A New Hazard Identification Method-State Transition Graph

Hu Min
Sydney Institute of Language & Commerce
Shanghai University, SHU
Shanghai, China
minahu@163.com

Wu Fangfang, Zhu Bo, Lu Bo, Pu Jinglei
Sydney Institute of Language & Commerce
Shanghai University, SHU
Shanghai, China
summer_wu0704@163.com

***Abstract*—*It is important and difficult to identify the Hazard before a disaster happen because disaster often happens suddenly. This paper proposes a new method – **State Transition Graph**, which based on visual data space reconstruction, to identify hazard. The change process of the system state movement from one state to another in a certain period is described by some state transition graphs. The system state, which is safe or hazard, could be distinguished by its state transition graphs. This paper conducted experiments on single-dimension and multi-dimension benchmark data to prove the new method is effectiveness. Especially the result of stimulation experiments, based on the Yangtze River tunnel engineering data, showed that state transition graph identifies hazard easily and has better performances than other method. The State transition graph method is worth further researching.*

Keywords- Hazard; Identification; State Transition Graph; Tunnel

I. INTRODUCTION

In recent years, the geological disaster, financial crisis, coalmine explosion and various emergent accidents bring huge human, economic and social losses in the world. Therefore, it's particularly important to identify hazard and take some preventive measures to decrease the losses. Identifying hazard means to find the abnormal information or tracks among massive data. In view of the fact that the disaster is an evolution process, which happens abruptly, so it is difficult for the traditional forecasting methods to find it. Some new perspectives, for example, artificial intelligence and data space reconstruction are introduced to this field. Onoda Takashi [1] adopted support vector machine (SVM) to diagnose the hydropower station abnormalities. Sven S. Groth [2] used SVM for text-mining to find risk in capital market. Cao[3] introduced the chaos theory into the disaster forecasting, also combined the phase space reconstruction theory with neural network to identify the flood long-term signs. Bai [4] used fractal theory to analyze monitoring data curves to find that the curve distortion is related to the pregnant process of earthquake. Ouyang [5] ignored the time dimension and transfer time series date to the non-time series space structure for weather forecast. Although these methods have obtained some achievements, hazard identification is still a difficult problem in the world

Visualization technology has advantages on high-dimension mass data mining. Inspired by this idea and the data space reconstruction concept, this paper proposes a new method for identifying hazard: State transition Graph. The new method provides a new way to forecast hazard based on high-dimension mass data.

This paper is divided into five sections. The first section introduces the concept and principle of the state transition graph. Section 2 describes the drawing steps of state transition graph. Section 3 illustrates the process of graphical feature extraction and hazard identification. Section 4 shows the experiments of benchmark data and engineering stimulation processed and results. And the conclusion and future work are drawn in the last section.

II. THEORETIC CONCEPT OF STATE TRANSITION GRAPH

A. State transition

State refers to a way of being that exists at a particular time, which can describe the nature of the system. It can be represented by a group of physical quantities. State transition refers to the process that the system transits from one state to another state. According to the basic laws of thermodynamics, the system belongs to a certain state when the various properties of the system remain certain numerical value. Otherwise, the system state is changed when the value of properties changes. The properties and the states of the system are related. The first law of thermodynamics also points out that the state transition process means the energy change. According to the energy transfer theory proposed by Haddon [6], the nature of disasters or accidents is the abnormal energy transfer. The hazard can be identified through observing the style and phenomenon of state movement.

The concept of state transition graph is inspired from visualization. It focuses on the changes of the whole system form the aspect of state transition. State transition graph uses a sequence of graphics to describe the route that the system transfers from one state to another in a period of time. The system state of a particular moment will be mapped into a spot or a polygonal line in the graph. The last spot and current spot are connected to embody the state transition trajectory while the data are single dimension, while the data are multi- dimension, connection between the start and end points of the two polygonal lines expresses the state transition trajectory. In this way, state transition graph are generated.

B. Time Window

If all the system state transition trajectories in the whole life cycle are drawn on one picture, the graphs might not reflect the inherent laws clearly when the data size are massive. Therefore the state transition graph method segments an original data set into several data sets according to period and progressive step. Each period will be drawn to one graph to get a series of graphs. Each graph contains the properties of particular stage, and researcher can observe and analyze whether the system state is safe or not through comparing the typical characteristics of graph. If the typical graphical characteristics are computed, the hazard could be identified automatically through classification algorithms.

III. DRAWING OF THE STATE TRANSITION GRAPH

A. Data Preprocessing

Assume there is an M dimensions system and the state variables at the moment K is X_k:

$X_k=\{x_{k,1},x_{k,2},\ldots,x_{k,M}\}$ where $k\geq 1$, $i=1,2,\ldots,M$

Then number of sampled data is N, which is expressed as $\{X_1, X_2,\ldots,X_N\}$

Also define the range of the ith variable is from Min_i and Max_i.

And for $x_{i,k}$ -- variable i at the moment k, its value can be zoom to the range from 0 to 2π according to (1).

$$x'_{i,k}=\frac{x_{i,k}-Min_i}{Max_i-Min_i}*2\pi \qquad (1)$$

B. Graphic Representation of the State Variables

Select one spot in the plane as origin to establish a Cartesian coordinate system.

For $x'_{i,j}$, mapped to point $G_{i,j}(G\_X_{i,j}, G\_Y_{i,j},)$ by (2).

$$\begin{cases} G\_X_{i,j}=i*sin(x'_{i,j}) \\ G\_Y_{i,j}=i*cos(x'_{i,j}) \end{cases} \qquad (2)$$

In the same way, state variable X_k at the moment K is mapped into M points, they are $G_{1,k},G_{2,k},\ldots,G_{M,k}$. Then connect these points in order to generate a polygonal line L_k, which can reflect the state variables at moment K. Fig. 1(a) displays a system with one variable(one dimension), and the state of moment K is expressed by $G_{1,k}$. And Fig.1(b) displays a system with five variables (five dimensions), the states of moment K can be described as a polygonal line O-$G_{1,k}$-$G_{2,k}$-$G_{3,k}$-$G_{4,k}$-$G_{5,k}$, so N sampled data can get N polygonal lines.

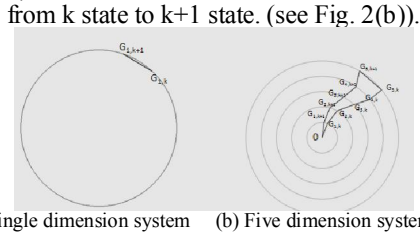
(a) Single dimension system (b) Five dimension system

Figure 1. System state at moment K

C. Graphical Representation of the State transition Process

For a single dimension system, if $G_{1,k}$ is the state of moment K, and $G_{1,k+1}$ is the state of moment K+1, the connection between the two spots to express the change of the system transition process from k state to k+1 state. (see Fig.2(a)).

For a single dimensions system, if O-$G_{1,k}$-$G_{2,k}$-$G_{3,k}$-$G_{4,k}$-$G_{5,k}$ is the state of moment K, and if O-$G_{1,k+1}$-$G_{2,k+1}$-$G_{3,k+1}$-$G_{4,k+1}$-$G_{5,k+1}$ is the state of moment K+1, A lines between $G_{5,k}$ and $G_{5,k+1}$ expresses the trajectory of the system transferring from k state to k+1 state. (see Fig. 2(b)).

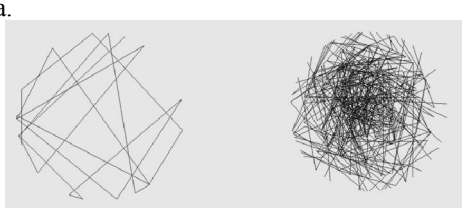


(a) Single dimension system (b) Five dimension system

Figure 2. System state transition from moment K to moment K+1

Thus, the trajectory of state transition can be reflected by graph. Assumed the period of data segment in one graph is T, data series $\{X_k,X_{k+1},\ldots, X_{k+T}\}$ can be used to generate a state transition graph.

Fig. 3(a) is the state transition graph, which includes 20 one dimension time-series data. And Fig. 3(b) is the state transition graph, which includes 100 five dimensions time-series data.



(a) Single dimension system (b) Five dimension system

Figure 3. System state transition graphs

D. Drawing a Sequence of the State transition graph

Different series state transition graphs can be drawn as cycle time T. The first data series in the first graph is $\{X_1,X_2,\ldots, X_T\}$, the second data series is $\{X_D,X_{D+1},\ldots, X_{T+D}\}$, and the N data series is $\{X_{D*(N-1)},X_{D*(N-1)+1},\ldots, X_{T+D*(N-1)}\}$. D is the progressive step and it is an integer between 0 and T. And when D equals to T, the two adjacent state transition graphs have no coincide in time. That is, a series of graphs should be drawn for a long period, which represents different properties of the system state in different period.

E. Feature Extraction of the State transition Graph

State transition graph is black and white two-color graph, which are easily to separate the background and the foreground. The shape and texture features are the main low-level vision characteristic. Therefore, they should be considered with priority during feature extraction. 15 common main characteristics were chosen first (table I). In order to analyze the graph space distribution characteristics, each graph is divided into four regions equally (see Fig. 4).

There are 15 features in the each region, and then 60 features are obtained in total finally.

TABLE I. FEATURE OF THE STATE TRANSITION GRAPH

Class	Name	Explanation
Internal Shape Features	EulerNumber	The number of objects in the region minus the number of holes in those objects
	Major AxisLength	The length (in pixels) of the major axis of the ellipse that has the same normalized second central moments as the region
	Minor AxisLength	The length (in pixels) of the minor axis of the ellipse that has the same normalized second central moments as the region
	Eccentricity	The eccentricity of the ellipse that has the same second-moments as the region. It is the ratio of the distance between the foci of the ellipse and its major axis length.
	Orientation	The angle (ranging from -90 to 90 degrees) between the x-axis & the major axis of the ellipse that has the same second-moments as the region
External Shape Features	Centroid	The center of mass of the region (including both the horizontal & the vertical coordinate
	Area	The number of black pixels in the region
	Density	The ratio between black and total pixels
	Perimeter	The distance around the boundary of each contiguous region in the image
Texture Features	Entropy	Entropy of grayscale image, a statistical measure of randomness that can be used to characterize the texture of the input image.
	deviation	Image matrix standard deviation, represents the color dispersion degree
	Contrast	Color contrast which describes the image definition and texture rills
	Correlation	Computes the cross-correlation of the GLCM elements with analogous columns or rows
	Energy	Sum of the GLCM elements squares, reflects the image gray distribution level & texture coarseness
	Homogeneity	Reflects the homogeneity of the image texture, measures the local change of texture

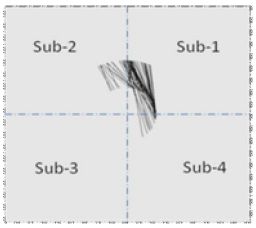


Figure 4. Division of the sub-state transition grpahs

F. Hazard identification based on Classification Algorithm

Each graph can be defined as an 60 dimension feature vector P= $\{p_1,p_2,\ldots,p_{60}\}$ basing on the feature extraction. The issue of hazard identification turns out to be a classification issue, which classify these graphs into two categories: safety or hazard. The work on hazard identification can be conducted through certain classification algorithm.

IV. EXPERIMENT AND EVALUATION

In this section, this paper provides three different dimension time-series benchmark datasets to validate the effectiveness of the state transition graph.

A. Synthetic Control Chart Pattern Recognition

1) Data Information

The first experiment used the dataset which contains 600 examples of control charts synthetically generated by the process in Alcock and Manolopoulos [7]. These data are useful to analyze whether the current manufacture process is normal and abnormal. There are six different classes of control charts: normal, cyclic, increasing trend, decreasing trend, upwards shift and downward shift. And each condition contains 100 examples.

2) Experiment Process and Results

This paper used all examples of "normal" and "decreasing trend" classes as experiment data, which are simplified into a two-category classification problem. That is, the normal ones are similar to the safe data and the abnormal ones are similar to hazard data. The experiment was divided into two conditions: normal and abnormal, and each class had 100 examples. Each example has 60 time-series data and can draw one state transition graph. Fig. 5 shows part of the state transition graphs, and Fig. 5(a)-Fig. 5(e) are normal and Fig. 5(f)-Fig. 5(j) are abnormal.

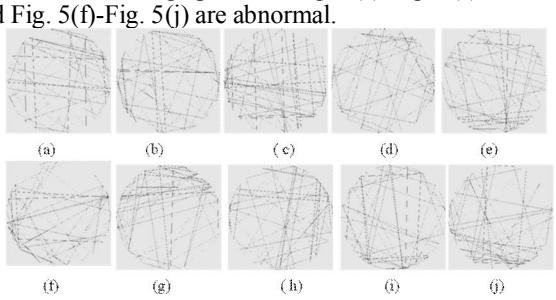


Figure 5. State transition graphs of synthetic control dataset

After the features of all graphs were computed, select certain proportion data randomly for training the decision tree classification model, and the rest data was used for testing. The classification ability of the model is showed in table II. The false positive rate is the proportion of absent abnormal result, while the false negative rate is the proportion of present abnormal result that actual outcomes are normal.

TABLE II. ACCURACY OF THE SYNTHETIC CONTROL DATASET

Training set proportion	Accuracy	False positive rate	False negative rate
10%	25.49 %	73.80%	75.22%
20%	58.82 %	26.50%	55.85%
30%	88.24 %	21.11%	2.42%
40%	98.82 %	1.95%	0.41%
50%	100.00 %	0.00%	0.00%

3) Analysis and Evaluation

Fig. 5 shows that there are obvious differences between the normal and abnormal state. In the normal state graphs, the lines are denser and the distribution of black pixels is evenly while the abnormal ones are sparse and distribution is more similar in each direction. Table II shows that the classification accuracy is low when the train set proportion is below 30%. However, the accuracy rises rapidly when the test proportions are larger than 30%. And the accuracy finally achieves 100% when the proportion is 50%. This may relate to the 60 features in the dataset. Ideal classification model may not be created if there is not enough training data and the original features are too much.

B. Ozone Level Detection

1) Data Information

The ozone level detection dataset is the one hour peak environment monitor data provided by Kun Zhang, Wei Fan and Xiaojing Yuan. Those data were collected from 1998 to 2004 at the Houston, Galveston and Brazoria area [8]. There are 2536 days instances in the dataset and each instance has 73 attribute, including temperature and wind speed. These data are used to forecast whether it is ozone day or normal day. Rearranged the data into a two dimension time-series dataset which including two attributes: temperature and wind speed. The sampling interval of this data set is one hour. In this way, a two dimension dataset with 60864 records was formed.

2) Experiment Process and Results

The 60864 records were divided into 2 classes, training set and the testing set. Each state transition graph was drawn with every 72 records (3 days record). The progressive step between two graphs is 24 records, which means previous 3 days' monitoring data were used to forecast the ozone situation of the later day. There were 1616 graphs in all, 192 graphs were normal (see Fig. 6(a)-Fig. 6(e)) and 424 graphs were abnormal (see Fig. 6(f)-Fig. 6(j)).

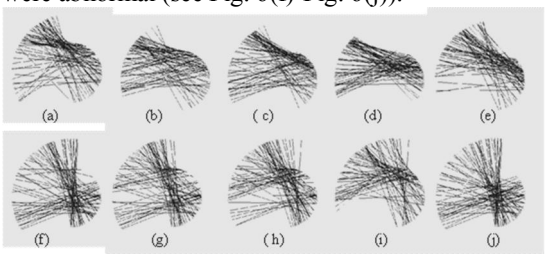


Figure 6. State transition graphs of ozone level detection dataset

The classification accuracy is showed in table III.

TABLE III. ACCURACY OF THE OZONE LEVEL DETECTION DATASET

Training set proportion	Accuracy	False positive rate	False negative rate
5%	60.58 %	39.41%	39.44%
10%	67.37 %	24.80%	40.46%
20%	73.16 %	20.01%	33.67%
25%	86.84 %	9.34%	16.97%
30%	92.11 %	6.10%	9.69%
35%	97.37 %	2.34%	2.92%

3) Analysis and Evaluation

Fig. 6 shows that the area of black pixel is small and the density of lines is small in the normal state transition graphs, while the pattern is like expanding fans with larger area in the abnormal state transition graphs. Table III shows the classification accuracy is low when the training set proportion is smaller than 25%. And it rises after the proportion exceeding 30% and reaches 97.27% at 35%. Comparing to the results of the synthetic control, the accuracy of the ozone level detection rises slower and floats up and down. This may due to the change of the ozone weather is continuous. And this experiment is according to days, so the features of some state transition are not obvious.

C. Electroencephalogram (EEG) Detection

1) Data Information

This data arises from a large study to examine EEG correlates of genetic predisposition to alcoholism by Lester Ingber [9]. It contains measurements from 64 electrodes with 256 records each sampled at 256 Hz. This paper chose the large data set version of the EEG dataset. The training set contains data for 10 alcoholic and 10 control subjects, with 20 runs per subject per paradigm. The test data used the same 10 alcoholic and 10 control subjects as with the training data, but with 30 runs per subject per paradigm.

2) Experiment Process and Results

In order to control the size of the picture, six electrodes F3, F4, T7, T8, P7, P8 were selected for drawing the state transition graph [10]. And each sample (including 256 records) draws one graph. Thus, the model drawn 500 pictures, among which the normal state referred to the EEG of the control subject and the abnormal state referred to the alcoholic ones. Fig. 7 shows part of the state transition graphs, and Fig. 7(a)-Fig. 7(e) are normal and Fig. 7(f)-Fig. 7(j) are abnormal.

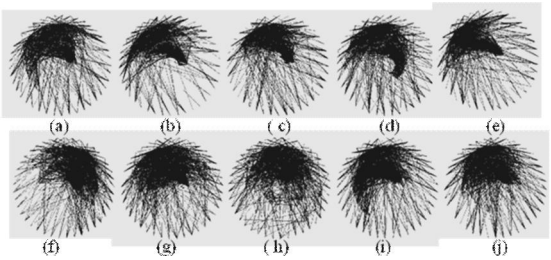


Figure 7. State transition graphs of EEG dataset

The classification accuracy is showed in table IV.

TABLE IV. ACCURACY OF THE EEG DATASET

Training set proportion	Accuracy	False positive rate	False negative rate
5%	22.58 %	81.97%	72.88%
10%	43.16 %	50.89%	62.79%
20%	61.84 %	40.56%	35.75%
30%	82.11 %	11.03%	24.76%
40%	100.00 %	0.00%	0.00%

3) Analysis and Evaluation

Fig. 7 shows that the black high density area of the normal graphs is larger than abnormal. Table IV shows that the classification accuracy is low when the train set proportion is below 20%. However, the accuracy rises rapidly when the proportions are larger than 30%. And the accuracy achieves 100% when the proportion is 40% finally.

V. ENGINEERING STIMULATION

A. Project Background

The Shanghai Yangtze River tunnel is 8.95 km's long and the diameter is 13.7m. Due to its long-distance excavation length, large shield diameter, high depth and complex geology, the risk of this is high. There were two stimulation experiments, one used the earth-pressure monitoring data and another used the five soil pressure value of shield sealed cabin for hazard identification.

B. Single-Dimension Earth Pressure Risk Forecasting

1) Data Process and Results

The earth pressure data were collected from 2008/5/12 12:00 to 2008/5/23 0:00 and sampled at 30 minutes. The time period was 24 hours and the progressive step length was 1 hour. Thus 228 state transition graphs were drawn. The training set includes 432 records before 2008/5/15 0:00 and all of them were non-risk. And the rest were test set and transferred to 170 graphs. Considering the training set only contained safe data, immune neglect selection algorithm was adopted after feature extraction. Fig. 8 shows the state transition graphs from 2008/5/15 00:00 to 2008/5/17 00:00 every 6 hours. Fig. 9 is the forecasting result from 2008/5/15 0:00 and the dashed line is the time that real accident phenomena emerged.

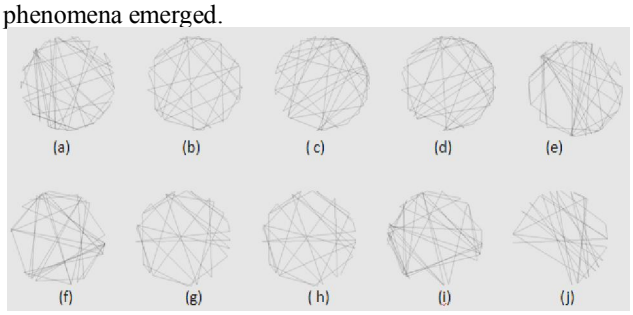


Figure 8. State transition graphs of the earth pressure

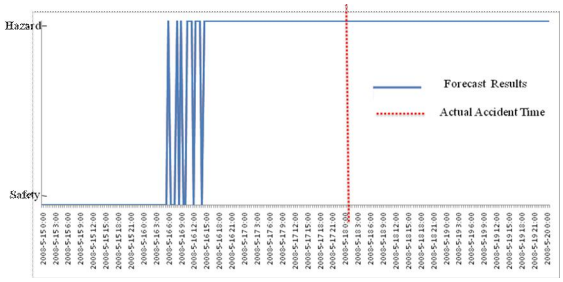


Figure 9. Risk forecasting of the earth pressure

2) Analysis and Evaluation

Fig. 8 shows that during 2008/5/15 00:00 to 2008/5/17 00:00, Fig. 8(a)-(e) are normal with even distribution. But from 2008/5/16 6:00, the lines are spare in picture (f), (g) and (h). And from 2008/5/16 18:00 picture (i) and (j) shows significant asymmetry. Fig. 9 shows that the risk alarm stared at 2008/5/16 6:00 and kept continuous alarming from 008/5/16 18:00. This result was 30 hours in advance than the real risk took place. And it will bring great benefit to do some measures to prevent hazard.

C. Five-Dimension Cabin Pressure Risk Forecasting

1) Data Process and Results

The Yangtze River Tunnel 16040 up-line cabin pressure data were from 2008/5/1 0:02:12 to 2008/5/31 17:42:11. The sampling interval was 3 minutes and the progressive step is one hour. This paper set the 1990 records from 2008/5/1 0:02:12 to 2008/5/5 0:02:18 as training set and the rest were test set. The time period was 12 hours and the progressive step was 2 hours. Thus the model drew 266 state transition graphs. Immune neglect selection algorithm was adopted to do hazard identification. Fig. 10 shows the state transition graphs from 2008/5/15 00:00 to 2008/5/17 00:00 every 6 hours. Fig. 11 is the forecasting result from 2008/5/15 0:00 and the dashed line is the time that real accident phenomena emerged.

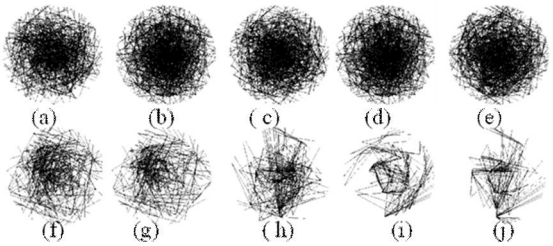


Figure 10. State transition graphs of the cabin pressure

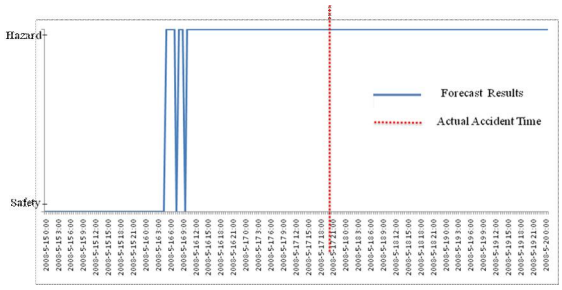


Figure 11. Risk forecasting of the cabin pressure

2) Analysis and Evaluation

Fig. 10 shows that during 2008/5/15 00:00 to 2008/5/17 00:00 this continuous period, the former 4 graphs are normal with even distribution and round shape. But from 2008/5/16 6:00, the graphs show obvious inhomogeneity and irregularity. Fig. 11 shows that the risk alarm stared at 2008/5/16 3:00, which is earlier than the earth pressure result. And it kept continuous alarming from 2008/5/16 9:00, which is 42 hours earlier than the real time risk. This explains that the using shield excavation parameter for hazard identification is quicker than the using the earth pressure data.

VI. CONCLUSION

This paper proposed the state transition graph method, and try g to solve the difficult problem of identifying hazard. The basic concept of the new method is reflecting the transfer trend of the system energy through the trajectory of state movements. Comparing to the traditional data curve, the state transition graph emphasizes on the process of state transition. And it also draws the single or multi-dimension data in one graph, which raises the analysis efficiency of the high frequency sampling monitoring data. Moreover, it has good robustness for missing or abnormal data.

The experiments on various standard dataset proved the effectiveness of the new method. And the stimulation results of the Yangtze River Tunnel showed that the state transition graph method can pre-alarm much earlier than the real risk happened time.

The research of the state transition graph method just starts at the beginning. Thus, the drawing method needs to be improved to better reflect the changes of the state. And it also needs further instances and experiments to validate.

ACKNOWLEDGMENT

This work was supported by grants from the Natural Science Foundation of China (50778109).

REFERENCES.

[1] Takashi Onoda, Norihiko Ito, Yamasaki Hironobu Trouble Condition Sign Discovery based on Support VectorMachines for Hydroelectric Power Plants Proceedings of International Joint Conference on Neural Networks, Atlanta, Georgia, USA, June 14-19, 2009,pp2358-2365

[2] Sven S. Groth, Jan Muntermann, Discovering Intraday Market Risk Exposures in Unstructured Data Sources:The Case of Corporate Disclosures, Proceedings of the 43rd Hawaii International Conference on System Sciences – 2010,pp1-pp10

[3] Cao Lianhai, Cao Bo, Chen Nanxiang, Xu Jianxin. Application of Phase Space Reconstruction and Neural Networking Flood Disaste rLosing Forcasting[J]. Journal of Earth Sciences and Environment, 2006. 28(02):89-92

[4] Bai Chaoying, Zhu Lingren, Wang Haitao, Zhou Shiyong. Significance Of Seismically Active Information Entropy And Time-Space Fractional Dimension In Earthquake Prediction [J]. Earthquake, 1992. 04:7-11

[5] Ouyang Shoucheng etc., Structure Transformation of Irregular Information for Time-sequence and Elaborate Analysis of Evolution [J], Engineering Science, 2005.07(04): 36-41.

[6] Haddon W. Energy damage and the ten countermeasure strategies. J Trauma 1973; 13: 321-31

[7] http://archive.ics.uci.edu/ml/databases/synthetic_control/synthetic_control.data.html

[8] http://archive.ics.uci.edu/ml/machine-learning-databases/ozone/

[9] http://archive.ics.uci.edu/ml/databases/eeg/

[10] L. Ingber. (1997). Statistical mechanics of neocortical interactions: Canonical momenta indicators of electroencephalography. Physical Review E. Volume 55. Number 4. Pages 4578-4593.

[11] Eser Kandogan, Visualizing multi-dimensional clusters, trends, and outliers using star coordinates, Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, p.107-116, August 26-29, 2001, San Francisco, California

[12] Strand S.,Forecasting the future: pitfalls in controlling for uncertainty. Futures, Volume 31, Number 3, April 1999 , pp. 333-350(18)

[13] Jan G. De Gooijer, Rob J. Hyndman,25 years of time series forecasting International Journal of Forecasting Volume 22, issue 3, 2006,pp443 – 47