ICIC Express Letters
Volume **10**, Number **7**, July **2016**

ICIC International ©2016 ISSN 1881-803X
pp. 1755–1761

THE KNOWLEDGE DISCOVERY OF NEAR-INFRARED SPECTROSCOPY BASED ON ATTRIBUTE PARTIAL ORDERED STRUCTURE DIAGRAM

Jialin Song[1], Boni Wang[1], Wenxue Hong[1,2,*] and Shaoxiong Li[1]

[1]School of Electrical Engineering
Yanshan University
No. 438, West of Hebei Avenue, Qinhuangdao 066004, P. R. China
*Corresponding author: hongwx@ysu.edu.cn

[2]Big Data Visualization Technology Center
Northeastern University at Qinhuangdao
No. 143, Taishan Road, Eco. and Tech. Development Zone, Qinhuangdao 066004, P. R. China

Received December 2015; accepted March 2016

Abstract. *Big data brings new challenges to knowledge discovery. The value of the big data lies in the complex structure of large data networks based on data correlation between established inherence in knowledge. In this paper, under the philosophical thought of human beings' cognition, attribute partial ordered structure diagram is applied to knowledge discovery. First, radar charts are used to express 80 samples of near-infrared (NIR) spectral data. Then, these radar charts' feature information is extracted in order to construct formal context. Finally, attribute partial ordered structure diagram is generated according to its corresponding formal context under its algorithm principle. It is proved that attribute partial ordered structure diagram can clearly show relations between attributes and objects, and attributes and attributes. This kind of diagram has better visual effects. The paper shows that the attribute partial ordered structure diagram is an effective tool for NIR spectral data's knowledge discovery and provides a new method for big data analysis.*

Keywords: Knowledge discovery, Attribute partial ordered structure diagram, Near-infrared spectral data

1. Introduction.

Formal concept analysis was proposed by German Professor Wille in 1982 [1]. It is an effective mathematical tool using formal context to generate concept lattice for data analysis. This theory has been successfully applied to knowledge discovery, data mining, machine learning, software engineering, information retrieval and visualization areas and so on [2]. However, due to complex relations among concepts, there are many cross lines in concept lattice, especially when dealing with large formal context [3]. Hong et al. proposed attribute partial ordered structure diagram which is based on the formal concept analysis [4]. Luan et al. used the theory for social classification and pattern recognition [5]. Liu et al. applied it to knowledge discovery of classical prescriptions [6]. In this paper, attribute partial ordered structure diagram is used for near-infrared (NIR) spectral data analysis and knowledge discovery. This kind of diagram has better visual effects. And it is an effective tool and new method for NIR spectral data's knowledge discovery. First, authors try to express the NIR data by radar chart and extract feature information of the radar charts. Then, formal context is constructed with the feature information. Finally, under the philosophical view of human cognition, the corresponding attribute partial ordered structure diagram is generated for data analysis.

2. **NIR Spectral Data.** When NIR light shines on the hydrogen-containing groups (such as C-H, N-H, S-H, O-H) of the substance, these groups will stretch, swing and do other group movements. So the substance can generate characteristic absorption in near infrared region. In other words, after absorbing or reflecting NIR light, NIR spectroscopy with the information of substance is obtained. Different substance has different molecular structure, composite state and other information, which leads to specific absorption characteristics in NIR spectroscopy. All of these above provide a theoretical basis for the analysis of NIR spectroscopy [8].

In this paper, the NIR spectroscopy data contains 80 samples of corn's NIR spectroscopy. The wavelength ranges from 1100 nm to 2498 nm. And the data point is obtained every 2 nm [11]. That is to say, there are 700 points in each sample. Figure 1 shows the spectrum of the NIR spectral data in this paper.

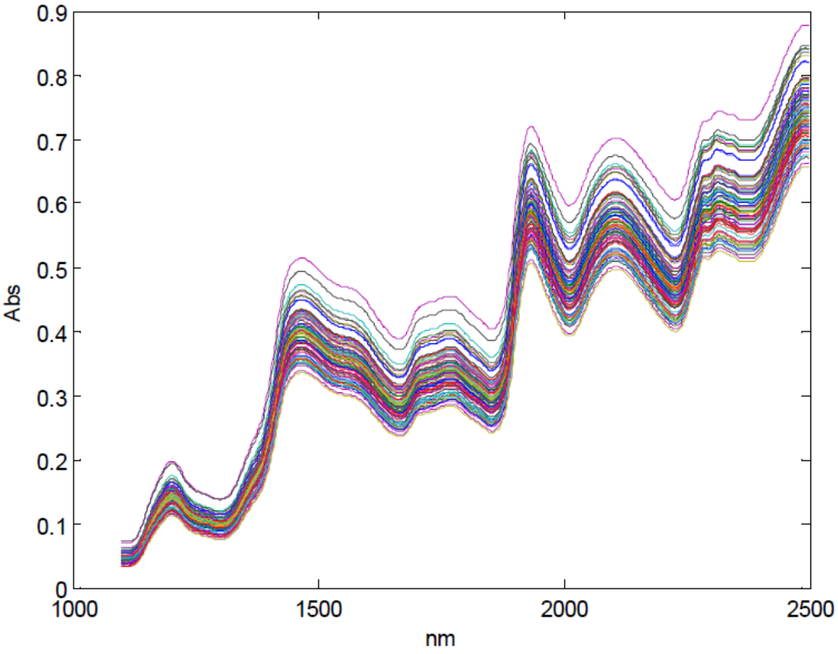


Figure 1. The NIR spectrum of 80 samples

3. **Radar Chart of NIR Spectroscopy and Feature Extraction.** Multi-dimensional radial coordinate chart, also known as radar chart, is one of early multivariate statistical charts. Due to its simple drawing and intuitive expression, it is widely used in data analysis [3]. Classic drawing of a radar chart has 4 steps:

(1) Draw a circle, and then put the circle into p equal parts;

(2) Connect center of the circle to the points on the circumference. Then p radiuses are obtained and define them as p coordinate axis and mark them with appropriate scale;

(3) For a given group data, put the p index value onto the corresponding axis, and then they are connected together to polygon with p sides;

(4) For n groups of data, n polygons with p sides can be got.

Radar chart of the first sample is drawn by Matlab in Figure 2.

High-dimensional data can be expressed in structured way through radar chart, which contains both figure information and features of figures.

Feature extraction is the process of mapping (or transforming) high-dimensional data to low-dimensional data [3]. High-dimensional data can lead to "dimension disaster" and it brings great challenges to data processing. High-dimensional data can be easily submerged useful data in useless data. Therefore, the basic task of feature extraction is

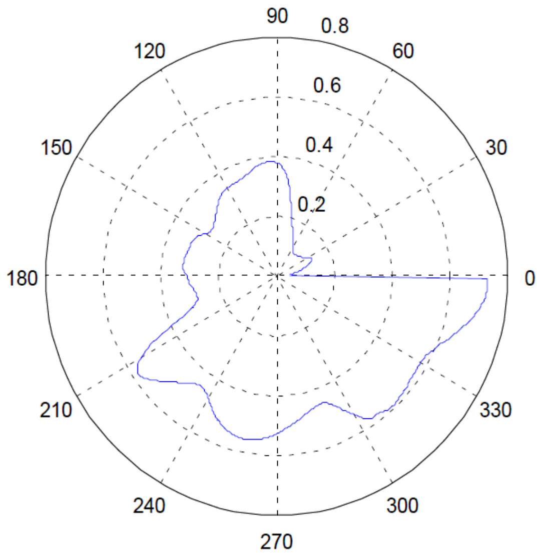


Figure 2. Radar chart of the first sample

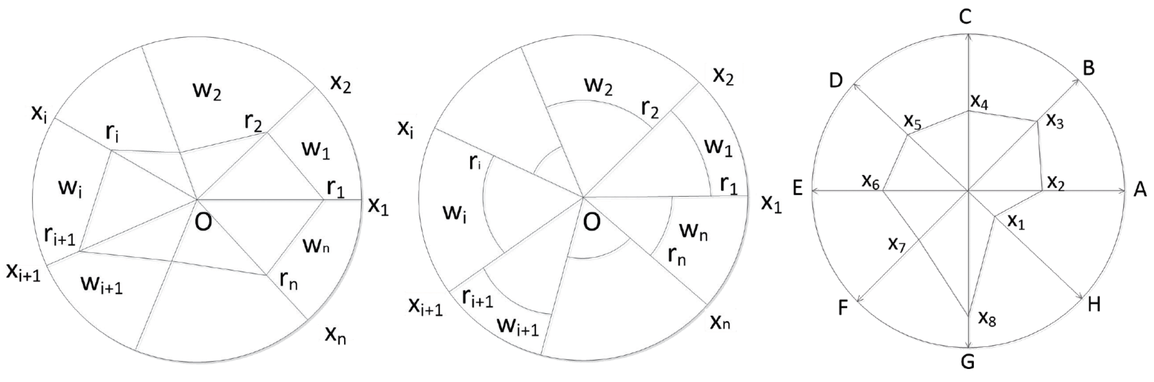
to distinguish those most representative and most effective features from many features and give up the redundant ones.

Multiple chart features include local features and overall features. The former includes the partition area ratio, adjacent amplitude ratio, and symmetry. The latter includes orientation, area, and center of gravity vector and location information. Next, overall features are discussed as follows.

Area feature is the information of evaluating the overall quality of radar chart. It is a function of quantitative information. Contributions of quantitative information to the whole information are different.

Some variables of information are too large or too small and they are not the optimal ones. According to such a situation, the concepts of monotonic function and interval function are established. Monotonic function refers to the ones that the effect of input variables on the overall quality is monotonic. Interval function refers to the ones that the contribution of the input variables to the whole quality is positive within certain range.

For the area features, it can be calculated by the method of triangle area and sector area. As is shown in Figure 3, the related symbols are: the area is S, the ray is r_i the radians is w_i and the dimension is n. The triangular area is called graphical features (the dimension and the original features are the same) and the area of polygon is called total area graphical features (the dimension is 1).



(a) Principle of triangle area (b) Principle of sector area (c) The extraction of radar chart's direction

Figure 3. Principle of triangle area and sector area

(1) The method of monotonic function for calculating area:

$$S = \sum\nolimits_{i=1}^{n} S_i \tag{1}$$

Triangle area:

$$S_i = \frac{1}{2} r_i r_{i+1} \sin w_i \tag{2}$$

Sector area:

$$S_i = w_i \pi r_i^2 \tag{3}$$

(2) The method of interval function for calculating area:

Here the optimum value of quantitative and qualitative information is the center area S_{center}. If the input values are out of this value's interval, ΔS is subtracted from S_{center}. ΔS is produced by $\Delta r_i = r_{center} - r_i$. In this way, the contribution to the whole quality of interval function can be expressed correctly. The method is as follows:

$$S_i = S_{center} - \Delta S \tag{4}$$

Triangle area:

$$\Delta S = \frac{1}{2} |\Delta r_i| r_{i+1} \sin w_i, \quad \Delta r_i = r_{center} - r_i \tag{5}$$

Sector area:

$$\Delta S = w_i \pi |\Delta r_i|^2, \quad \Delta r_{center} - r_i \tag{6}$$

The direction feature of radar chart includes maximum direction, minimum direction and so on. For an n-dimensional radar chart, there are n encoding directions and each direction. So each direction matches with a unique direction. Each n-dimensional radar chart's direction can be represented with a certain symbol of a string of length n. For example, the maximum direction or the minimum direction of a radar chart can be expressed by a symbol.

Figure 3(c) shows the extraction of radar chart's direction. The 8-dimension ($x_1, x_2, \ldots,$ x_8) radar chart's direction feature is ABCDEFGH. So the maximum direction is G and the minimum is H.

In this paper, Matlab 2009b is used to extract the 80 NIR samples features including total area, maximum and its direction, minimum and its direction.

4. **The Formal Context.** Let $K = \{U, M, I\}$ be a formal context, where U represents a set of objects, M represents a set of attributes and I is the relationship between U and M. Usually, a formal context can be expressed by a table. The row of the table represents objects and the column represents attributes. If an object has a certain attribute, the crossing cell of the object and attribute can be marked with 1, otherwise marked with 0.

After discerning the features of the radar charts, the formal context of the NIR data is calculated as Table 1. In Table 1, s_1 to s_4 represent 4 equal parts of the total area, M_1 to M_4 represent 4 equal parts of the maximum value, m_1 to m_4 represent 4 equal parts of the minimum value, I_1 to I_3 represent 3 equal parts of maximum direction and i_1 to i_3 represent 3 equal parts of minimum direction.

5. **Attribute Partial Ordered Structure Diagram.** Attribute shows the feature of all kinds of things. Common attribute which exists in all objects shows the common expression. Unique attribute which can distinguish something from others shows the personality. From the perspective of human cognitive model, a hierarchical graph called attribute partial ordered structure diagram which contains attributes and objects can be constructed [7].

In Figure 4, there are two axes in the graph, one represents attributes and the other represents objects. While on the object coordinate, those near the origin of coordinates represent similarities, and those away from the origin represent personalities.

Table 1. Formal context of the NIR data

	s_1	s_2	s_3	s_4	M_1	M_2	M_3	M_4	m_1	m_2	m_3	m_4	I_1	I_2	I_3	i_1	i_2	i_3
1	1	0	0	0	0	1	0	0	0	1	0	0	1	0	0	0	1	0
2	1	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0
3	1	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	1	0
4	1	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	1	0
5	0	1	0	0	0	0	1	0	0	0	1	0	0	1	0	0	1	0
6	0	1	0	0	0	0	1	0	0	0	1	0	0	0	1	0	1	0
7	0	1	0	0	0	0	1	0	0	0	1	0	0	1	0	0	1	0
8	0	1	0	0	0	0	1	0	0	0	1	0	0	1	0	0	1	0
9	1	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	1	0
10	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0
...																		
78	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1
79	0	1	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	1
80	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1

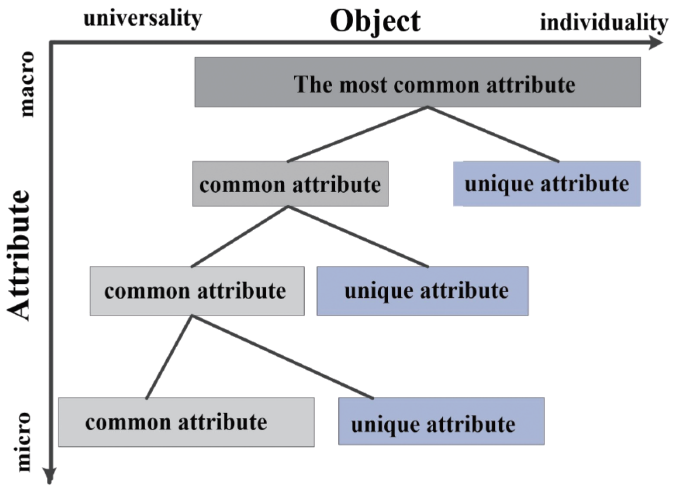
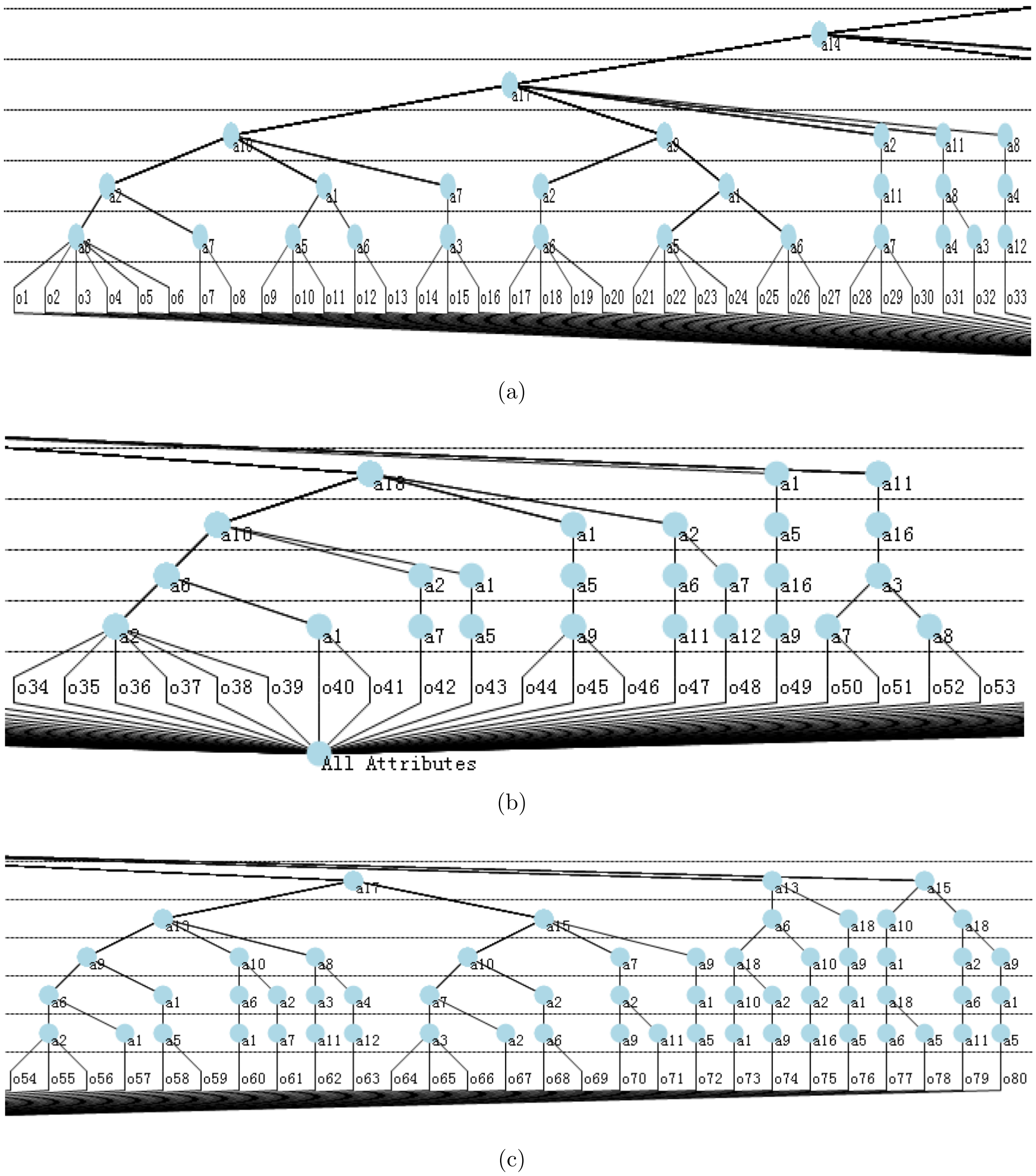


Figure 4. Philosophical principle of knowledge discovery

This kind of diagram puts common things near to the origin and personalized things away from the origin. From the perspective of cognition, relations among attributes show the structure and relations among objects express similarity. It is easier to recognize and distinguish objects from the view of the similarity.

According to the generation method of partial ordered structure diagram, the attribute partial ordered structure diagram of the NIR data is generated as shown in Figure 5.

In Figure 5, it can be easily seen that the samples are clustered to 2 parts. They are the objects under the attribute a_{14} and a_{17}. In the third layer, it can be seen that the objects are clustered into 4 bunches. The first bunch shows objects from o_1 to o_{33} under the attribute a_{17} and the corresponding samples are $\{$*10, 17, 19, 52, 53, 60, 21, 34, 3, 4, 9, 2, 70, 20, 23, 24, 33, 59, 64, 67, 31, 37, 40, 66, 13, 56, 68, 5, 7, 8, 11, 16, 77*$\}$; the second bunch shows the objects from o_{34} to o_{48} under the attribute a_{18} and the corresponding samples are $\{$*48, 51, 71, 76, 78, 80, 49, 50, 72, 74, 42, 54, 55, 46, 47*$\}$; the third bunch shows the objects from o_{54} to o_{63} under the attribute a_{19} and the corresponding samples are $\{$*14, 32, 58, 61, 38, 39, 1, 25*$\}$; the fourth bunch shows the objects from o_{64} to o_{72} under the attribute a_{15} and the corresponding samples are $\{$*18, 22, 35, 63, 62, 69, 57, 6, 36*$\}$.



(a)

(b)

(c)

where: a1:s1, a2:s2, a3:s3, a4:s4, a5:M1, a6:M2, a7:M3, a8:M4, a9:m1, a10:m2, a11:m3, a12:m4, a13:I1, a14:I2, a15:I3, a16:i1, a17:i2, a18:i3; o1:10, o2:17, o3:19, o4:52, o5:53, o6:60, o7:21, o8:34, o9:3, o10:4, o11:9, o12:2, o13:70, o14:20, o15:23, o16:24, o17:33, o18:59, o19:64, o20:67, o21:31, o22:37, o23:40, o24:66, o25:13, o26:56 , o27:68, o28:5, o29:7, o30:8, o31:11, o32:16, o33:77, o34:48, o35:51, o36:71, o37:76, o38:78, o39:80, o40:49, o41:50, o42:72, o43:74, o44:42, o45:54, o46:55, o47:46, o48:47, o49:30, o50:15, o51:29, o52:27, o53:28, o54:14, o55:32, o56:58, o57:61, o58:38, o59:39, o60:1, o61:25, o62:12, o63:75, o64:18, o65:2, o66:35, o67:63, o68:62, o69:69, o70:57, o71:6, o72:36, o73:73, o74:65, o75:26, o76:43, o77:44, o78:45, o79:79, o80:41

Figure 5. Attribute partial ordered structure diagram of the NIR data

Traditional methods such as BP neural network need mounts of training sets to ensure the correctness; K-means must give the number of clusters in advance; support vector machines (SVM)'s visual effects are not that good.

6. **Conclusion.** In this paper, authors apply attribute partial ordered structure diagram to NIR spectral data's knowledge discovery. Radar chart is used to express NIR spectral data. These radar charts' feature information is extracted in order to construct formal context. It is proved that attribute partial ordered structure diagram can clearly show relations between attributes and objects, and attributes and attributes. This kind of diagram has clear structure, no cross lines and good visual effects. The paper shows that attribute partial ordered structure diagram is an effective tool for knowledge discovery of NIR spectral data and provides a new method for big data analysis.

Acknowledgment. This work is partially supported by the National Natural Science Foundation of China (NSFC) under Grant 61273019, 61201111, 81273740 and 81373767. It is also supported by the Natural Science Foundation of Hebei Province under Grant F2013203368. The authors gratefully acknowledge the supports. The authors also gratefully acknowledge the reviewers for their valuable comments and suggestions to the paper.

REFERENCES

[1] R. Wille, *Restructuring Lattice Theory: An Approach Based on Hierarchies of Concepts Ordered Sets*, Reidel, Dordrecht, 1982.

[2] W. Hong, J. Luan, T. Zhang, S. Li, C. Zheng and J. Liu, Complete definition of attribute and object feature in formal context, *Journal of Yanshan University*, no.5, pp.381-387, 2014.

[3] W. Hong and J. Wang, *Pattern Recognition Based on Visualization*, National Defense Industrial Press, 2014.

[4] W. Hong, S. Li, T. Zhang, J. Luan and W. Liu, Generation principle of partial ordered structure towards big data, *Journal of Yanshan University*, no.5, pp.388-393, 2014.

[5] J. Luan, C. Wang, E. Yan, J. Yu, J. Song and W. Hong, The classification of Hsyes-toth dataset based on structural partial-ordered attribute diagram, *ICIC Express Letters*, vol.7, no.3(B), pp.965-970, 2013.

[6] S. Liu, S. Xu, R. Li, S. Li, W. Hong, Z. Zhu and M. Liu, Knowledge discovery of cough treatment pattern of ZHANG Zhong-jing's classical prescriptions based on partial ordered structure theory, *Journal of Yanshan University*, no.5, pp.455-459, 2014.

[7] W. Hong, J. Luan, T. Zhang, S. Li and E. Yan, A new method for knowledge discovery based on partial ordered structure theory, *Journal of Yanshan University*, no.5, pp.394-402, 2014.

[8] J. C. Hirsch, J. R. Charpie, J. G. Gurney and R. G. Ohye, Role of near infrared spectroscopy in pediatric cardiac surgery, *Progress in Pediatric Cardiology*, vol.29, no.2, pp.93-96, 2010.

[9] T. Zhang, H. L. Ren, X. M. Wang, Y. Y. Zhang and W. X. Hong, A calculation of formal concept by attribute topology, *ICIC Express Letters, Part B: Applications*, vol.4, no.3, pp.793-800, 2013.

[10] J. Cui, *Research on Pattern Analysis Methods Based on Multiple Graph Presentation of Traditional Chinese Medicine Fingerprint*, Yanshan University, 2012.

[11] W. Hong, H. Gao, J. Cui, X. Li and S. Ji, Extraction and representation of gragh and feature primitives of multivariate gragh, *Journal of Yanshan University*, no.5, pp.405-411, 2008.

[12] *http://www.eigenvector.com/data/index.htm.*