Forecasting Model of Mass Incidents in China

An Explorative Research Based on Suppport Vector Machine

Jiashu Zhou [1,3], Erping Wang[1]*, Yiwen Chen[1], Xuanna Wu [1,3]
1:Institute of Psychology,
Chinese Academy of Sciences,
Beijing, 100101, P.R. China
zhoujs@psych.ac.cn,
wangep@psych.ac.cn,
chenyw@psych.ac.cn,
wuxn@psych.ac.cn
3:Graduate University of Chinese
Academy of Sciences (GUCAS),
Beijing, 100049, P.R. China

Yujie Ma[2]*
2: Institute of Mathematics,
AMSS, Chinese Academy of
Sciences,
Beijing, 100190, P.R. China
yjma@iss.ac.cn

*Correspondent Authors:
EPW, YJM and YJT

Yingjie Tian[4]*
4:Research Center of Fictitious
Economy and Data Science,
Chinese Academy of Sciences
Beijing, 100190, China P.R
tianyingjie1213@163.com

***Abstract*- [Purpose] Mass incidents have emerged as a serious social problem concerning national security in China. So, it is necessary to construct a forecasting model to predict such public events. In this paper, Support Vector Machines are applied to the model. [Method] Based on the social surveys conducted in 119 counties of Shanxi, Gansu and Hubei provinces, 3 multi-class classification problems were proposed, and then 3 multi-class Support Vector Classification forecasting models were constructed. [Results] Preliminary experiments have proved that our method, compared with multiple cumulative logistic regression, should be more effective and accurate(enter method as well as the stepwise one). [Conclusion] It can be concluded from the results that irrationally behavioral intentions can be predicted more accurate than those rational ones. When the collective attitudes are applied to the forecast of the collective behavioral intentions, SVM method was approved to be the most effective approach. This paper represents an originally explorative research.**

Keywords- Mass incident; Collective action; Classification; Support Vector Machine; Forecasting Model;

I. INTRODUCTION

Mass incidents in China, violent or nonviolent, highlight the confrontation and conflict with governments at all levels and/or some powerful social groups, which are subsets of collective actions.

Many previously researches concerning the collective actions are available in scholar literature database, such as PsychInfo, Elservier, Web of Knowledge… In the amount of research literatures in disciplines of sociology [1][2], psychology [3][4], politics [5] (public administration), economics, social psychology, collective attitude and the history of social phenomena can be used to predict or forecast behaviors that have not happened yet [6][7][8][9]. For the attitude-behavior consistency (ABC), there is a recent review [10] for a literature reference. All these articles proposed that it is possible to use attitude's measurements to estimate, predict or forecast behaviors, even in the cases of collective attitude and behaviors. Based on this statement, it is plausible for researchers to predict or forecast the occurrence of 'mass incidents'(so as the behavioral intentions of collective actions) in China by social survey concerning the local residents' attitudes. Apparently, most feasible or accessible approach is multiple regression analysis, linear or logistic (binary; cumulative; nominal).

However, recently, some people [3] argued that '…regression is rarely useful for prediction in most social science contexts.' Therefore, improper linear models are disasters for prediction or forecasting. On the other hand, few support vector machine (SVM) techniques have been used as a resolution in field of neuroscience, psychology, behavioral sciences and medicine last years [11][12][13], and there is even no application of SVM technique in the field of social or collective behaviors' prediction. So, it may be necessary to take advantage of SVM to explore the probability of mass incidents' modeling in some regions of China as a methodology extension for psychology, statistics and computer engineering . Further more, since it has been not searched a single article focusing the predicting or forecasting of collective action in the PsychInfo Database, the research presented here may be an original one.

.

II. METHODS AND PROCEDURE

A. Social survey and data structure

Since 2004, a sociopsychological survey of collective attitudes' was annually conducted in 119 counties of Shanxi,

978-0-7695-3705-4/09 $25.00 © 2009 IEEE
DOI 10.1109/BIFE.2009.44

IEEE computer society

Gansu and Hubei provinces. For each county, about 40 to 80 participants were interviewed, and allowed to answer the questionnaires.

In the questionnaire of year 2008's , besides the personal information such as age, salary, education, sex, family condition, attitudes (about 30 items) and behavior intentions (9 items) questions were also asked. Attitude questions are viewed as the independent variables, and the behavior intentions are viewed as the dependent variables. For each question, 5 point Likert scale was recorded, such as "very satisfied=5; a little satisfied=4; unknown = 3; a little unsatisfied=2; very unsatisfied=1", apparently, the scale is an ordinal one [14][15]. The cases' number of the whole dataset is 5588. After the missing data by case (pairwise) was deleted, 5267 cases remained.

B. Brief Introduction of Support vector Machine

Since SVMs were proposed in the 1990s, they have been successfully applied to a wide range of pattern recognition problems including handwriting recognition, object recognition, face detection, text categorization and so on [16]. Now we briefly introduce the standard algorithm *C*-SVC for classification problems:
For the given training set

$$T=\{(x_1,y_1),\cdots,(x_l,y_l)\} \tag{1}$$

with input $x_i=(x_{i1},\cdots,x_{in})^T\in R^n$ and output $y_i\in\{-1,+1\}$, where x_{ij} represent the *j*th feature of the *i*th feature vector *xi*. Let $\phi:R^n\rightarrow \mathrm{H}$ be a mapping from input (feature) space to a Hilbert space H . *C*-SVC finds a hyperplane $(w\cdot\phi(x))+b=0$ which can separate the two classes with the maximal margin and minimal training errors in the Hilbert space. By applying a kernel function to replace the inner product in *H*, the corresponding decision function is

$$f(x)=\operatorname{sgn}(\sum_{i=1}^{l}\alpha_i^* y_i K(x_i,x)+b^*), \tag{2}$$

where α^* is the solution of the following optimization problem

$$\min_{\alpha}\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}y_i y_j\alpha_i\alpha_j K(x_i,x_j)-\sum_{j=1}^{l}\alpha_j, \tag{3}$$

$$\text{s.t. } \sum_{i=1}^{1}y_i\alpha_i=0,\ \ 0\leqslant\alpha_i\leqslant C,\ \ i=1,\cdots,l, \tag{4}$$

and b^* can be obtained as follows: if there exist $\alpha_j^*\in(0,C), j=1,\cdots,l,$ then

$$b^*=y_j-\sum_{i=1}^{l}y_i\alpha_i^* K(x_i,x_j). \tag{5}$$

Many kernel functions can be used in *C*-SVC, including polynomial kernel $K(x_i,x_j)=(1+(x_i\cdot x_j))^d$, RBF kernel $K(x_i,x_j)=\exp(-\frac{\left\|x_i-x_j\right\|}{2\sigma^2})$ and so on. In [17], RBF kernel is used because it perform well than other kernels, so in this paper, we use RBF kernel function to test the performance of our method.

C. Construction of Problems

First. The cases with any missing datum were deleted by case (pairwise). After the pairwise excluding, 5267 cases were left for further analyzing. Considering the huge contrast between the urban and suburbs in China, data from the two type regions were treated separately and analyzed independently.

For each dependent variable (attitude scale), the different satisfactory status were valued on a vector with five elements which represent the correspondent attitude of the question, for example, the vector (1,0,0,0,0) represents the "very satisfied", the vector (0,1,0,0,0) represents the "satisfied", ...and so as the vector (0,0,0,0,1) represents the "very unsatisfied" . The similar construction method was applied to other ordinal variables (dependent variables as well as independent ones) in the whole social survey dataset.

Since there are 5 vectors and about 81 attitudes questions or personal information questions in the survey, there are about 5 × 81 = 405 dimensions in the whole model of *C*-SVC.

1) Problem 1

Since there are 9 behavior intentions as the dependent variables, and for each variable, there are 5 point scaling as well, directly, the SVM-C is five classification problems. The five categories, adhering to the actual responses of the interviewees', are the subjective probabilities to take the true behaviors: very small, comparatively small, middle possibility, comparatively large, very large. So, in this problem, Dependent variables (Behavioral Intentions (BIs) in this research) has 9 classifications and each classification has 5 sub-classification/points.

2) Problem 2

Since five classifications SVM may have less accurate of prediction. Further merging of the categories was taken in Problem 2. In this case, only three categories were taken for each intention: smaller possibility (counting the sum of "very small" and "comparatively small" 's frequencies in Problem 1), middle possibility(the same as Problem 1), larger possibility (counting the sum of "very large" and "comparatively large"'s frequencies in Problem 1). So, in this problem, BI*s* have 9 classifications and each classification has 3 sub-classifications/points.

3) Problem 3

Based on Problem 2, in order to improve the accuracy further, binary classification was taken, that is, summing the "smaller possibility" and "middle possibility" as the "less possibility" category, and viewing the "larger possibility" as the "more possibility" one. So, in this problem, BI*s* have 9 classifications and each classification has 2 sub-classifications/points.

D. Experiments

1) Processing of Problem 1

a) Step 1.

Whole dataset was randomly separated into the training set and the test set.

b) Step 2.

Using Algorithm *C*-SVC to train the data in the training set. Choosing parameters of *C*-SVC by cross-validation.

c) Step 3.

Test the training model in test set. Result is given in Table 1 .

2) Processing of Problem 2 and Problem 3

The steps of Problem 2 and 3 are similar with the procedure presented in 1).

E. Results and Discussion(Comparison with other methods)

In Table I, the predicting accuracy of *C*-SVC is far higher than the determined coefficient-- analogy R^2 (in multiple cumulative logistic regression, MCLR). Even after the stepwise procedure, the analogy R^2 reduced. Although any research literature that argues the prediction/forecasting power of logistic regression has not been found yet, the validity of multiple linear regression for forecasting is low and deeply suspected[3]. Since the multiple linear regression (MLR) and logistic regression both belong to the generalized linear model (GLM), the limitation of MCLR in solving the multiple variable statistic model can be anticipated or estimated. While, the SVM-C, as a sub-method of SVM, may be a more powerful approach[18].

TABLE I . COMPARISON OF SOME ACCURACY OF DIFFERENT METHODS

Behavioral Intentions	Classifications' Numbers	analogy R^2 of enter logistic model	analogy R^2 of stepwise logistic model	C-SVC accuracy
On Striking	5	61.9%	61.2%	100%
	2	77.5%	77.6%	99.981%
Petition	5	57.7%	56.4%	99.6%
	2	66%	64.9%	99.981%

From Table II, it can be inferred that after selecting the optimal parameters in *C*-SVC, the predict accuracies (measured in percentages) can be increased for great intents, both in urban samples as well as suburb's. Further more, it can be discovered the improvement of the accuracies is larger for suburb sample which may attribute the sample size effect.

TABLE II . COMPARISON OF THE PREDICT ACCURACIES(%) OF *C*-SVC

Behavioral Intentions (DV)	Urban vs. Suburb	Parameters NOT SELECTED	Optimal SELECTED and RETEST
PN	S	61.49	84.96
	U	80.12	100
UR	S	67.15	99.88
	U	86.34	100
ST	S	77.20	99.98
	U	93.48	100
CD	S	69.13	100
	U	86.96	100
TL	S	67.80	95.36
	U	78.89	87.89
LT	S	65.34	96.10
	U	77.95	100
RN	S	79.10	99.98
	U	72.98	100
PT	S	65.66	99.98
	U	65.22	100
CL	S	94.14	98.10
	U	94.10	100

Table II DV=Dependent variables, PN=Personal negotiation; UR=Unify relatives and friends to resist; ST=Striking; CD=Collective demonstration; TL=Tolerance; LT=Litigating; RN=Reflecting to the news agencies; PT=Petition; CL=Leaving the local collectively; U=Urban; S=Suburb;

Table III implies that, overall, the accuracies of retest of suburb samples are much higher than urban ones in most cases. And most important, for those irrational "behavioral intentions ", such as Unify relatives and friends to resist (UR), Striking (ST), Collective demonstration (CD) and Leaving the local collectively (CL), their accuracies are much higher than other rational ones, esp. for the urban sample which means this model of SVM-C is sensitive to predict the behavioral intentions that are coherently associated with the actual collective actions from the collective attitudes collected from the social survey.

TABLE III . OPTIMAL PARAMETERS OF BEHAVIORAL INTENTIONS

Behavioral Intentions	Optimal *C*	Optimal g in RBF	Accuracy of retest (%)	Urban vs. Suburb
PN	2	0.0625	68.9873	Urban
UR	32	0.03125	91.1392	
SK	32	0.25	93.038	
CD	32	0.25	87.9747	
TL	2	0.015625	65.8228	
LT	32	0.25	56.3291	
RN	32	0.25	60.7595	
PT	32	0.25	60.7595	
CL	32	0.03125	95.5696	
PN	32	0.0078125	84.96	Suburb
UR	32	0.5	99.98	
SK	32	0.5	99.981	
CD	32	0. 5	100	
TL	2	0.0625	95.36	
LT	2	0.0625	96.10	
RN	32	0.25	85.0498	
PT	32	0. 5	99.981	
CL	32	0.0125	98.10	

Table III. DV=Dependent variables, C: linear parameter of the SVM-C, g: nonlinear model of SVM-C, PN=Personal negotiation; UR=Unify relatives and friends to resist; ST=Striking; CD=Collective demonstration; TL=Tolerance; LT=Litigating; RN=Reflecting to the news agencies; PT=Petition; CL=Leaving the local collectively.

III. CONCLUSION AND DISCUSSION

A. The High Predict Power of SVM in The Collective Action Forecasting Model

Since the application of SVM in the field of collective actions' prediction has not been searched in many database—PsychInfo & PsychArticle, Elservier, Web of Knowledge and MedLine, it seems that, authors doubt, it is the originally explorative and quantitative research presented in this paper. However, compared with other proper regression method, either logistic cumulative regression or stepwise logistic cumulative regression, the SVM methods have some shortcomings [18][19]. Even so, the SVM-C approach has its incredible advantage in the collective actions' forecasting model.

B. Summary

This original research proposed the accessibility of applying SVM-C method to study the predicting factors of collective intentions even actions which concern the social harmony and national security of China. It has been found that irrationally behavioral intentions can be forecasted more accurate than rational ones from attitudes by method of SVM-C.

C. Future Study

For further research, the modeling between collective attitudes and collective actions directly is anticipated urgently. Because the actual data concerning mass incidents' behaviors have not been counted in the research, next, SVR (support vector regression) approach is anticipated to explore the accuracy of the predict model in a similar way.

ACKNOWLEDGMENT

This study is founded by both the Natural Science Foundation of China (No. 70731004 and No.10601064) and the Knowledge Innovation Project of the Chinese Academy of Sciences (Grant No. KSCX2-YW-R-130), and all authors thank to M.S. Pang, Li-Rong for her diligent work of statistic programming in this research.

REFERENCES

[1] S. Kanazawa. "A new solution to the collective action problem: The paradox of voter turnout". American Sociological Review, vol. 65(3), , pp. 433-442, 2000.

[2] D.S. Crystal and M. DeBell. "Sources of civic orientation among American youth: trust, religious valuation, and attributions of responsibility". Political Psychology, vol. 23(1), pp. 113-132 , 2002.

[3] J. Dana and R.M. Dawes. "The superiority of simple alternatives to regression for social science predictions". Journal of Educational and Behavioral Statistics, vol. 29(3), pp.317-331 , 2004

[4] A. Bliuc, C. McGarty K. Reynolds and D. Muntele. "Opinion-based group membership as a predictor of commitment to political action". Eur.J.Soc.Psychol., vol. 37, pp. 19-32, 2007.

[5] T.A.T. Eyck. "Does information matter? A research note on information technologies and political protest". The Social Science Journal, vol. 38, pp. 147-160,2001.

[6] L. Musgrove and C. McGarty. "Opinion-based group membership as a predictor of collective emotional responses and support for pro- and anti-war action". Social Psychology, vol. 39(1), pp. 37-47,2008.

[7] S.D. Reicher. " ' The Battle of Westminster': developing the social identity model of crowd behavior in order to explain the initiation and development of collective conflict". Eur.J.Soc.Psychol., vol. 26, pp. 115-134,1996.

[8] M. Lubell. "Familiarity breeds trust: collective action in a policy domain". The Journal of Politics, vol. 69(1), pp. 237-250 , 2007.

[9] D. Baldassarri and P. Bearman. "Dynamics of political polarization". Am. Soc.Rev., vol. 72, pp. 784-811 , 2007.

[10] W.D. Crano and R. Prislin. "Attitudes and persuasion". Ann.Rev.of Psychol., vol. 57, pp. 345-374 , 2006.

[11] X. Li, D. Lord, Y. Zhang and Y.C. Xie. "Predicting motor vehicle crashes using support vector machine models". Accident Analysis and Prevention, vol. 40, pp. 1611-1618 , 2008.

[12] H. Fröhilich, A. Hoenselaar and J. Eichner et.al. "Automated classification of the behavior of rats in the forced swimming test with support vector machines". Neural Networks, vol. 21, pp. 92-101,2008.

[13] B. Ou, P. Lin and S. Vorbach. "Exploration of computational methods for classification of movement intention during human voluntary movement from single trial EEG". Clinical Neurophysiology, vol. 118, pp. 2637-2655, 2007.

[14] S.S.Stevens. "On the theory of scales of measurement." Science, vol. 103, pp. 667-680, 1946.

[15] S.S.Stevens. "Mathematics, measurement and psychophysics". In S.S.Stevens(1st. ed.), Handbook of experimental psychology (pp.1-49). New York: Wiley, 1951.

[16] C. J.C. Burges. "A tutorial on support vector machines for pattern recognition". Data Ming and knowledge Discovery, vol. 2(2), pp.121–167 , 1998.

[17] P.D.Dobson and A.J. Doig. "Distinguishing enzyme structures from non-enzymes without alignments". Journal of Molecular Biology, vol. 330(4), pp.771–783 , 2003.

[18] H.M. Tzeng, Y.L. Lin and J.G. Hsieh. "Forecasting violent behaviors for schizophrenic outpatients using their disease insights: development of a binary logistic regression model and a support vector model". International Journal of Mental Health, vol. 33(2), pp. 17-31, 2004,.

[19] J.D. Long. "Omnibus hypothesis testing in dominance-based ordinal multiple regression". Psychological Methods, vol. 10(3), pp. 329-351, 2005.