Prediction of Neural Tube Defect Using Support Vector Machine[1]

JIN-FENG WANG[#,2], XIN LIU[#], YI-LAN LIAO[#], HONG-YAN CHEN, WAN-XIN LI, AND XIAO-YING ZHENG[*,2]

[#]State key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China; [□]Institute for Sustainable Water, Integrated Management & Ecosystem Research (SWIMMER),University of Liverpool, Liverpool, L69 3GP, UK; [△]City University of Hong Kong, Tsinghua Graduate School at Shenzhen 518055, China; [*]Institute of Population Science, Peking University, Beijing 100187, China

Objective To predict neural tube birth defect (NTD) using support vector machine (SVM). **Method** The dataset in the pilot area was divided into non overlaid training set and testing set. SVM was trained using the training set and the trained SVM was then used to predict the classification of NTD. **Result** NTD rate was predicted at village level in the pilot area. The accuracy of the prediction was 71.50% for the training dataset and 68.57% for the test dataset respectively. **Conclusion** Results from this study have shown that SVM is applicable to the prediction of NTD.

Key words: NTD; Prediction; Small sample; SVM

INTRODUCTION

Neural tube defect (NTD) refers to any functional or structural anomaly present in infancy or later in life which represents a leading cause of infant mortality and disability in the world[1]. NTD affected people and their families incur the cost of health care for the whole life and thus usually become economically deprived, especially those living in rural areas.

The study on prediction of NTD would bridge the gap between the sampling survey sites[2] and estimation of its true distribution. Although previous epidemiological studies focused on the distributional pattern[3-4] and determinants of NTD[5-7], few dealt with its prediction[8]. NTD is an event of small probability, while conventional statistical methods are applicable to large sample size[9-10]. SVM is a machine learning model and is supposed to be useful for small sampling. The objective of this study was to apply SVM to the modelling and predicting of NTD. The dataset in the pilot area was divided into non overlaid training set and testing set. SVM was trained using the training set and the trained SVM was then used to predict the classification of NTD in a pilot area, namely Heshun county in Northern China. The prediction results reflected well the real life situation and thus the approach adopted in the exercise was promising for more general applications.

MATERIALS AND METHODS

Study Population

Heshun county located at the Tai Hang mountain area of Shanxi province in northern China was selected as a pilot area for this study (Fig. 1), where NTD occurrence is the highest in the world. Our purpose was to test the applicability of SVM to predict the NTD prevalence rate in Heshun. Heshun county consists of 326 administrative villages with an area of 2 250 km^2. Most of the people in this county are farmers and their living environment seldom changes for a long time. There was no large-scale human immigration in the history of this region. The

[1]This study was supported by CAS (KZCX2-YW-308), the MOST (2007DFC20180; 2007AA12Z233), and NSF (40471111; 70571076).

[2]Correspondence should be addressed: Jin-Feng WANG, State key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China, Tel: 86-10-64888965, Fax: 86-10-64889630, E-mail: wangjf@Lreis.ac.cn; and Xiao-Ying ZHENG, Institute of Population Research, Peking University/WHO Collaborating Center of Reproductive Health and Population Science, Beijing 100871, China, Tel: 86-10-62759185, Fax: 86-10-62751976, E-mail: xzheng@pku.edu.cn

Biographical note of the first author: Jin-Feng WANG, a professor who works in the Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China.

0895-3988/2010
CN 11-2816/Q

inherited and congenital causes of birth defects are believed to be similar among the people in this region. Most types of birth defects designated by WHO, which include anencephaly, spina bifida, encephalocele, holoprosencephaly, and hydrocephalus, can be found in Heshun. During 1998-2005, there were 7 880 births in Heshun with 187 NTD cases. Births occurred at the hospital or at home, and mothers were residents of the county during that time period. Also included were all therapeutic abortions performed among residents of the area whose estimated delivery date fell within the time period of interest. All NTD cases, regardless of pregnancy outcome, were verified by doctors in the hospital. Records of NTD cases were collected from local family planning departments.

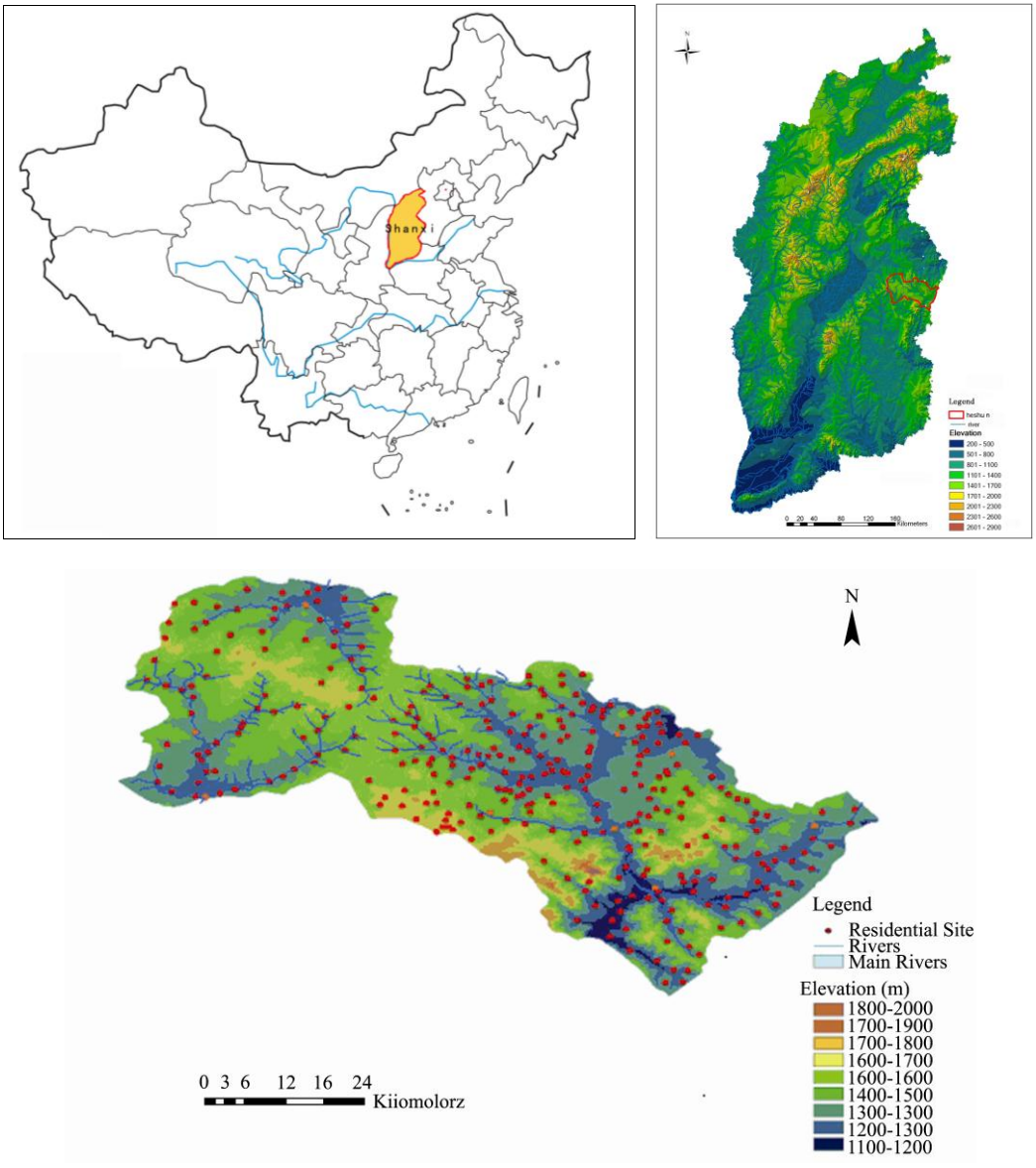

FIG. 1. Location of Heshun County. (a. China, boundary of provinces and major rivers; b. Shanxi province, elevation; c. Heshun county, elevation overlaid by the 326 villages, in which the data is collected).

In the study, input data includes the spatial distributions of both NTD rate and the surrogates of its suspected determinants in Heshun county[13-14]. The suspected factors in various villages were classified into socioeconomic and geographical factors. The socioeconomic factors reported useful information on medical conditions (the number of doctors), the per-capita incomes (per-capita net incomes), the agricultural chemical exposures (the use of fertilizers and pesticides), and the crop yields (vegetable productions) of every village. All socioeconomic data were provided by the Heshun statistical bureau. The

geographical factors included elevation, gradient, access condition (distance to main roads), water shortage condition (distance to rivers) and the geological background (distance to faults, type of soil and lithological classes) of the villages.Villages with less than a total of 5 births were excluded from the calculation in order to use stable rate data. The birth defect rate was divided into the following three categories: 0, (0, 0.08), [0.08, 1], namely, 1 = no birth defects, 2 = birth defect rate not high, 3 = high incidence of birth defects.

Categorical variables are treated as dummy variables. For examples lithology has seven categories (1, 2, 3, 4, 5, 6, 7), represented by dummy variables: lithology1, lithology2, lithology3, lithology4, lithology5, and lithology6, as shown in Table 1. Similarly, we introduced dummy variables to represent soil classes. Usually, we introduced n−1 dummy variables to represent those variables with n categories.

TABLE 1

The Introduction of Dummy Variables for Variable Lithology Types

Before	After					
Lithology Types	Lithology 1	Lithology 2	Lithology 3	Lithology 4	Lithology 5	Lithology 6
1	0	0	0	0	0	0
2	1	0	0	0	0	0
3	0	1	0	0	0	0
4	0	0	1	0	0	0
5	0	0	0	1	0	0
6	0	0	0	0	1	0
7	0	0	0	0	0	1

SVM

SVM (Support Vector Machine) is a machine learning method[11-12], which is based on the statistical theory and integrates the largest interval hyperplane, Mercer nucleus, convex quadratic programming and relaxation variable technologies. Abiding by structural risk minimization principle, SVM can effectively solve the practical problems such as small sample size, non-linearity, high dimension, and local minimum. The basic idea of SVM is to convert an input sample dataset to a high-dimensional feature space through a nonlinear transformation, then is used to calculate the optimal separating surface that separates samples linearly in the feature space (Fig. 1).

Support Vector Machine developed from the optimal separating surface in the circumstances of linearly separable, the so-called optimal separating surface, is able not only to separate two classifications correctly, but also to make the interval between the two classifications the largest. In Fig. 1, the right side of the H_1 is divided into positive category, while the left side of H_2 is divided into the negative category, and the samples located in the middle of two types refused classification (an alternative explanation of refusing classification is that classifying into either positive or negative domain is reasonable, and therefore it is actually unreasonable to limit to any one classification). The points on the border have special meanings. In fact, it is the border points that determine the optimal separating hyperplane. These points (imagining they are the points that are exact on the H_1 and H_2 in Fig. 1), in the issue of text classification, are vectors themselves, and are called support vector. The H_1 and H_2 are modelled by equations in the following functional form:

$$xw+b=0.$$

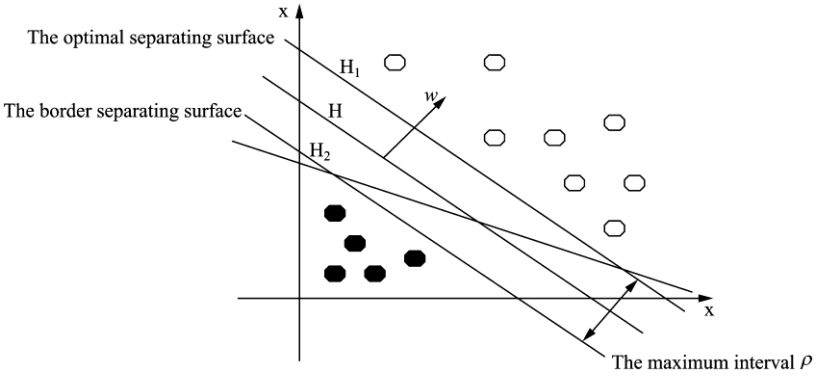

FIG. 2. The optimal separating surface in two dimensions.

where b and w are parameters. The distance between H_1 and the origin is $|-1-b|/\|w\|$, the distance between H_2 and the origin is $|1-b|/\|w\|$, therefore, the distance between H_1 and H_2 is $2/\|w\|$, which means the interval distance equals to $\rho=2/\|w\|$. The maximum interval is equivalent to the minimum of $\|w\|^2$.

Constraint conditions. The sample set (x_i, y_i), i=1, 2,…,n , $x\in R^d$, $y\in(+1, -1)$ that can be linearly separable must satisfy:

$(wx_i)+b\geqslant 1$, if $y_i=1$ and $(wx_i)+b\geqslant 1$, if $y=-1$

that is,

$$y_i\left[\left(w\cdot x_i\right)+b\right]\geqslant 1,\ i=1,2,\ldots,n \tag{1}$$

That is, samples must be on H_1 or H_2 side, rather than at the middle between the two. The sample satisfying the above equation is support vector.

The separating surface that minimizes the $(1/2)\|w\|^2$ and satisfies the condition (1) is called optimal separating surface, that is,

$$\min \frac{1}{2}\|w\|^2$$

$$\text{subject to}\quad y_i\left[\left(w\cdot x_i\right)+b\right]-1\geqslant 0\ \ (i=1,2,\cdots l)$$

(l is the sample size) (2)

where w is independent variables, the objective function is the above quadratic equation (2) w; all constraints are linear functions of w. Note, x_i here is not a variable, but it refers to a sample element, and is known.

Using the method of Lagrange multipliers (a_i) to solve this constrained optimal problem,

$$L(w, b, a)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{N}a_i[y_i(wx_i+b)-1] \tag{3}$$

differentiating w leads to $w-\sum_i a_i y_i x_i=0$, or

$$w=\sum_i a_i y_i x_i \tag{4}$$

differentiating b leads to $\sum_i a_i y_i=0$ (5)

consequently, problem (2) can be rewritten by Eqs (3)-(5) as

$$\text{maximize: } L=\sum_{i=1}^{n}a_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}a_i a_j y_i y_j x_i x_j \tag{6}$$

$$\text{subject to: } \sum_i a_i y_i=0 \text{ and } a_i\geqslant 0,\ \forall i \tag{7}$$

This is an optimization problem for quadratic function constrained by inequality, therefore there exists an unique solution.

In the case of linearly inseparable surfaces, we can add a relaxation item $\xi_i\geqslant 0$ in the condition (1), and it becomes,

$$y_i\left[\left(w\cdot x_i\right)+b\right]-1-\xi_i\geqslant 0, i=1,2,\cdots,n \tag{8}$$

$$\left(w,\xi\right)=\frac{1}{2}\|w\|^2+C\left(\sum_{i=1}^{n}\xi_i\right) \tag{9}$$

Then, change the target to find the minimum of equation (9) with the limitation of $0\leqslant a\leqslant C$, in which C is the punishment factor. It means, considering the smallest misclassification and the largest classification intervals, we can get the generalized optimal separating surface.

For nonlinear problems, we only need to non-linearly map the input vector to a higher dimension feature space, and further to construct the optimal separating hyperplane. We do not need to know the expression of the specific mapping function $\phi(x_i)$, since in this high-dimension space it only involves the inner product operation. If $K(x_i, x_j)=\phi(x_i)\cdot\phi(x_j)$, then $K(x_i, x_j)$ is called kernel function, the condition that a function is kernel function has been given by Mercer theorem. The corresponding optimal decision function becomes:

$$f\left(x\right)=\operatorname{sgn}\left(\sum_{i=1}^{n}a_i^* y_i K\left(x_i, x\right)+b^*\right) \tag{10}$$

SVM needs not to carry out characteristics reduction! It has a strong anti-interference ability.

We divided the data in line with the format of LIBSVM, a software of SVM, into two categories, NTD_ training (200 sample data) and NTD testing (70 sample data). During the process of classification training, svmscale.exe was called first to transform the original sample vector, then traverse the default c (Cost) and g (Gamma) parameter; next, svmtrain.exe was called to calculate the precision of parameters c and g, after reaching the best accuracy, then the model was ready for prediction use with the corresponding c and g.

RESULTS

The output of training and testing: c is punish factor in Eq. (9) and equal to 2048, g is objective Eq. (9) under constraint Eq. (8) and equal to 0.00012. First, the data was rescaled to the proper range so that training and predicting would be faster. Second, cross validation was used to find the best parameters cost (c) and gamma (g) for the model in order to reach the optimal accuracy. After that, the model with optimal parameters was obtained, which then was employed to predict the classification. Finally, the scaled testing data was applied to test the accuracy of model.

Fig. 3 suggests the process of cross validation. It is apparent that log2 (gamma) and log2 (C) valued -11 and 10 first. Also, -11.5 and 14 was tried. But after several validation tests, they changed to -13 and 11, and the accuracy increased from 69.5 to 71.5 correspondingly.

The experimental results showed that, when the parameters of RBF kernel were functioning, Cost and Gamma, were 2 048 and 0.00012207 respectively,

and the classification was most accurate and the accuracy rate was 71.5% for the training dataset and 68.5714% for the test dataset. This means that of the 200 sample training dataset and 70 sample testing dataset, respective 143 and 48 villages, were correctly classified using the support vector machine model. Table 2 lists part of the predictions, compared with the actual observed data.

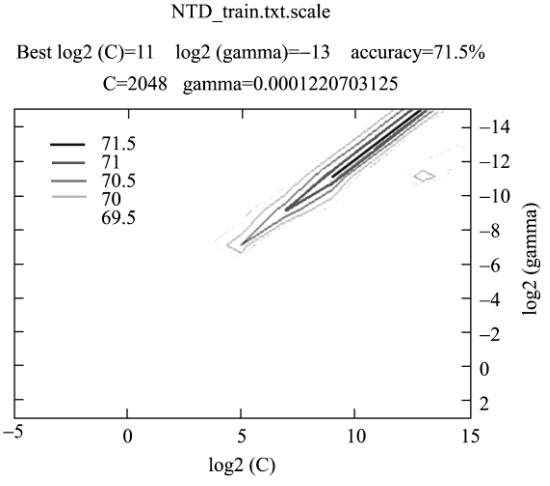

FIG. 3. Searching the optimal cost and gamma.

TABLE 2

The Predicted and Observed Classification of NTD Rate of Part of Villages in Test Dataset

Village Name	Observed Classification	Predicated Classification
Baimuzhai	1	1
Bomugua	1	1
Beicun	1	1
Ximaquan	1	1
Hedi	1	1
Xingcun	2	2
Caijiacun	2	2
Dongnao	2	2
Xinao	2	2
Renyuanzhi	2	2
Jiajiagou	1	2
Jiujing	1	1
Jingyugou	2	2
Raocun	2	2
Hougou	1	1
Hebei	3	1
Liujiayoa	1	2
Dongyaogou	3	2
Houshimengou	1	1
Fengtia	2	1
Baizhen	2	2
Huili	2	2
Yangpozhuang	1	2
Nanyao	2	1
Qiannanyao	2	1
Hounanyao	2	2
Taiyangpo	1	1
Qingbei	2	1

CONCLUSION

Neural tube birth defects are a group of rare but severe diseases, which usually impose high life burden on the patients and their families and impoverished them, especially in rural China. Spatial sampling survey monitors the situation at sampled sites, the overall dynamic of NTD in a region, including the values at unsampled sites which need to be predicted based on the observed sample values and using suitable models.

As a tool to handle small sampling redients, the support vector machine (SVM) was proved to be efficient to estimate the prevalence of neural tube birth defects at unsampled sites in the study area, and displayed its potential to be applicable to other areas o predict the rare events.

DISCUSSION

Although lots of tools could be used for prediction of NTD, few are suitable for small sampling featured by high dimension and nonlinear simultaneously. Unfortunately, the NTD occurrence has the above three features, which is small probability event, often effected linearly or nonlinearly by many environmental and social factors[7, 14-15]. The SVM algorithm addresses the above challenge by using a support vector and kernel function. Although SVM outperforms others in prediction for a small sampling, high dimension and nonlinear dataset, it is conceptually a black box approach. The explicit relationship between NTD and its suspect factors as input into the model, which leads to a good prediction, deserved to be explored in the future studies.

REFERENCES

1. Carmona R H (2005). The global challenges of birth defects and disabilities. *Lancet* **366**, 1142-1144.
2. Wang J F, Jiang C S, Li L F, *et al*. (2009). *Spatial Sampling Design and Statistical Inference*. Beijing: Science Press.
3. Rushton G, Lolonis P (1996). Exploratory spatial analysis of birth defect rates in an urban population. *Stat Med* **15**, 717-726.
4. Wu J L, Chen G, Song X M, *et al*. (2008). Spatiotemporal property analysis of birth fefects in Wuxi, China. *Biomedi Environ Sci* **21**, 432-437.
5. Wu J L, Wang J F, Meng B, *et al*. (2004). Spatial exploratory data analysis of birth defect risk factors' identification. *BMC Public Health* **4**, doi:10.1186/1471-2458-4-23.
6. Gu X, Lin L M, Zheng XY, *et al*. (2007). High prevalence of NTDs in Shanxi province: a combined epidemiological approach. *Birth Defect Research (Part A)* **79**, 702-707.
7. Wang J F, Li X H, Christakos G, *et al*. (2010). Geographical

detectors-based health risk assessment and its application in the neural tube defects study of the Heshun region, China. *International Journal of Geographical Information Science* **24**(1), 107-127.

8. Jiang Y, Du H Z, Zhu W Y, *et al*. (2008). Effects of a regional Chinese diet and its vitamin supplementation on proliferation of human esophageal cancer cell lines. *Biomed Environ Sci* **21**, 442-448.
9. Haining R (2003). *Spatial Data Analysis: Theory and Practice.* Cambridge: Cambridge University Press.
10. Christakos G (2005). *Random Field Models in Earth Sciences*. NY: Dover Publ.
11. Vapnik V, Chervoknenkis A Y (1971). On the uniform convergence of relative requencies of events to their probabilities. *Theory of Probabilities and its Application* **16**, 263-280.
12. Vapnik V (1995). *The Nature of Statistical Learning Theory*. New York: Springer-Verlag.
13. Buttenheim A M (2008). The sanitation environment in urban slums: implications for child health. *Population and Environment* **30**, 26-47.
14. Li XH, Wang J F, Liao Y L, *et al*. (2006). A geological analysis for the environmental cause of human birth defects based on GIS. *Toxicological & Environmental Chemistry* **88**, 551-559.
15. Wang J F, Christakos G, Han W G, *et al*. (2008). Data-driven exploration of 'spatial pattern-time process-driving forces' association of SARS epidemic in Beijing, China. *Journal of Public Health* **30**, 234-244.

(*Received August* 20, 2009 *Accepted June* 9, 2010)