International Journal of Database Theory and Application
Vol.9, No.3 (2016), pp. 229-238
http://dx.doi.org/10.14257/ijdta.2016.9.3.22

Generalization Threshold Optimization of Fuzzy Rough S*et al*gorithm in Healthcare Data Classification

Beibei Dong, Yu Liu, Benzhen Guo and Xiao Zhang*

The College of Information Science and Engineering, Hebei North University, Zhangjiakou Hebei, 075000, China
**Corresponding author*
E-mail:xz1965cn@aliyun.com

Abstract

There is ineffective classification problem in application of K-means clustering algorithm in massive data cluster analysis. This paper presents a K-means algorithm based on generalization threshold rough set optimization weight. Firstly, utilize attribute order described method, using the average distance calculation with Laplace method to optimize the generalization threshold of fuzzy rough set , then the Euclidean distance metric is used in the calculation of the similarity of K-means algorithm, introducing the variation coefficient into the cluster analysis, clustering the Euclidean distance weighted K-means algorithm totally based on data, finally, combine the rough set algorithm based on the generalization threshold optimization and K-means clustering algorithm, applied to medical and health data classification. The K-means algorithm based on generalization threshold rough set optimization weight presented by this paper has a better effect on medical and health data classification.

Keywords*: clustering mining; K-means clustering algorithm; fuzzy rough sets; generalization threshold; weighted Euclidean distance; healthcare data*

1. Introduction

In recent years, with the gradual development of medical information, most of the hospital have established digital medical information systems and electronic patient health records. After the generation and collection of massive medical data, how to store massive, heterogeneous, real-time and multiple medical data efficiently so that large scale complicated medical data can be searched rapidly and responded accurately. Analysis the real-time information and historical data comprehensively, afford medical workers reference of diagnosis and treatment, or provide nursing project to the end-users directly by Cloud service [1]. Meanwhile, the storage and processing platform of medical large data can help analyze historical medical data by data mining techniques theory, mining key physiological characteristics of the patient from the mass of medical information in big data, detecting early disease and predicting possible health risks reliably and efficiently to provide patients valuable medical services [2].

As the foreign health industry is more advanced, data warehouse technology has been widely used in clinical decision support and management decision support fields. Kerkri *et al*. did research in the cross-platform data integration method of medical data warehouse, processing the methods architecture to restructure and consolidate the heterogeneous autonomy patient medical data in different information systems, totally considering the security of personal information privacy protection [3]. Wisniewski *et al*. proposed construction methods and examples of hospital infection control system disease analysis based on three hospital data integration, totally considering the security of Personal Information

ISSN: 2005-4270 IJDTA

Privacy [4]. Scotch *et al*. established a cost estimation system based on data warehouse in the platform of a hospital in Canada to reduce medical malpractice rates, reduce the cost of health care costs, improve efficiency of the system target and have a good effect [5]. Lyman *et al*. designed and analyze drug misuse data mining with a data warehouse of Japanese hospital and other clinical as a data source, provided decision support to improve the quality of medical care [6]. In addition, decision support system based on data warehouse is also widely used in medical quality management, disease management and effectiveness evaluation in foreign countries. JH Peng *et al*. expressed that as the health information technology accelerated, the type and size of medical is increasing at an unprecedented rate, so the health sector has entered the "era of big data."[7] This paper summarizes the new challenges in health information technology of big data era based on the analysis of the medical large data basic concepts, introduces the measurement to restore, manage, integrate, and utilize the medical data in big data era in Shanghai Zhabei, and think about the next step of the work.GH Zhou *et al*. first described the status of health care in the field of data resource management , then discussed the big data technology application in health care areas, including disease management, public health, health management, medical research and other aspects combining the applications of the big data technology in medicine and health at home and abroad, and presented challenges and recommendations of health area in the big data Era [8]. CW Chang *et al*. sued artificial neural network to predict the diagnosis of Parkinson's disease with accuracy of 92.9% [9]. Nan Li used rough set data mining process to diagnose and evaluate the lung cancer, presenting a new idea for the future of medical diagnostics [10].

This paper proposes a weighted K-means algorithm based on the generalization threshold rough set optimization weight aimed at K-means clustering algorithm defects, and carries out simulation experiments to verify the effectiveness of improvement strategies.

2. Advantages and Disadvantages of K-Means Clustering Algorithm

The K-means clustering algorithm is a dynamic clustering algorithm based on the partition, which is both an important branch of the data mining and one of the most commonly used clustering algorithm in practice. The sum of the square error criterion function is often used as the clustering criterion function. The sum of the square error criterion function is defined as:

$$E = \sum_{i=1}^{k} \sum_{p \in c_i} \left| p - m_i \right|^2 m_i c_i \qquad (1)$$

Among them, p is the point in sample space, E is the sum of all objects' square error, p is the point represents the given data objects in the space. m_i is the average value of c_i .

K-means clustering algorithm also has some defects. The algorithm based on the k determined in advance to divide the sample need to be clustered in k kinds K-mean, and minimizes the square sum of the distance between all sample and clustering centers in clustering domain. Therefore, K-means algorithm requires the users to specify the number of the clusters k to be generated before running the algorithm.

In addition, K-means algorithm selects the clustering centers by the random method. The clustering effect will be poor if the initial clustering center isn't chosen appropriately. The initial dependence may lead to instability of the clustering results. Meanwhile, the local optimal solution is obtained instead of global optimal solution.

Furthermore, as the outlier and the noise point keep away from data-intensive area, it will lead the cluster center deviate the real data-intensive area, consequently, K-means algorithm. So K-means algorithm is sensitive to outlier data of noise, but a little data have a huge effect on average. In intrusion Detection, outlier means invasion. So the detection and analysis of outlier is some of valuable direction.

As an unsupervised clustering Intrusion Detection Technology, K-means algorithm is based on two assumptions: first, assume that during all of the main activities, the invasion is little with respect to other normal activities; second, invasion is different from normal activities. Based on these two assumptions, we can assume a dataset to record the main activities, k data is marked as initial cluster centers, then divided into k clusters based on the distance to the k center, unmarking the initial cluster centers, calculating the new cluster center value according to data from the cluster. k clusters are divided according to the k' cluster center to select new center value. Repeat the steps until data convergence.

Though K-means algorithm has a lot of shortcoming, for example it is sensitive to initial value, susceptible to data variation and easy to be trapped into optimum, and the algorithm is sample fast and suitable for large data detection. So, in this paper, the improved algorithm is used in medical health data classification.

3. A K-Means Algorithm Based on Generalization Threshold Optimization Rough Set

3.1 A Rough Set Based on Generalization Threshold Optimization

In order to get useful data from storage equipment to make the hidden information valuable, a rough set can rebuild attribute set and the use the new attribute set to replace the original.

In probability and statistics, correlation coefficient represents the correlation of two-dimensional random variable, while weights are just between 0~1, so we can get the right value by the correlation coefficient. Assume Z_i is condition attribute, Y_i is solution decision attribute. The correlation coefficient of condition attribute and categorical attribute is as follow:

$$\rho(Z,Y) = \frac{\mathrm{cov}(Z_i,Y)}{\sqrt{D(Z_i)D(Y)}} \qquad (2)$$

Then, assume the weight is:

$$W_{A_i} = \left|\rho(Z,Y)\right| = \left|\frac{\mathrm{cov}(Z_i,Y)}{\sqrt{D(Z_i)D(Y)}}\right| \qquad (3)$$

Secondly, mutual information can also be used to reflect the effect of condition attribution on decision attribution. Assume C condition attribution, D decision attribution, and mutual information I is:

$$I(C,D) = \sum_{C_i \in C} \sum_{D_j \in D} P(C_i,D_j) \log_2 \frac{P(C_i,D_j)}{P(C_i)P(D_j)} \qquad (4)$$

Then, assume the weight of condition attribution C_i is:

$$W_{B_i} = \frac{I(C_i,D)}{\frac{1}{n}\sum_{j=1}^{n} I(C_i,D)} \qquad (5)$$

In data mining for domain user, contracted condition attribution will be arranged in the order of the effect on decision attribution, as $c_1 \succ c_2 \succ \ldots \succ c_{|C|}$,and calculate the weight using average distance method with Laplace:

$$W_{K_i} = \frac{|C| - i + 1}{|C| + 1}, i \le |C| \qquad (6)$$

Finally, the new weight is the average of two.

3.2 K-Means Algorithm Based on Weighted Euclidean Distance

K-means algorithm requires the users to select the cluster number k, and it is easy to provide a local optimal solution in the process of cluster. This paper takes the Euclidean distance metrics to calculate the similarity, the algorithm steps are as follows:

(1)Determine a set S_n contains n data objects. $S_n = \{x_1, x_2, ..., x_n\}$ .Firstly, select two data objects w , v , which have maximum distance as the initial clustering center:

$$d_{wv} = \max\{d_{ij}, i, j \in 1, 2, ..., n\} \qquad (7)$$

Set: $x_1^* = x_w$, $x_2^* = x_v$, $d_{wv} = d_1^*$;

(2)Calculate the other $n-2$ data objects in set S_n according to the Euclidean distance, and take x_1^* , x_2^* as clustering center to classify kinds, namely:

$$\forall i \in \{1, 2, ..., n / w, v\} \qquad (8)$$

If the formula (4) is satisfied, then divide x_i into kind x_1^* , or divide x_i into kind x_2^* .Take the x_1^* , x_2^* as clustering center, set S_n is divided into two categories, denoted as S_{21}^* , S_{22}^* ,respectively.

$$\left|x_i - x_1^*\right| < \left|x_i - x_2^*\right| \qquad (9)$$

(3)Calculate the distance from all data objects to x_1^* in kind S_{21}^*

$$d_{21} = \max\{|x_i - x_1^*|, x_i \in S_{21}^*\} \qquad (10)$$

Calculate the distance from all data objects to x_2^* in kind S_{22}^*

$$d_{22} = \max\{|x_i - x_2^*|, x_i \in S_{22}^*\} \qquad (11)$$

Take $d_2^* = \max\{d_{21}, d_{22}\}$, the corresponding data object is denoted as x_3^* .

(4)If $d_2^* > hd_1^*$ (h is the input parameter, usually gained by the experience),take x_3^* as the third point of clustering center, take x_1^* , x_2^* , x_3^* as kind center, divide S_n into three categories, denoted as S_{31}^* , S_{32}^* , S_{33}^* .

(5)Calculate the distance from all data objects to x_1^* in kind S_{31}^* , then

$$d_{31} = \max\{|x_i - x_1^*|, x_i \in S_{31}^*\} \qquad (12)$$

Calculate the distance from all data objects to x_2^* in kind S_{32}^* , then

$$d_{32} = \max\{|x_i - x_2^*|, x_i \in S_{32}^*\} \qquad (13)$$

Calculate the distance from all data objects to x_1^* in kind S_{33}^* , then

$$d_{33} = \max\{|x_i - x_3^*|, x_i \in S_{33}^*\} \qquad (14)$$

$$d_3^* = \max\{d_{31}, d_{32}, d_{33}\} \qquad (15)$$

The corresponding data object is denoted as x_4^* .

(6)If $d_3^* > h \cdot average(d_1^* + d_2^*)$, then take x_4^* as the fourth clustering center, return to (4),or finish the algorithm. The final clustering center is x_1^* , x_2^* , x_3^* .

After optimizing the K-means algorithm by Euclidean distance,considering the operators usually regard some index as important, it's necessary to give a certain weight to the index, so that each different variable effect on the data can be reflected, Better effect can be obtained to improve clustering result.

Let p dimension vectors coordinates of point m , n , s be $(x_{m1}, x_{m2},..., x_{mp})$, $(x_{n1}, x_{n2},..., x_{np})$, $(x_{s1}, x_{s2},..., x_{sp})$,satisfy:

$$(x_{s1}, x_{s2},..., x_{sp}) = (x_{m1}, x_{m2},..., x_{mb} + c,..., x_{mp}) \quad (16)$$

Where, $1 \le b \le p (b \in Z)$, c is a constant, from the definition of weighted Euclidean distance between the two points, then:

$$d_{mn}^{\#} = \sqrt{\sum_{i=1}^{P} \partial_i (x_{mi} - x_{ni})^2} \quad (17)$$

$$d_{sn}^{\#} = \sqrt{\sum_{i=1}^{P} \partial_i (x_{si} - x_{ni})^2} \quad (18)$$

Therefore:

$$\begin{aligned} (d_{mn}^{\#})^2 - (d_{sn}^{\#})^2 &= \sum_{i=1}^{P} \partial_i (x_{mi} - x_{ni})^2 - \sum_{i=1}^{P} \partial_i (x_{si} - x_{ni})^2 \\ &= \sum_{i=1}^{P} \partial_i (x_{mi} - x_{ni} + x_{si} - x_{ni})(x_{mi} - x_{ni} - x_{si} + x_{ni}) \end{aligned} \quad (19)$$

From formula （11）:

$$\begin{aligned} &\sum_{i=1}^{P} \partial_i (x_{mi} - x_{ni} + x_{si} - x_{ni})(x_{mi} - x_{ni} - x_{si} + x_{ni}) \\ &= 2\partial_b c (x_{mb} - x_{nb} + c/2) \end{aligned} \quad (20)$$

Let p dimension vectors coordinates of point e , f , h be $(x_{m1} + c, x_{m2},..., x_{mp})$, $(x_{m1}, x_{m2} + c,..., x_{mp})$, $(x_{m1}, x_{m2},..., x_{mb} + c,..., x_{mp})$,then:

$$\begin{aligned} &[(d_{en}^{\#})^2 - (d_{mn}^{\#})^2] : [(d_{fn}^{\#})^2 - (d_{mn}^{\#})^2] : [(d_{hn}^{\#})^2 - (d_{mn}^{\#})^2] \\ &= \partial_1 (x_{m1} - x_{n1} + c/2) : \partial_2 (x_{m2} - x_{n2} + c/2) : \partial_b (x_{mb} - x_{nb} + c/2) \end{aligned} \quad (21)$$

So, the difference of weighted Euclidean distance has direct ratio relations with the weight given by us in this paper. Under such specific condition, the data is given and the data is coincided with formula(21). This is in line with people's subjective consensus so they acknowledge the feasibility and the rationality of weighted Euclidean distance.

However, the subjective experience weighted Euclidean distance has strong subjectivity and casualness and it needs both the professional trait and brilliant understanding of their fields. It's the defect of subjective weighted method. Therefore, this paper introduces the variation coefficient method into clustering analysis to cluster the Euclidean distance weighted K-means algorithm completely based on the data.

Let the matrix contains p dimension vectors of n samples be:

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \quad (22)$$

The average value of index j is:

$$\overline{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij} \quad (23)$$

The standard deviation of index j is:

$$\sigma_j = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2} \quad (24)$$

Variation coefficient is:

$$V_j = \frac{\sigma_j}{\overline{x}_j} \quad (25)$$

Then the weight of each index can be defined as :

$$w_i = \frac{V_j}{\sum_{j=1}^{p} V_j} \quad (26)$$

3.3 Improved K-Means Algorithm Based on Rough Set Optimization

The improved K-means algorithm based on rough set optimization proposed in this paper combined the algorithm based on generalization threshold rough set optimization weight with K-means clustering algorithm. The idea is to combine the rough set with the fuzzy cluster, and improve the fuzzy clustering algorithm with the concept of lower approximation and upper approximation.

The improved algorithm is mainly reflected in the clustering center. The clustering center of improved fuzzy clustering algorithm based on rough set is:

$$V_j = \begin{cases} w_{lower} \cdot \dfrac{\sum_{X_i \in \underline{A}(C_j)} g(u_{i,j}) X_i}{\left|\underline{A}(C_j)\right|}, if\ \overline{A}(C_j) = \underline{A}(C_j) \\ w_{upper} \cdot \dfrac{\sum_{X_i \in (\overline{A}(C_j) - \underline{A}(C_j))} g(u_{i,j}) X_i}{\left|\overline{A}(C_j) - \underline{A}(C_j)\right|}, else \end{cases} \quad (27)$$

Among them,

$$g(u_{i,j}) = \frac{1-\beta}{1+\beta} \cdot u_{i,j}^2 + \frac{2\beta}{1+\beta} \cdot u_{i,j}^2 \quad (28)$$

Define the membership function, the objective function is:

$$E = \sum_{i=1}^{c} \sum_{j=1}^{n} g(u_{i,j}) d_{i,j} \quad (29)$$

Among them,

$$d_{i,j} = (v_i, x_j)^2 = \left\| v_i - x_j \right\|^2 \quad (30)$$

The membership degree is:

$$u_{ik} = \left[\sum_{j=1}^{C} \left(\frac{d_{ik}}{d_{jk}} \right)^{2/(m-1)} \right]^{-1}, x_k \neq v_i \quad (31)$$

The improved algorithm flow is:

(1)data preprocessing. We'd better to take dimension reduction method to process much complicated high dimension data. Therefore, we use principal component analysis to reduce dimension.

(2)Determine the number of class K. The initial number of class is uncertain in K-means algorithm. It is determined by the experimenter. So the clustering effect will be better if we determine the number of class K use the algorithm.

(3)Calculate the clustering center with the fuzzy clustering algorithm based on rough set.

(4)Obtain clustering results.

The number of class K is determined by initialization in K-means clustering, instead of being calculating, K is obtained by the experimenter according to the different requirement. However, different clustering results are obtained with different K. When the number of class K is uncertain, the results will be quite different if we defined K randomly. Therefore, we need to show the results of K with fuzzy clustering validity function.

$$S = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n}(u_{i,j})^{m}\left\|V_i - X_j\right\|^2}{n\left[\min_{i,j} d^2_{w(i,j)}\right]} \qquad (32)$$

Where, $u_{i,j}$ is membership degree, V is clustering center, x is objects, d is Euclidean distance.

4. Algorithm Performance Simulation

In order to demonstrate the improved performance of the proposed algorithm, we carry out simulation. Firstly, do convergence simulation on K-means algorithm based on generalization threshold rough set optimization compared with the standard rough set algorithm comparison, as the Figure 1 shows.

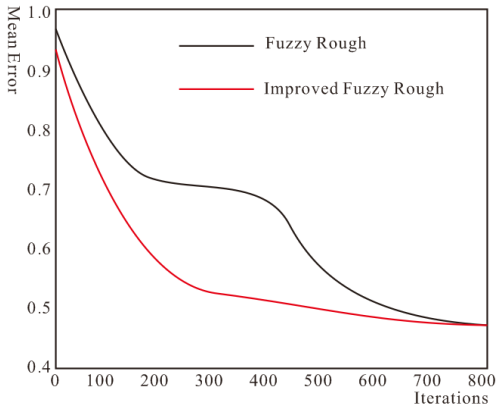


Figure 1. Improved Convergence of Rough Algorithm

Then select 20 healthcare datasets to make clustering simulation based on weighted Euclidean distance K-means algorithm, compared with K-means algorithm, as the Figure 2 shows.

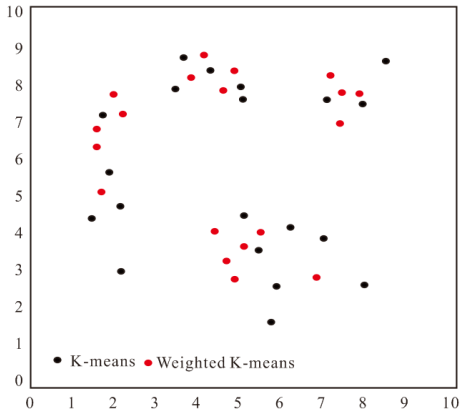


Figure 2. Weighted K-Means Clustering Algorithm

Finally, make clustering effect simulation of K-means algorithm based on improved rough optimization, compared with the weighted Euclidean distance K-means algorithm, as the Figure 3 shows.

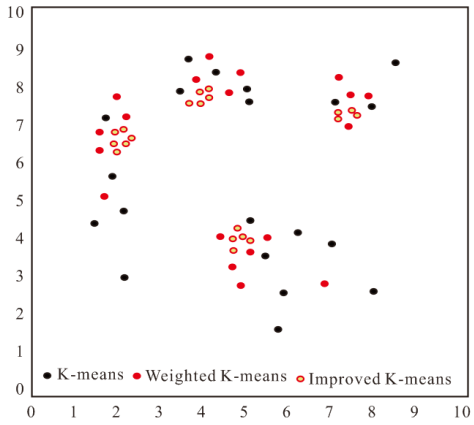


Figure 3. Improved K-Means Clustering Algorithm

As can be seen from the results above, The K-means algorithm based on generalization threshold rough set optimization weight presented by this paper has a better effect on medical and health data classification, compared with K-means algorithm based on weighted Euclidean distance and the K-means algorithm.

5. Conclusion

With the gradual development of medical information, after massive medical data undergoes generation and collection, how to store massive, heterogeneous, real-time, varied medical data efficiently; how to analyze historical medical data by data mining techniques theory, detecting early disease and predicting possible health risks reliably and efficiently to provide patients valuable medical services, is the problem to be solved of medical field in big data era.

This paper proposes a K-means algorithm based on the generalization threshold rough set optimization weight aimed at K-means clustering algorithm defects, and carries out simulation experiments to verify the effectiveness of improvement strategies. The K-means algorithm based on generalization threshold rough set optimization weight presented by this paper has a better effect on medical and health data classification, compared with K-means algorithm based on weighted Euclidean distance and the K-means algorithm.

Acknowledgment

This work was supported by: Youth Foundation of the Education Department of Hebei Province (QN2014182) Major project of Hebei North University (ZD201301)

Hebei Province Population Health Information Engineering Technology Research Center Youth Foundation of the Education Department of Hebei Province (QN2015225)

Youth Foundation of Natural Science of Hebei North University (Q2014008) Zhangjiakou Department of Science and Technology Project (1421012B)

Research and Design for Prairie Wireless Fire Alarm System Based on ZigBee (1411073B)

References

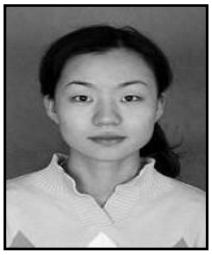
[1] H. Chen, "Design of intelligent medical model based on sensor clustering data mining in IoT", Transducer and Microsystem Technology, vol. 33, no. 4, **(2014)**, pp. 76-79.

[2] Z. Bo, "Moving object tracking algorithm based on variable scale global search", Laser Journal, vol. 36, no. 2, **(2015)**, pp. 30-34.

[3] H. Suyu, "Tracking and Detection Algorithm of Motion Targets Based on Improved Gaussian Mixture Model", Tracking and Detection Algorithm of Motion Targets Based on Improved Gaussian Mixture

Model, vol. 23, no. 3, **(2015)**, pp. 861-863.
[4] X. Tianwei, "Robot Moving Target Tracking Algorithm Based on Adaptive Kalman Filter", Computer Measurement & Control, vol. 23, no. 1, **(2015)**, pp. 173-175.
[5] H. Fengjun, Z. Yanwei and C. Jian, "SIFT Feature Points Detection and Extraction of Three-Dimensional Point Cloud", WIT Transactions on Information and Communication Technologies, no. 60, **(2014)**, pp. 603-611.
[6] H. Fengjun and Z. Yanwei, "Comparative research of matching algorithms for stereo vision", Journal of Computational Information Systems, vol. 9, no. 13, **(2013)**, pp. 5457-5465.
[7] S. Yanzha, "Visual tracking of moving object based on double layer features optimization", Journal of Optoelectronics•laser, vol. 26, no. 1, **(2015)**, pp. 162-169.
[8] Z. Chunrong, "Bayesian Networks for Knowledge Discovery in Large Medical Data Set", Microelectronics & Computer, vol. 25, no. 6, **(2013)**, pp. 112-115.
[9] X. Jianjun, "A Study of Artificial Neural Networks in Medical Imaging Data Mining", Journal of Practical Radiology, vol. 22, no. 11, **(2014)**, pp. 1416-1418.
[10] Y. Haibin and L. Suyun, "Intelligent Integration and Application of Traditional Chinese Medicine Clinical Medical Research Information-sharing Systems", Journal of Huazhong University of Science And Technology. World Science and Technology: Modernization of Traditional Chinese Medicine, vol. 15, no. 6, **(2013)**, pp. 1480-1482.

Authors



Beibei Dong, The author is now working in Hebei North University as a teacher. She is engaged in signals and network in several areas, especially in medical.

Yu Liu, A lecturer in the School of Information Science and Engineering, Hebei North University, China. Her research interests are in the field of medical information and internet of things.

Benzhen Guo, A lecturer in the School of Information Science and Engineering, Hebei North University, China. His research interests are in the field of medical information and internet of things.

Xiao Zhang, A Professor in the School of Information Science and Engineering, Hebei North University, China. His research interests are in the field of medical information and internet of things.