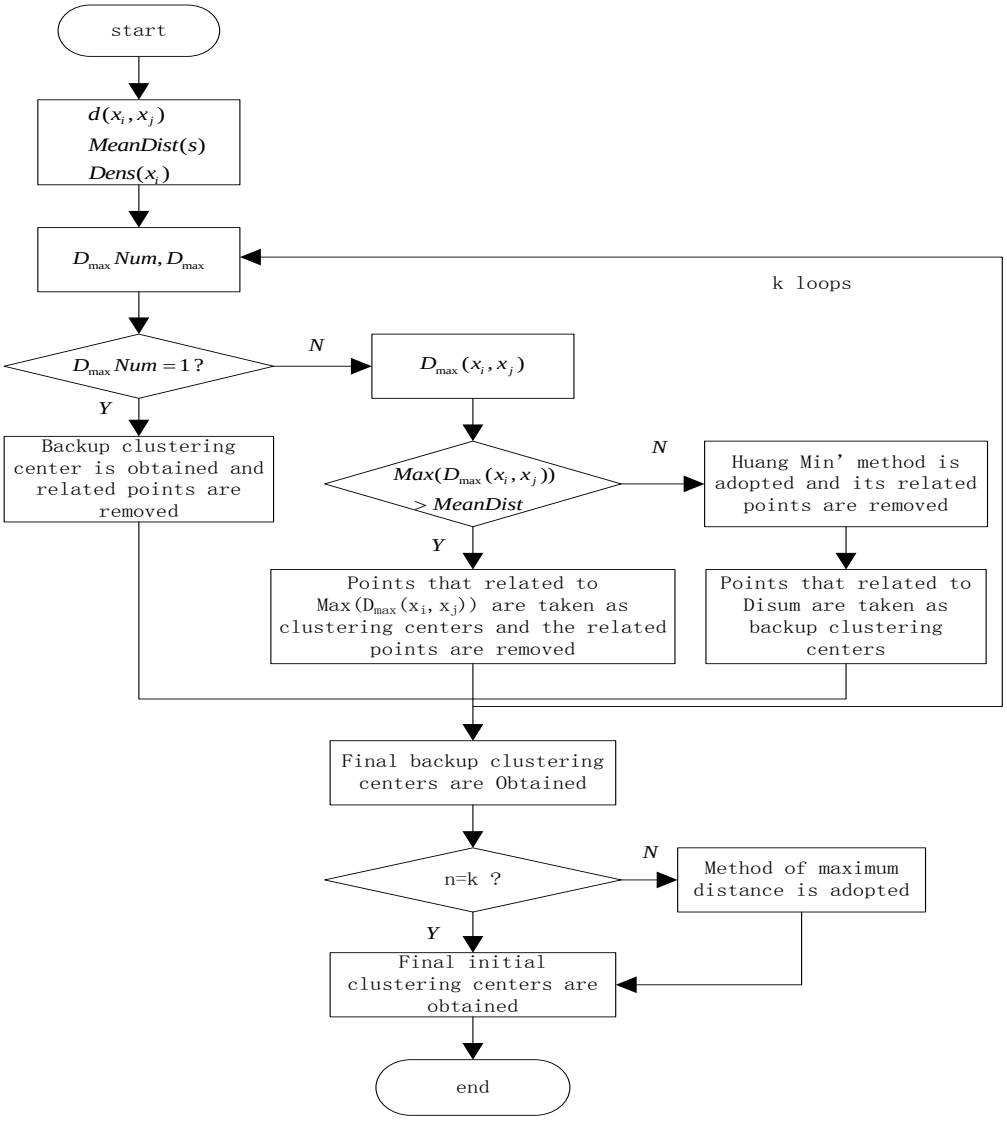

Figure1. Algorithm Flow Chart

2. The k-means Algorithm based on the Distribution of Maximum Density Points

The clustering data samples set are assumed as below:

$S = \{x_1, x_2, \cdots, x_n\}$, $x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$, $i = 1, 2, \cdots, n$;

The number of initial clustering centers: k.

Definition 1, the Euclidean distance between two samples of the data set S:

$$d(x_i, x_j) = \| x_i - x_j \| \tag{1}$$

Where, $x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$.

Definition 2, the average distance of data set S:

$$MeanDist(s) = \frac{\sum d(x_i, x_j)}{C_n^2}, \tag{2}$$

Where n stands for the number of samples in S, C_n^2 stands for the number of combination of any two samples.

Definition 3, the density parameters of each sample in data set S:

$$Dens(x_i) = \sum_{j=1}^{n} u(MeanDist - d(x_i, x_j)) \quad (3)$$

Where, $u(z) = \begin{cases} 1 & z \geq 0 \\ 0 & z < 0 \end{cases}$.

The k-means algorithm based on the distribution of maximum density points can be obtained in reference [4].

Here, only the main part is shown as follows:

⋮

for $i = 1:k$

Each sample's density parameter is calculated according to formulate (3), and all density parameters are taken into density parameter set D.

The max one is found out and taken as clustering center. The points' density parameters whose distance to the clustering center is less than the average distance of the dataset are removed from the density parameter set D.

end for

⋮

To solve the problem that the number of samples with same maximum density parameter may be not the only one, Huang Min gets some improvements as follows:

⋮

for $i = 1:k$

Each sample's density parameter is calculated according to formulate (3), and all density parameters are taken into density parameter set D.

The maximum density parameter D_i is found out from D, and the number *maxl* of D_i is checked out:

if $maxl == 1$

The detail process should consult the step 3 of reference [4].

else

All the samples that related to the maximum density parameter D_i of dataset S are found out.

The candidate clustering centers are assumed to be x_i, x_j, x_k, and then the sum of the candidate centers' distance to all samples is calculated separately when $d(x_i, x_j) \leq meansDist(S)$, and taken into set *SUM*.

The related sample is found out and taken as the initial clustering center after the $Min(SUM)$ is calculated.

The density parameters are removed from set D under the condition when $d(x_i, x_j) \leq meansDist(S)$.

end for

⋮

In Huang Min's advanced k-means algorithm, she calculates the sum of candidate clustering centers' distance to all samples separately when $d(x_i, x_j) \leq meansDist(S)$, and chooses samples related to $Min(SUM)$ as clustering center. The $Min(SUM)$ represents that on the condition of same density parameter, and thus its relationship to around samples appears more compact. It meets the requirement to take them as one cluster. However, Huang has overlooked the relationship among the samples with the same maximum density parameter. Samples with the same maximum density parameter may belong to the same cluster or may not. To solve the

relationship of these samples, an optimized k-means algorithm for selecting initial clustering centers is proposed in this paper.

3. The Optimized k-means Algorithm for Selecting Initial Clustering Center

The specific technical process of the optimized k-means algorithm for selecting initial clustering centers proposed in this paper shows as follows:

Input dataset: $S=\{x_1, x_2, \cdots, x_n\}$, K :the number of clustering centers;

Output: *centorid* , initial clustering centers;

Steps:

The distance of any two samples and the average distance $MeanDist(s)$ of data set S are calculated according to formulate (1) and (2).

for $i=1:k$

Each sample's density parameter is calculated according to formulate (3), and taken into density parameter set D.

The maximum density parameter $D_{\max}$ of set D is found out and the number $D_{\max}Num$ of $D_{\max}$ is checked out.

if $D_{\max}Num==1$

The sample x_i related to $D_{\max}$ is taken as clustering center.

Distances between x_i and all samples are calculated. Then the samples whose distance to x_i is less than $MeanDist(s)$ are found out and the related density parameters are removed from density parameter set D.

else

The samples x_i, x_j, x_k , etc that have the same maximum density parameter are found out. Distances of any two of these samples are calculated. The maximum distance $Max(d)$ is calculated and compared with $MeanDist(s)$.

if $Max(d) < MeanDist(s)$

The sum of distance of x_i, x_j, x_k and all samples are calculated separately on the condition that the distance is less than $MeanDist(s)$. These distances are taken into set SUM.

$Min(SUM)$ is calculated and the related sample which can be one of the initial clustering centers is found out.

The related parameters are removed from set D.

else

Related samples where distance of any two of those on the condition when $Max(d) > MeanDist(s)$ are found out and taken as initial clustering centers.

end if

end if

end for

Assuming that after K times of loop, it gets N candidate clustering centers and the centers are in *centroidBackup* .

if $N == K$

centorid = *centroidBackup* ;

else

Distance $d(x_i, x_j)$ of any two of *centroidBackup* and the maximum distance $Max(d)$ are calculated. The two centers related to $Max(d)$ are taken as

initial clustering centers. Here, it assumes that the two related samples are x_1, x_2 and put them in *centorid* .

for $i = 3:k$

The remaining $(N-(i-1))$ points' distance to the samples that does not include the *N* candidate clustering centers is calculated. These distances are multiplied separately and the sample related to maximum result is putted in *centorid*.

end for

end if

Samples with the same density parameter may belong to the same cluster or not. Calculating the distance of any two of these samples and compare the distance with *MeanDist*(*s*) can distribute these samples to the related cluster properly. On the basis of this assumption, it may distribute some of *K* clustering centers into the same class by Min Huang's algorithm. The algorithm of Changzheng Xing has reduced the isolated points' effect to k-means algorithm. However, the problem found by this paper has not been resolved.

The problem can be solved by adopting the optimized k-means algorithm for selecting initial clustering center proposed in this paper. After obtaining the samples with the same maximum density parameter, the distance of any two of related samples is compared with the average distance of the dataset. If $d(x_i, x_j)$ is less than *MeanDist*(*s*) , the proposed algorithm follows the method of Min Huang. If $d(x_i, x_j)$ is greater than *MeanDist*(*s*), all of the related samples on the condition that $d(x_i, x_j)$ is greater than *MeanDist*(*s*) are found out and taken as candidate clustering centers. In order to find out all the candidate clustering centers, it lunch the loop *K* times from step (2) to step (26) and the *N* candidate clustering centers are assumed to be gotten after *K* times of loop. It is obviously that *N* is greater than *K* possibly. If *N* is equal to *K* , the candidate clustering centers are taken as initial clustering centers. However, if *N* is greater than *K* , the maximum distance method of Donghai Zhai [8] is adopted. The distance of any two of the candidate clustering centers is calculated and the two samples with the maximum distance are taken as initial clustering centers. The distance of the remaining $(N-2)$ points to the exiting initial clustering centers is computed separately, and each remaining point's distance to the existing center is multiplied. Then the maximum is found out and taken as the third initial clustering center. And so on until *K* efficient initial clustering centers are found out.

4. Experimental Results and Analysis

The three UCI dataset: iris, balance and soybean-small are used for testing the performance of the traditional algorithm, the Min Huang's algorithm and the optimized k-means algorithm for selecting initial clustering centers proposed in this paper. The related introduction is show in table 1.

Table 1. Introduction to Dataset of UCI

	Number of classes	Number of samples
iris	3	150
balance	3	625
soybean-small	4	47

As the traditional algorithm selects the initial clustering centers by random method, therefore the experiment of the traditional algorithm is carried out 10 times to reduce the possibility of falling into the local-minimum. The maximum, minimum and mean value of the 10 times experiment are counted. Huang's algorithm and the proposed algorithm do not adopt the random method and lunch only one time. The statistical results are compared between the traditional algorithm, the Huang Min's algorithm and the proposed algorithm in table 2.

The experimental results of three kinds of algorithms on different dataset are shown below in figure 2, figure 3 and figure 4. The red line stands for the result of traditional algorithm. The green line stands for the result of the Min Huang' algorithm and the blue line stands for the algorithm proposed in this paper.

It can be seen from figure 2, figure 3 and figure 4 that the traditional k-means algorithm has the feature of instability, and the clustering result varies with different kind of dataset. Therefore, there is a big drawback of the traditional k-means algorithm using the random selection of initial cluster, which will cause inconvenience to research and application. Especially when the traditional k-means algorithm is used for large scale dataset, it will waste amount of time definitely.

Table 2. Experiment Result of Three Kinds of Algorithms

		balance	iris	soybean-small
Traditional k-means algorithm	1	48.16%	89.33%	72.34%
	2	49.76%	89.33%	72.34%
	3	50.08%	89.33%	72.34%
	4	51.84%	86.00%	72.34%
	5	50.56%	89.33%	89.36%
	6	55.36%	89.33%	72.34%
	7	54.56%	86.00%	72.34%
	8	56.48%	89.33%	72.34%
	9	52.96%	89.33%	72.34%
	10	52.16%	86.00%	72.34%
	minimum	48.16%	86.00%	72.34%
	maximum	56.48%	89.33%	89.36%
	average value	52.19%	88.33%	74.04%
Huang Min's algorithm		45.76%	88.67%	93.62%
Algorithm of this paper		62.44%	88.67%	96.60%

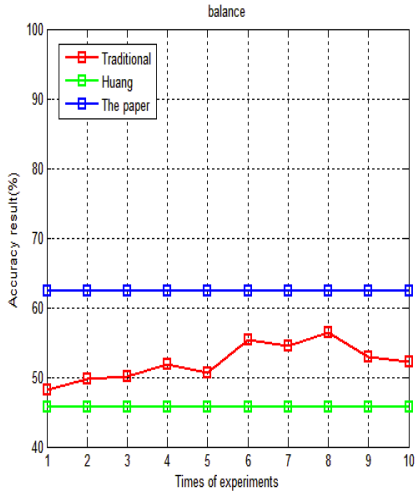

Figure 2. Balance Dataset

From figure 2, it can be observed that the algorithm of Huang Min gets a poor performance than the traditional k-means algorithm. The minimum of the traditional k-means algorithm is slightly higher than the algorithm of Min Huang. The performance of the proposed algorithm in this paper obtains a better one obviously. Moreover, this paper's performance is greater than the algorithm of Yuan Fang [9] whose final result is 52.20% testing with the dataset balance.

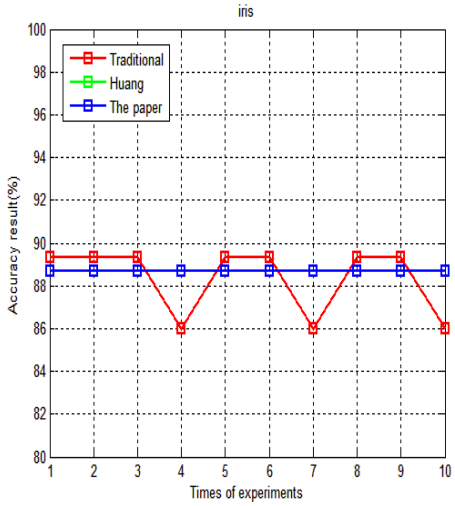

Figure 3. Iris Dataset

From figure 3, it can be seen that there only exist the blue line and the red line. The reason can be found in table2, the experimental result of dataset iris of Huang Min's algorithm and the optimized k-means algorithm proposed in this paper are the same; the green line is covered by the blue line in figure 3. By observing figure 3 and table 2, the results of the Huang Min's algorithm and the optimized k-means algorithm proposed in this paper are lower than the maximum value of the traditional algorithm, but slightly greater than the minimum and average value.

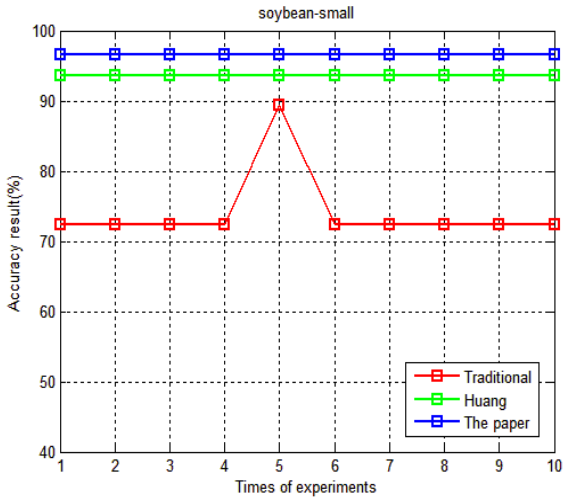

Figure 4. Soybean-small Dataset

By observing figure 4, the experimental results of the three algorithms lunched on dataset soybean-small can be clearly seen. The number that the traditional k-means algorithm reaches the maximum value is only once; the performance of the Min Huang's algorithm and the proposed algorithm in this paper are higher than the traditional algorithm. Although, the traditional algorithm gains the stability on this dataset, however, its performance cannot meet the requirement. The algorithm of Huang Min and the proposed algorithm in this paper gain a better performance and the proposed algorithm in this paper is better than Huang's, which illustrates the effectiveness of the proposed algorithm in this paper.

In summary, the experiment proves the necessity and effectiveness of this paper to solve the shortcoming of reference [5]. The optimized k-means algorithm for selecting initial clustering centers proposed in this paper can distribute the clustering centers into different clusters properly and can obtain better clustering result.

5. Conclusions

Traditional k-means algorithm obtains the initial clustering centers by using random method, which makes it easy to fall into local-minimum and affect the final clustering result. Although the traditional k-means algorithm can lunch many times, it may not get the ideal result and it is a waste of time.

To solve this problem, many researchers have lunched the related improvements. Min Huang acquires good results by her improvements, but has overlooked the relationship among the samples that have maximum density parameters. Thus Huang's idea cannot perform better on some certain datasets. In response to this issue, the paper puts forward related solutions that the distance of samples with maximum density parameters is compared with the average distance of dataset. By this paper's algorithm, the clustering centers can be distributed into different clusters properly. Finally, the number of candidate clustering centers with the default number is compared. If the former is greater than the latter, it selects the needed clustering centers by adopting the method of max-min distance. The experiment has proved the reliability of the proposed algorithm in this paper.

Acknowledgements

The study is funded by the 2014 Education Department of Liaoning Province (Project No.:L2014079).

References

[1] S. B. Zhou, Z. Y. Xu and X. Q. Tang, "New method for determining optimal number of clusters in K-means clustering algorithm", Computer Engineering and Applications, vol. 46, no. 16, **(2010)**, pp. 27-31.
[2] B. J. Frey and D. Dueck, "Clustering by passing messages between data points", Science, vol. 315, no. 5814, **(2007)**, pp. 972-976.
[3] X. J. Tong, F. R. Meng and Z. X. Wang, "Optimization to k-means initial cluster centers", Computer Engineering and Deaign, vol. 32, no. 8, **(2011)**, pp. 2721-2723+2788.
[4] L. B. Han, Q. Wang, Z. F. Jiang and Z. Q. Hao, "Improved k-means initial clustering center selection algorithm", Computer Engineering and Applications, vol. 46, no. 17, **(2010)**, pp. 150-152.
[5] M. Huang, Z. S. He, X. L. Xing and Y. Chen, "New *k*-means clustering center select algorithm", Computer Engineering and Applications, vol. 47, no. 35, **(2011)**, pp. 132-134.
[6] C. Z. Xing and H. Gu, "*K*-means algorithm based on average density optimizing initial cluster centre", Computer Engineering and Applications, vol. 50, no. 20, **(2014)**, pp. 135-138.
[7] L. J. Zhou, H. Wang, W. B. Wang and N. Zhang, "Parallel K-means algorithm for massive data", Huangzhong Univ. of Sci. &Tech.(Natural Science Edition), vol. 40(S1), **(2012)**, pp. 150-152.
[8] D. H. Zhai, J. Yu, F. Gao, L. Yu and F. Ding, "K-means text clustering algorithm based on initial cluster centers selection according to maximum distance", Application Research of Computers, vol. 31, no. 3,**(2014)**, pp. 713-715.
[9] F. Yuang, Z. Y. Zhou and X. Song, "K-means Clustering Algorithm with Meliorated Initial Center", Computer Engineering, vol. 33, no. 3, **(2007)**, pp. 65-66.
[10] H. S. Park and C. H. Jun, "A simple and fast algorithm for K-medoids clustering", Expert Systems with Applications, vol. 36, no. 2, **(2009)**, pp. 3336–3341.

Authors

Jianhui Song, she received the Bachelor degree in Control Technology and Instrument from Harbin Institute of Technology, China, in 2004, the Master and PhD degree in Instrument Science and Technology from Harbin Institute of Technology, China, in 2006 and 2010. Now she is an associate professor in Shenyang ligong University, China. Her main research interests include multi-sensor information fusion and intelligent detection technology.

Xuefei Li, he received the Bachelor degree in Communication Engineering from Liren College of Yanshan University, China, in 2013. Now he is pursuing a master degree in Shenyang ligong University, China. His main research interests include multi-sensor information fusion and object recognition.

Yanju Liu, she received the Bachelor degree from Shenyang University of Technology, China, in 1987, the Master degree in Automation Engineering from Northeastern University, China, in 1996, and PhD degree in Motor and Electrical Engineering from Shenyang University of Technology, China, in 2011. Now she is a professor in Shenyang ligong University, China. Her main research interests include intelligent instrument, multi-sensor detection and Information fusion technology.

