International Journal of Hybrid Information Technology
Vol.8, No.10 (2015), pp.63-72
http://dx.doi.org/10.14257/ijhit.2015.8.10.06

The Outlier Detection Algorithm Based on Cumulative Holoentropy in Clustering Subspace

Zhang Zhong-ping[1,2], Sun Ying[1], Fang Chun-zhen[1] and Wang Ying[1]

[1]*School of Information Science and Engineering, Yanshan University, Qinhuangdao, Hebei 066004, China*
[2]*The Key Laboratory for Computer Virtual Technology and System Integration of Hebei Province, Qinhuangdao, Hebei 066004, China)*
zpzhang@ysu.edu.cn, sunyingdear@163.com, 1021613517@qq.com

Abstract

Subspace outlier mining has a very important significance in big data analysis. To a large extent, subspace clustering algorithm has impact on the efficiency of mining outliers in subspaces. To solve the problem that CMI method selects best clustering subspaces unstably and complexly, formulas of chain rule of Cumulative Entropy, Cumulative Total Correlation and Cumulative Holoentropy were given. Cumulative Holoentropy was used to mine the best clustering subspaces on continuous data sets in which outliers were detected. Subspace outlier detection algorithm based on Cumulative Holoentropy was then proposed. Finally, the validity and scalability of proposed method were tested on real datasets and virtual datasets. Experiment shows that the efficiency of mining outliers in subspaces is enhanced by the proposed algorithm.

Keywords: *Big Data Analysis, Outlier Detection, Subspace Clustering, Cumulative holoentropy*

1. Introduction

In recent years, the large data is widely applied to biological information research, network intrusion detection, bank fraud analysis, medical service, finance and other fields. However, many data in the practical application are high dimensional data and this data in the high dimensional space, whose characteristic is uniformly distributed, is difficult to distinguish between normal data and abnormal data, leading to hides lots of abnormal data and valuable information. At present, scholars have done a lot to explore for the outlier detection.

The traditional methods of outlier detection can be divided into the following four categories: (1) the outlier detection based on distance; (2) the outlier detection based on local density; (3) the outlier detection based on the deviation; (4) the method based on statistics.

Here, Subspace Outlier Detection Based on Cumulative Holoentropy was proposed, This algorithm uses the k-means method to cause the subspace of CMI values associateed with dimension order for CMI methods, leading to optimal subspace clustering happens to change in the same data sets and the situation of computational complexity has been improved. Using the Cumulative Holoentropy(CH) as an indicator to select the best subspace clustering when looking for a subspace, there will improve the efficiency mining of subspace outliers.

2. Basic Definitions

Definition 1 Entropy [8] is a thermodynamic function which is disorder degree of microscopic particles in system. Symbol is S, expressed as $S = k \ln \Omega$.

ISSN: 1738-9968 IJHIT

Definition 2 Information Entropy [9] is a measure of the average random variable of uncertainty.

Definition 3 Entropy $H(X)$ of discrete random variable X is defined as [11]

$$H(X) = -\sum_{x \in \chi} p(x) \log p(x) \tag{1}$$

Definition 4 Cumulative Entropy[6] $h_{CE}(X)$ of a continuous random variable X is defined as $h_{CE}(X) = -\int_{dom(X)} P(X \le x) \log P(X \le x) dx$ (2)

Definition 5 Condition Cumulative Entropy[6] of any continuous random variables X when random vector $V \in R^B$ (B is positive integer) selected v is defined as

$$h_{CE}(X \mid v) = -\int_{dom(X)} P(X \le x \mid v) \log P(X \le x \mid v) dx \tag{3}$$

Theorem 1 Entropy of the chain rule [11]

If random variable $X_1, X_2, \cdots, X_n$ obey $p(x_1, x_2, \cdots, x_n)$, and then

$$H(X_1, X_2, \cdots, X_n) = \sum_{i=1}^{n} H(X_i \mid X_{i-1}, \cdots, X_1) \tag{4}$$

Definition 6 The total correlation [12] relationship between a set of random variables is defined as

$$C(X_1, X_2, \cdots, X_n) = \sum_{i=1}^{n} H(X_i) - H(X_1, X_2, \cdots, X_n) \tag{5}$$

Definition 7 Holoentropy [5]($HL(V)$) of a random variable is defined as sum of entropy of random vectors and its total correlation. It is

$$HL(X_1, X_2, \cdots, X_n) = H(X_1, X_2, \cdots, X_n) + C(X_1, X_2, \cdots, X_n) = \sum_{i=1}^{m} H(X_i) \tag{6}$$

Here, $V = X_1, X_2, \cdots, X_n$, when $X_1, X_2, \cdots, X_n$ are independent each other, so

$$HL(X_1, X_2, \cdots, X_n) = H(X_1, X_2, \cdots, X_n) = \sum_{i=1}^{n} H(X_i)$$

Get three inferences based on definitions and theorems, as follows:

Corollary 1 cumulative entropy of Chain rule

$$h_{CE}(X_1, X_2, \cdots, X_n) = \sum_{i=1}^{n} h_{CE}(X_i \mid X_1, X_2, \cdots, X_{i-1}) \tag{7}$$

Corollary 2 Cumulative Total Correlation

$$c_{CE}(X_1, X_2, \cdots, X_n) = \sum_{i=1}^{n} h_{CE}(X_i) - \sum_{i=1}^{n} h_{CE}(X_i \mid X_1, \cdots, X_{i-1}) = \sum_{i=1}^{n} h_{CE}(X_i) - h_{CE}(X_1, X_2, \cdots, X_n)$$

Corollary 3 $c_{CE}(\mathrm{X}_1, \mathrm{X}_2, \cdots, \mathrm{X}_n) = \sum_{i=1}^{n} h_{CE}(\mathrm{X}_i) - h_{CE}(\mathrm{X}_1, \mathrm{X}_2, \cdots, \mathrm{X}_n)$

3. Holoentropy Measure Subspace Clustering

The given data set DB (N, D), N indicates the number of records in the data, D indicates the number of attributes in a data set. Each record can be expressed as (x_1, x_2,... , x_D), the total space of data set can be expressed as F={X_1, X_2,…, X_D}, the subspace of data set can be expressed as S={ X_1,X_2,⋯, X_d}, here 1≤ d≤ D.

The definition of information entropy shows that the information entropy of attribute set of data set can describe uniform degree of distribution for data of the data set in total space, information entropy of attribute subset of data set can measure uniform degree of distribution for data of the data set in attribute subspace. It is in the subspace clustering situation. The information entropy of Attribute set (or attribute subset) is smaller, to explain the data is better in the total space (or subspace clustering) clustering situation, otherwise the clustering situation is worse.

Literature[5] shows that it is not easy to determine ultimate effect which have the same information entropy of the subspace outliers mining when two or more sub-space information entropy are the same, That is, whether to more accurately to detect outliers, so using holoentropy to judge subspace clustering situation. Holoentropy is smaller, subspace data distribution is more compact, formation of clusters is better; otherwise data distribution of subspace is more uniform, formation of clusters is worse.

Holoentropy calculated subspace clustering correctness and superiority as follows:

Table 1 is a three-dimensional data set, all the entropy of its two-dimensional attributes subset is used to measure distribution degree of two-dimensional subspace data.

Table 1. Discrete Data Sets

	X_1	X_2	X_3
1	a_1	a_2	d_3
2	a_1	a_2	d_3
3	a_1	a_2	d_3
4	a_1	a_2	e_3
5	a_1	b_2	f_3
6	b_1	c_2	g_3

All two-dimensional subspace holoentropy for this discrete data set is calculated as follows:

In the subspace S(X_1, X_2), Its values is a_1, b_1 for attribute X_1. Its values is a_2, b_2, c_2. So the subspace S(X_1, X_2) holoentropy is calculated as follows:

$HL(X_1, X_2)=H(X_1, X_2)+C(X_1, X_2)=H(X_1)+H(X_2)=1.902$

Similarly, $HL(X_1, X_3)= H(X_1, X_3)+C(X_1, X_3)=H(X_1)+H(X_3)=2.4425$;

$HL(X_2, X_3)= H(X_2, X_3)+C(X_2, X_3)=H(X_2)+H(X_3)=3.0441$

So $HL(X_1, X_2) < HL(X_1, X_3) < HL(X_2, X_3)$, subspace S($X_1$, X_2) clustering is the best, S(X_1, X_3) clustering followed, S(X_2, X_3) clustering is the worst. The distribution of data from attribute subset (X_1, X_2), (X_1, X_3) and (X_2, X_3), so with the holoentropy measure data clustering in the subspace is correct.

Here use an example to illustrate the advantages of using whole entropy measure subspace clustering. This case is the 95th character sequence (three-dimensional, 178 records) in the Character Trajectories discrete data set of UCI dataset Character Trajectories discrete data sets. The 95th characters sequence scattergram tracks is shown in Figure 1, (x represents X-axis tracks of the 95th characters sequence, y represents Y-axis tracks of the 95th characters sequence, ptf represents tip velocity of the 95th character sequence on the point(X,Y)).

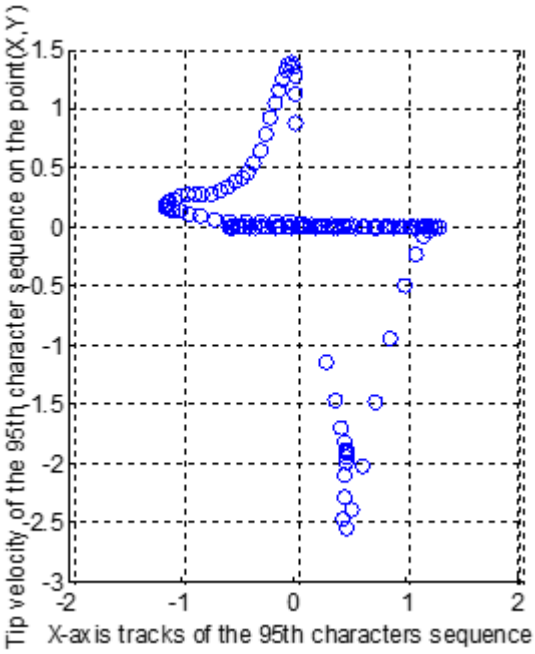

Figure 1(a). Subspace S(x, ptf) H(x, ptf)=6.3554, HL(x, ptf)=11.0334

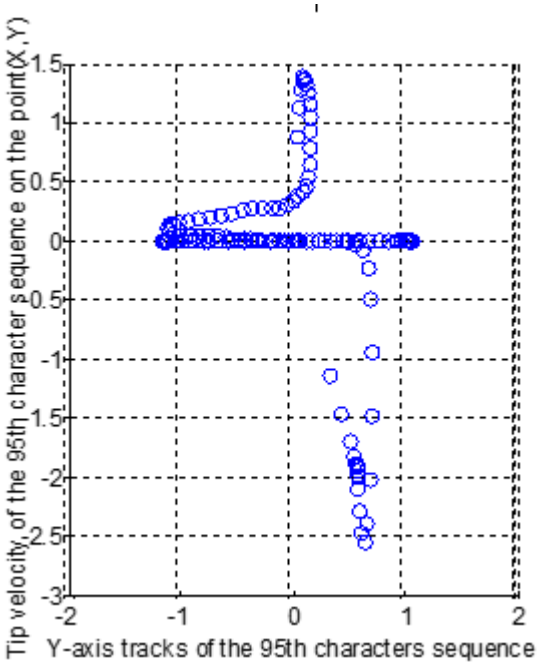

Figure 1(b). Subspace S(y, ptf) H(y, ptf)=0.4756, HL(y,ptf)=11.0334

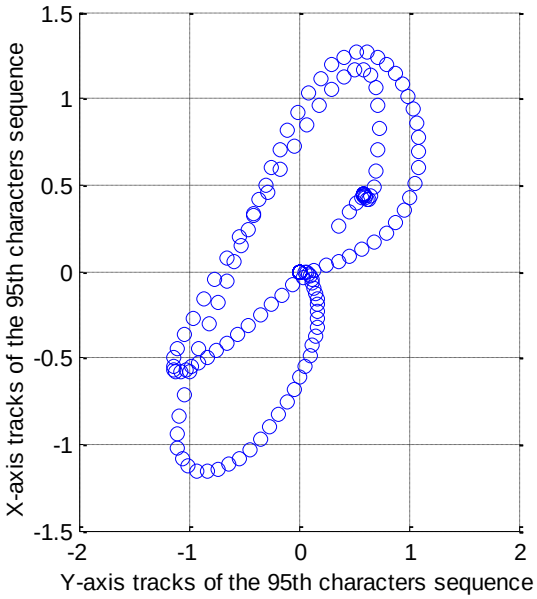

Figure 1(c). Subspace S(x, y) H(x, y)=6.3554, HL(x, y)=12.7108

The 95th character sequence track data is a three-dimensional data set (x, y, ptf) in the data set, information entropy of each dimension is H(x)=6.3554, H(y)=6.3554, H(ptf)=4.6780. Its three two-dimensional subspace S(x, ptf), S(y, ptf),S(x, y), information entropy of S(x, y) is H(x, ptf)=H(x)+H(ptf|x)=6.3554, H(y, ptf)=H(y)+H(ptf|y)=0.4756, H(x, y)=H(x)+H(y|x)=6.3554 . It can be seen if only considering the information entropy of subspace S(x, ptf) and subspace S (x, y), have H(x, ptf)= H(x, y)=6.3554, so cluster situation of the two subspace is the same, but compare Figure 1(a) and Figure 1(c), can visually see Figure 1(a) clustering situation is better.

Holoentropy in the subspace S(x, ptf) and S(x, y) respectively is HL(x, ptf)=H(x)+H(ptf)=11.0334, HL(x, y)=H(x)+H(y)=12.7108, HL(x, ptf)<HL(x, y), Description of subspaces S (x, PTF) clustering better than the subspace S (x, y), Therefore, holoentropy of subspace values can describe clustering situation of subspace well.

For discrete data sets, using a discrete random variable holoentropy to calculating subspace clustering situation; if it is a continuous data set, we use cumulative holoentropy to calculateng subspace clustering situation. According to this idea, subspace outlier detection based on cumulative holoentropy(SODCH) was proposed.

4. SODCH algorithm

4.1 Descripted of SODCH Algorithm

SODCH algorithm for CMI method uses k-means clustering method in data preprocessing phase, resulting in the total subspace clustering on the same data set changes and the emergence of the computational complexity of the problem has been improved.

(1) SODCH algorithm inspired to start a two-dimensional subspace begin using cumulative holoentropy formula (formula (9)) is calculated for each CH value of the two-dimensional (current dimension d) subspace within search range CH value in ascending order; Setting the width of the beam search to limit the number of obtained two-dimensional (current dimension d) subspace, if the current number of two-dimensional (current dimension d) subspace exceeds beam search width, then the excess subspace will be removed to obtain the total two-dimensional (current dimension d) subspace clustering;

(2) Subsequently, the obtained two-dimensional (current dimension d) subspace combined pairwise combinations as required to get the candidate item of three-dimensional (the

current dimension (d + 1)) subspace, at this time, re-use the cumulative holoentropy form ula (formula (9)) calculated for each CH value of the three-dimensional (the current dime nsion (d + 1)) subspace, performs (1), and get the total three-dimensional (the current dim ension (d + 1)) subspace clustering;

(3) Then, simplify the total subspace clustering which have the presence of containing and contained relations, remove the included subspace. Similarly, on this basis, get all the total subspace circularly that the dimensions less than the total space dimension;

(4) Finally, by using the LOF method of outlier mining in all the total subspace clusteri ng, output the degree of the outlier results for each record.

4.2 Process of SODCH Algorithm

Algorithm 1 Subspace Outlier Detection Based on Cumulative Holoentropy
Input: Continuous data collection DB(n, m), Cumulative Holoentropy(CH) Upvalue, Beam search width t;
Output: outlier degree of each record in the DB.
SODCH(DB,Upvalue,t)
BEGIN
(1) ReadData(DB); //Read each record in the data set(DB), data set(DB) has n records, m attributes;
(2) For i from 1 to m; // reach entropy of each column with circule;
(3) Counted $h_{CE}(Xi)$;
(4) End For;
(5) Calculate cumulative holoentropy of each subspace;
(6) Calculate cumulative holoentropy CH(S) of current d-dimensional subspace S;
(7) If CH(S) less than Upvalue;
(8) Reserve S;
(9) End;
(10) With the combination of all the current d-dimensional subspace S get subspace S + 1 of (d+1) dimensional ; / / get the (d + 1) dimensional subspace set of candidates;
(11) End For;
(12) Return all the total subspace clustering;
(13) Each record in for dataset;
(14) Carry out outlier detection using LOF method in all the total subspace clustering of obtained;
(15) End;
(16) Output the outlier degree of each record in dataset;
(17) End;

5. The Experimental Results and Analysis

Experimental operation of the machine configuration: Intel(R) Pentium(R) CPU G2020 2.9GHz,2G memory,500G HDD, Operating System Windows XP.

5.1 Real Data Sets

Real data sets in the literature[6] include :
(1) Ann_thyroid data set(6-dimensional +1-dimensional, 3772 records);
(2) Glass data set (7-dimensional +1-dimensional, 214 records);
(3) Diabetes data set (8-dimensional +1-dimensional, 768 records);
(4) Lymphonorm data set (18-dimensional +1-dimensional, 148 records);
(5) Ion data set (32-dimensional +1-dimensional, 351 records);
(6)Segmentnorm data set (19-dimensional +1-dimensional, 2013 records).

In order to verify the effectiveness of the algorithm, In the above-described real data sets, compared SODCH algorithms with CMI methods in running total time of select the

best subspace clustering and excavate outlier, as is shown in Figure 2, SODCH runs total time of select the best subspace clustering and excavate outlier better than CMI. Figure 3 shows that excavate outlier effects of SODCH and CMI are essentially the same.

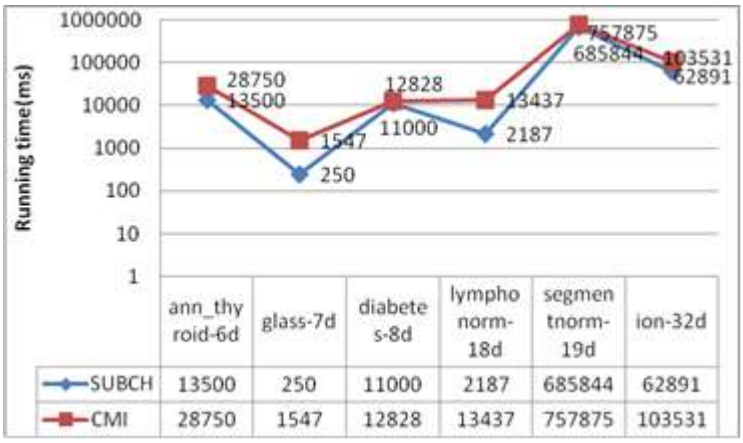

Figure 2. Run Total Time of Selecting the best Total Subspace Clustering and Outlier Detection for SODCH and CMI

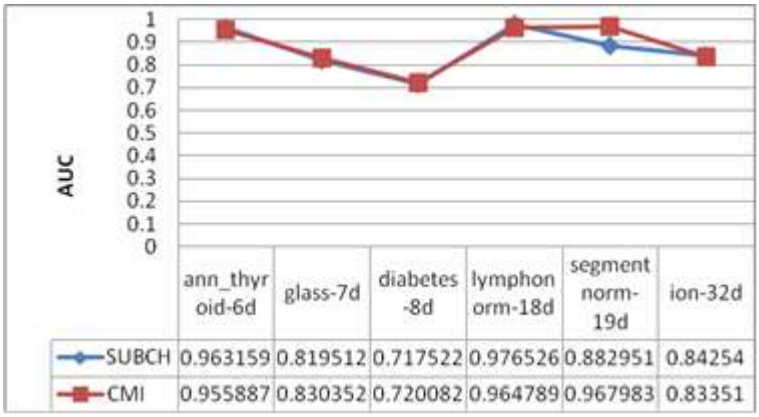

Figure 3. AUC of Outlier Detection for SODCH and CMI on the Total Cluster Subspace

5.2 Virtual Data Sets

The four virtual data set in the literature[6] :

D20_5120 data set (20 dimensional, 5120 records);

D40-5120 data set (40 dimensional, 5120 records);

D80_5120 data set (80 dimensional, 5120 records);

D120_5120 data set (120 dimensional, 5120 records).

To validate the SODCH and CMI for excavate, the total clustering subspace is effective and scalable in the above 4 virtual data set in Figure 4. Operational efficiency of SODCH is better than CMI in excavate the total subspace clustering.

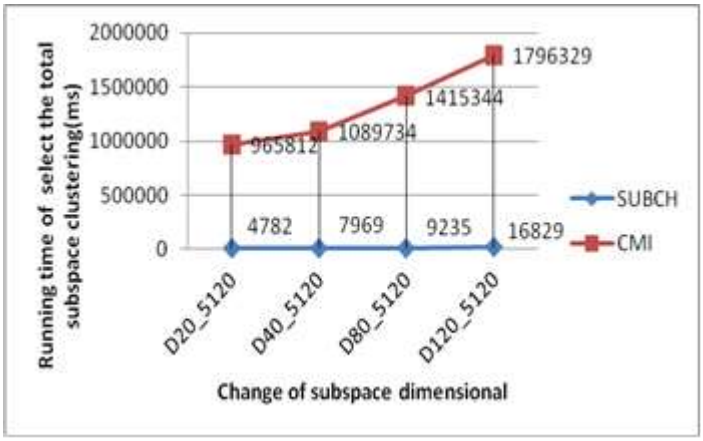

Figure 4. SODCH and CMI Excavates Subspace Clustering Optimal Effectivity and Scalability

As is shown in Figure 2, Figure 3 and Figure 4: SODCH behavior better than CMI in selecting the total subspace clustering, in the outlier detection accuracy of SODCH and CMI is basically the same in subspace clustering. With increasing number of dimensions, SODCH algorithm has good scalability when selecting the total subspace clustering, subspace excavating of high-dimensional data set is very suitable.

ACKNOWLEDGEMENTS

This work was financially supported by Hebei Provincial Natural Science Foundation of China (F2012203087), National Natural Science Foundation of China (61272124), and National Natural Science Foundation of China (61073060).

References

[1] J. W. Han and M. Kamber, Translator: F. Ming and M. Xiaofeng, "Data Mining: Concepts and Techniques", Mechanical Industry Press, Beijing, **(2007)**.
[2] A. W. Mohemmed, M. Zhang and W. N. Browne, "Particle swarm optimization for outlier detection", Proceedings of the 12th annual conference on Genetic and evolutionary computation, ACM, **(2010)**, pp. 83-84.
[3] W. Meijing and Y. Dongyi, "Improved PSO-based algorithm for outlier detection", Journal of Computer Applications, vol. 23, **(2012)**, pp. 139-143.
[4] L. Jinjiang, Z. Caiming and F. Hui, "Point cloud denoising algorithm based on swarm intelligent", Computer Integrated Manufacturing Systems, vol. 17, no. 5, **(2002)**, pp. 935-945.
[5] W. Shu and W. Shengrui, "Information-Theoretic Outlier Detection for Large-Scale Categorical Data", IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 3, **(2013)**, pp. 589-602.
[6] H. V. Nguyen, E. Müller and J. Vreeken, "CMI: An information-theoretic contrast measure for enhancing subspace cluster and outlier detection", SDM, **(2013)**, pp. 198-206.
[7] W. Lee and D. Xiang, "Information-theoretic measures for anomaly detection", IEEE Symposium on Security and Privacy, IEEE, **(2001)**, pp. 130-143.
[8] "Inorganic Chemistry Department of Beijing Normal University", Inorganic Chemistry Department of Huazhong Normal University, Inorganic Chemistry Department of Nanjing Normal University, Inorganic Chemistry, Higher Education Press, Beijing, **(2002)**.
[9] C. E. Shannon, "A mathematical theory of communication", ACM SIGMOBILE Mobile Computing and Communications Review, vol. 5, no. 1, **(2001)**, pp. 3-55.
[10] A. D. Crescenzo and M. Longobardi, "On cumulative entropies", Journal of Statistical Planning and Inference, vol. 139, no. 12, **(2009)**, pp. 4072-4087.
[11] T. M. Cover and J. A. Thomas, Translator: R. Jishou and Z. Hua, "Elements of Information Theory", Mechanical Industry Press, Beijing, **(2005)**.
[12] S. Srinivasa, "A Review on Multivariate Mutual Information", Univ. of Notre Dame, Notre Dame, Indiana, vol. 2, **(2005)**, pp. 1-6.

Authors

Zhongping Zhang, Male, Born in 1972, professor, Ph.D., post-doctoral, CCF Senior Member (E20-0006458S). His main research interests are the grid computing, data mining and semi-structured data *etc*. He has undertaken 1 project of provincial level and participated in 2 projects funded by national natural science foundation of China. He rewarded the provincial Scientific and Technological Progress second-class Award. On the domestic and international academic conferences and journals, He published more than 80 papers, 15 of them are cited by EI.

Ying Sun, Female, Born in 1989, Current Master Students, the main research interest is data mining and the outlier detection.

Chunzhen Fang, Female, Born in 1987, Current Master Students, the main research interest is the outlier detection in data mining.

Ying Wang, Female, Born in 1980, associate professor, Ph.D., the main research interest is Business process management