Visual Data Mining of SARS distribution using Self-Organization Maps

Li Yanli, Hu Bisong, Gong Jianhua
State Key Laboratory of Remote Sensing Science
Institute of Remote Sensing Applications, CAS
Beijing 100101, China
liyl02@gmail.com

Cao wuchun, Fang Lique
State Key Laboratory of Pathogen and Biosecurity
Beijing Institute of Microbiology and Epidemiology
Beijing, 100101, China

Abstract—Background: From a public health perspective, the socioeconomic conditions correlate with occurrences of infectious diseases. Our premise is that the number of SARS patients is a non-linear function of socioeconomic effects that are not normally distributed among regions. The objective was to integrate multivariate data sets representing social and economic factors to evaluate the hypothesis that regions with similar socioeconomic characteristics exhibit similar distributions of SARS disease. Results: The SOM algorithm used the intrinsic distributions of 21 social and economic variables to classify 31 regions into five clusters. SOM determined clusters were compared with the distributions of SARS outcomes. The result picture shows that the variability between regions clusters was significant with respect to the distribution of SARS occurrence. Conclusion: Our study demonstrated a positive relationship between socioeconomic conditions and SARS outcomes in regions using the SOM method to overcome data and methodological challenges traditionally encountered in public health research. Results demonstrated that community health can be classified using socioeconomic variables and that the SOM method may be applied to multivariate socioeconomic health studies.

Keywords-Self-Organization Map, visual data mining, SARS, China, socioeconomic

I. INTRODUCTION

Huge amounts of data are collected every day, and common databases contain several terabytes of data while today’s data management systems provide only limited means to access to the data. Visual Data Mining (VDM) is a theory, methodology and also a technology, which combine traditional data mining and information visualization techniques [1]. It uses the technology of Visualization to transform the data in a large data set to graphic or images and shows the result to screen. It’s extended from the integration of Visualization and Data Mining. VDK takes advantages of both data mining and visualization. It could analysis large data set at a time and be potential for automation. And also, the user is directly involved in the exploration process. It’s applicable even if only little is known about the data or the exploration goals are vague, or highly inhomogeneous and noisy data is given. VDK combines the power of automatic calculations and the capabilities of human processing.

A Self Organizing Map (SOM) which is also called Kohonen networks [2] is a feed forward artificial neural network. It’s an important method in VDK. It uses an unsupervised training algorithm to perform non-linear non-parametric regression. It is capable of projecting multidimensional datasets onto low, usually 1- or 2- D displays while preserving the useful information within the raw data, in doing so they enhance analyst ability to extract knowledge in the form of novel patterns, structures and correlations hidden in the data. Kohonen’s SOM techniques provide an excellent tool for analyzing disparate multidimensional datasets from different sources [3]. Using SOM techniques, multi sourced datasets, even with inconsistent labeling could be collectively analyzed to discover implicit, previously unknown useful knowledge embedded in the complex data set. SOMs are increasingly seen as better data clustering, visualization, and dimensions-reduction tools in data mining and knowledge discovery applications. SOM applications to exploratory data analysis have produced considerable success in many fields, such as medical, financial market, customer segment, industrial engineering, manufacturing to name a few from the lists of thousands of SOM applications in Kaskiy et al (1998) and Oja et al (2002)[4-8]. In this paper, we use SOM net to analysis the relationship between socioeconomic variables and the outbreak of SARS.

II. METHODOLOGY

A. Self-Organization Map

The SOM algorithm, first introduced by Teuvo Kohonen was developed from the basic information processing modeling in the human brain's cortical cells, known from the neuro-physiological experiments of the late twentieth century (Kohonen, 1982). In the training process of the algorithm, initially the SOM output layer units or nodes are assigned with a random set of vectors, usually referred to as a code book. Each set of the input vectors is then presented to the SOM input nodes and matched against the output units to find the best matching unit (BMU) in the code book. Once a BMU is found, the particular set of input vectors is assigned to it, and then that output unit vector values are adjusted more close to the values of the input set. The neighboring units of this BMU are also adjusted close to the values of the latter. Similarly, the whole set of input data are assigned to their BMUs in the output layer, mapping the similar input data vectors together on the 2D display with most of the original attributes preserved. Hence, the trained SOM display enables analysts to view any implicit previously unknown useful knowledge within the raw data in the form of patterns, structures and relationships.
B. The principle of Self-Organization Map

A typical architecture of the SOM network is shown in figure 1. The essential constituents of feature maps are as follows[9-10]:
- an array of neurons that compute simple output functions of incoming inputs of arbitrary dimensionality.
- a mechanism for selecting the neuron with the largest output.
- an adaptive mechanism that updates the weights of the selected neuron and its neighbors.

The training algorithm proposed by Kohonen for forming a feature map is summarized as follow:

Step 1: Initialization: Choose random values for the initial weights $w_j(0)$

Step 2: Winner Finding: Find the winning neuron $*$ at time $k$, using the minimum distance Euclidean criterion:

$$ j^* = \arg \min _{j} \| x(k) - w_j \|, \quad j = 1, \ldots, N^2. \quad (1) $$

Where $x(k) = [x_1(k), \ldots, x_p(k)]^T$ represents the $k^{\text{th}}$ input pattern, $N^2$ is the total number of neurons, and $\| . \|$ indicates the Euclidean norm.

Step 3: Weights Updating: Adjust the weights of the winner and its neighbors, using the following rule

$$ w_j(k+1) = w_j(k) + \eta(k) \Lambda_j(k) (x(k) - w_j(k)) \quad (2) $$

Where $\eta(k)$ is a positive constant and $\Lambda_j(k)$ is the topological neighborhood function of the winner neuron $j^*$ at time $k$. A typical choice of $\Lambda_j(k)$ is the Gaussian-type function

$$ \Lambda_j(k) = \exp\left(-\frac{d^2_{j^*j}}{2\sigma^2}\right) \quad (3) $$

where the parameter $\sigma$ is the “effect width” of the topological neighborhood and $d^2_{j^*j}$ is the lateral distance between neurons $j^*$ and $j$ in the discrete output space. It should be emphasized that the success of the map formation is critically dependent on how the values of the main parameters (i.e., $\eta(k)$ and $\Lambda_j(k)$) initial values of weight vectors, and the number of iterations are predetermined.

III. CASE STUDY AREA AND VARIABLES

Severe acute respiratory syndrome (SARS) was first recognized as a global threat in mid-March 2003[11-12]. The first known cases and the last case of SARS occurred in November 2002 and in July 2003. The international spread of SARS resulted in 8098 SARS cases in 26 countries, with 774 deaths. As an example of visual data mining using SOM, we choose the socioeconomic data to be trained in the map. Then we search the relationship between SARS occurrence and socioeconomic factors.

The SARS data which include the number of SARS patients in every region in China was got from the report of Ministry of Health of the People’s Republic of China, then the socioeconomic data from China population and employment statistics yearbook 2007. The data formula is as table 1.

| TABLE I. THE SOCIOECONOMIC DATA IN CHINA 2003 |
|-----------------------------------------------|
| region       | Beijing | Tianjin | Hebei  | Shanxi |
|--------------|---------|---------|--------|--------|
| Total case   | 2521    | 175     | 215    | 448    |
| population   | 1456    | 1011    | 6769   | 3314   |
| Birth Rate   | 5.1     | 7.14    | 11.43  | 12.26  |
| Death Rate   | 5.2     | 6.04    | 6.27   | 6.04   |
| Natural Growth Rate | -0.1  | 1.1     | 5.16   | 6.22   |
| Life Expectancy | 76.1  | 74.91   | 72.54  | 71.65  |
| unemployment population | 6.9579 | 12.0264 | 25.6917 | 13.0514 |
| unemployment rate | 1.4   | 3.8     | 3.9    | 3      |
| Average Wage | 25312   | 18648   | 11189  | 10729  |
| Average Family Size | 84.33  | 75.73   | 38.44  | 43.01  |
| Collective Household | 1144  | 150     | 161    | 331    |
| Family Household | 4981   | 3228    | 19248  | 9080   |
| household population | 13150 | 9544    | 63514  | 30851  |
| Agricultural rate | 75.5957 | 60.6885 | 30.38624 | 31.35289 |
| Average Family Size | 2.588084 | 2.898982 | 3.443007 | 3.176328 |
| No Schooling | 1.490251 | 1.160141 | 3.35903 | 1.925748 |
| Primary School | 6.796657 | 15.37187 | 26.58186 | 19.4027 |
| Junior Secondary School | 31.51203 | 41.51647 | 56.8679 | 56.4844 |
| Senior Secondary School | 24.47057 | 24.67311 | 10.38036 | 14.09269 |
| College       | 16.29526 | 9.659463 | 2.893588 | 3.541584 |
| university    | 15.98866 | 7.022996 | 1.387818 | 2.576083 |
| Graduate and Higher | 3.412256 | 0.538637 | 0.03598 | 0.170476 |

The next step, correlation analysis was used to find out the related socioeconomic factors. The result is as follow table 2.

| TABLE II. CORRELATION ANALYSIS RESULTS |
|-----------------------------------------|
| Attribute 1   | Attribute 2   | Correlation |
|--------------|---------------|-------------|
| total case   | population    | -0.0408     |
| total case   | Birth Rate    | -0.2505     |
| total case   | Death Rate    | -0.349      |
| total case   | Natural Growth Rate | -0.1888 |
| total case   | Life Expectancy | 0.3249     |
| total case   | unemployment population | -0.1202 |
| total case   | unemployment rate | -0.6423    |
| total case   | Average Wage  | 0.4443      |
| total case   | Urban Population | 0.5205     |
| total case   | Collective Household | 0.5133    |
| total case   | Family Household | -0.032      |
The correlation analysis shows that, Population, Birth Rate, Natural Growth Rate, Unemployment population, Family Household, Household population, Average Family Size, No Schooling, Junior Secondary School are non-correlation with SARS occurrence. Death Rate, Life Expectancy, Average Wage, Agricultural rate, Primary School, Senior Secondary School are low correlation with SARS occurrence. Unemployment Rate, Urban Population, Collective Household, College, university, Graduate and Higher Level are significantly related to SARS occurrence.

IV. RESULT

Choose the significantly related data as input, the cluster was got as figure 2

In this cluster, the whole data was classified into three parts, in which Guangdong province was selected alone, Beijing city classified together with Shanghai city, and the other as a whole cluster. Put all the related factors into the net, get the follow results.

| total case        | Household population | -0.027 |
|-------------------|----------------------|--------|
| Agricultural rate |                      | 0.4933 |
| Average Family Size |                    | -0.2632 |
| No Schooling     |                      | -0.2403 |
| Primary School   |                      | -0.4994 |
| Junior Secondary School |              | -0.0061 |
| Senior Secondary School |            | 0.4635 |
| College          |                      | 0.5706 |
| university       |                      | 0.6251 |
| College and Higher Level |       | 0.7145 |

### TABLE III. CLUSTER RESULT USING ALL RELATED DATA

| C1  | Liaoning, Hebei, Heilongjiang, Jilin, Anhui, Inner Mongolia, Fujian, Xinjiang, Hainan, Shanxi, Ningxia |
|-----|---------------------------------------------------------------------------------------------------|
| C2  | Zhejiang, Jiangsu, Shandong, Henan, Hebei, Sichuan, Hunan, Guangxi, Shaanxi, Chongqing, Jiangxi, Gansu, Yunnan, Guizhou, Qinghai, Tibet |
| C3  | Tianjin, Shanghai |
| C4  | Beijing |
| C5  | Guangdong |

In these two cluster results, Guangdong province and Beijing city was separated out, just similar to the distribution of the epidemic situation of SARS. Look into figure 4. Consider from the two results, the factors that with low correlation are important in the net training, the more information got, the more exact results got.

In this result, Tianjin city and Shanghai city were collected into the same cluster, and they are both similar to Beijing city (more near, more similar), but in fact, the number of SARS cases of these three cities is 2521 (Beijing), 175 (Tianjin), 8 (Shanghai). The situation in Beijing is far different from Shanghai. As we all known, SARS has very close relation with air temperature and environment factors. But Beijing city and shanghai city have far different climate and environment which are important in SARS transmission, though they have high similar population and social structure. Limited to the data, the air temperature and other environment factors were not discussed in this paper. Then the result is a little different from reality. The result shows that, under the same temperature and environment, shanghai city has a high similar with Beijing city in the SARS transmission. To get more exact result, more data should be added.

V. CONCLUSION

In this paper, SOM net was used to visualize the relationship between socioeconomic and SARS occurrence, the result shows that the socioeconomic factors are related to the SARS transmission. The regions that more similar in socioeconomic have more similar attribute in SARS transmission. To get more information, detailed data should be collected. SOM net could find the hiding information from huge amounts of data, and give the intuitionistic visualization. It has simpler net structure, strong ability of auto learning, and
computes fast. It could be useful in visual data mining of public health.

ACKNOWLEDGMENT

This research is partially supported by the Plans of the National Sci-Tech Major Special Item 2009ZX10004-720 and 2008ZX10004-012, the National Natural Science Foundation of China (40871181). Beijing Institute of Microbiology and Epidemiology, State Key Laboratory of Pathogen and Biosecurity provided the SARS data.

REFERENCES

[1] Keim, D.A, Information visualization and visual data mining. Visualization and Computer Graphics 8(1)2002,1-8
[2] Teuvo Kohonen, Self-organizing maps, 3rd ed, Berlin, Springer, 2001
[3] Georg Pölzlbauer, Michael Dittenbach, Andreas Rauber, Advanced visualization of Self-Organizing Maps with vector fields, Neural Networks 19 (2006) 911–922
[4] Seth E. Spielmana, and Jean-Claude Thillb, Social area analysis, data mining, and GIS Computers, Environment and Urban Systems Volume 32, Issue 2, March 2008, Pages 110-122
[5] Heather G Basara and May Yuan, Community health assessment using self-organizing maps and geographic information systems, International Journal of Health Geographics 2008, 7:67
[6] Mu-Chun Su and Hisao-Te Chang, Self-Organizing Neural Networks for Data Projection, ICSC’99, LNCS 1749, pp. 206-215, 1999.
[7] Antonio Soriano-Asensi, Jose’ D. Marti'n-Guerrero, Emilio Soria-Olivas, Alberto Palomares, Rafael Magdalena-Benedito, Antonio J. Serrano-Lo’pez, Web mining based on Growing Hierarchical Self-Organizing, Expert Systems with Applications 34 (2008) 2988–2994
[8] Banu Diri, Songul Albayrak, Visualization and analysis of classifiers performance in multi-class medical dataExpert Systems with Applications 34 (2008) 628–634
[9] Orkun Ersoy, Erkan Aydar, Alain Gourgaud, Hasan Artuner, Hasan Bayhan, Clustering of volcanic ash arising from different fragmentation mechanisms using Kohonen self-organizing maps, Computers & Geosciences 33 (2007) 821–828
[10] Yan Li, Social Area Analysis Using SOM and GIS: A Preliminary Research, RCAPS Working Paper No.07-3 August 2007
[11] Sze-Bi Hsu a, Lih-Ing W. Roeger, The final size of a SARS epidemic model without quarantine, J. Math. Anal. Appl. 333 (2007) 557–566
[12] Wendi Wang, Shigui Ruan, Simulating the SARS outbreak in Beijing with limited data, Journal of Theoretical Biology 227 (2004) 369–379