Communication Records Analysis using Gray Prediction

Sun Xiaofei

(Department of Computer Science and Technology , Zaozhuang University, Zaozhuang 277160,China)

Keywords: communication records; personal privacy; combat crime; grey correlation degree; grey prediction

Abstract. Communication records are personal privacy, while the reasonable use of communication records can combat crime and protect the interests of citizens. This paper tries to find a balance between the two and predict the occurrence of criminal activities with the least communication records. A method is proposed to identify key contacts using grey correlation degree and predict that that something is off using the grey prediction model. Firstly, key contacts were identified according to communication frequency and other influencing factors, and then the occurrence of criminal activities was predicted using the grey prediction model. The experiments showed that this method is simple and it can predict criminal activities rapidly and accurately.

Introduction

With the development of society, criminals have more knowledge and higher anti investigation ability than ever, so investigators are faced with great challenges. Computer aided investigation software is used to solve cases. Communication records analysis is an important criminal investigation method which is based on simple data statistics. A method using grey correlation degree and gray prediction to analysis communication records is proposed.

Gray prediction can predict complex systems which have uncertainty factors, so it is getting more and more popular[1,2,3]. There are many studies of establishment of gray model using prediction model of interval gray number sequence in recent years. These studies focus on the expression of grey number sequence[4], gray number geometry[5] and gray number of synthetic gray[6]. Zeng proposed improved prediction model of interval grey number based on kernel and gray scale[7].A method calculates the time of the crime using gray prediction is proposed in this paper, and it can replace manual analysis. First, we monitor the communication of suspects, key contacts are selected and prediction model is established, then we can predict communication trend of suspects. If prediction differs greatly from actual values, there's reason to believe something's happening. The method is accurate, fast and practical.

1 Key contacts selection method

It’s important to select key contacts in case investigation, and our method has better accuracy using grey Correlation Degree. We got phone bills from telecom companies and created a database about contacts, calls and other information, which contained about 600 contacts. Criminal suspect’s detail lists about call and SMS were analyzed, which contain call time, SMS time, count, fee, latitude and longitude and so on. Key factors receive special attention such as count of call and SMS, count of contact during special time. The first sieving gets rid of less closely related contacts and choose top 10 contacts to analyze. Besides count of call and SMS, count of contact during special time and call time length, we can increase or decrease key factors according to need. Standard 0-1 transform for data is performed then. Variable i indicates the number of contacts which values from 1 to 10. Variable j represents four key factors which values from 1 to 4. m_{ij} represents the factor j for the contact i. The factor is n_{ij} after standardization. The standard formula is as follows.

$$n_{ij} = (m_{ij} - m_{j,\min}) / (m_{j,\max} - m_{j,\min}) \quad (1)$$

$m_{j,\min}$ is the smallest of the factor *j*, and $m_{j,\max}$ is the maximum of the factor *j*. Table 1 is the data sample after standardization. Count of call and SMS, count of contact during special time and call time length in table 1 are processed with 0-1 transformation. For example contact number 1586***1899 has the most number of calls, so it is 1 after 0-1 transformation. In this way, we can shield the huge difference between different parameters and get accurate predicted results. We assign different weights to each parameter to produce different results and the weights can be acquired through multiple trials and we also consider the factors of experience. Weights do not affect the choice of the experimental algorithm. The weights of the experiment are fixed, and the weight of each parameter is same. The formula for calculating the gray correlation coefficient is as follows.

$$\xi(ik)=\frac{\min\limits_{i}\min\limits_{k}\left|x_0(k)-x_i(k)\right|+\rho\max\limits_{i}\max\limits_{k}\left|x_0(k)-x_i(k)\right|}{\left|x_0(k)-x_i(k)\right|+\rho\max\limits_{i}\max\limits_{k}\left|x_0(k)-x_i(k)\right|} \quad (2)$$

$\xi(ik)$ is correlation coefficient. $\left|x_0(k)-x_i(k)\right|$ represents the absolute difference between x_0 sequence and x_i sequence in *k* point. $\rho\in[0,1]$, and it is the resolving coefficient. The bigger the resolution coefficient, the bigger the resolution, and the smaller the smaller[7].

$$r_i=\frac{1}{N}\sum_{k=1}^{N}\xi(ik) \quad (3)$$

r_i is the gray correlation degree between contact *i* and the ideal target, and the correlation degree of each parameter of ideal target is 1.

Table 1 Correlation coefficients and correlation degrees

contact	call count	SMS count	call time span	call count during special time	r_i	value rank
1876***0516	0.4791	0.5975	0.3989	0.3333	0.4571	6
1336***5856	1	0.3824	0.7346	0.3333	0.6126	2
1307***7958	0.3116	0.3333	0.4135	0.4234	0.3705	10
1567***1752	0.8236	0.8331	1	0.3333	0.7475	1
1307***5976	0.5641	0.3333	0.3571	1	0.5636	3
1388***2675	0.3968	1	0.3248	0.3333	0.5137	4
1573***4859	0.2963	0.3333	0.2731	0.7153	0.4045	9
1345***7836	0.6583	0.3333	0.6274	0.3333	0.4881	5
1568***6579	0.4816	0.3333	0.4927	0.3333	0.4102	7
1316***7285	0.5126	0.3333	0.4235	0.3569	0.4066	8

The experimental data involved user's privacy, we hide several intermediate. The results of the correlation analysis were evaluated. The greater the correlation degree, the higher the value of the contact person. According to the principle of correlation analysis, the contact sequence with highest correlation degree is the closest to the standard target sequence, so the number 1567***1752 is closest to the ideal target, and it has the largest investigation value. Secondly 1336***5856, and the investigation value of number 1307***7958 is the lowest.

2 Suspect trend analysis

2.1 Selection of prediction or data mining

The prediction method focuses on the forecasting ability of the algorithm while data mining pays more attention to the algorithm of the explanation. The predictive power is required to train the good algorithm on the training set and the error on the verification set should be as low as possible. The predictions given by the model should be as close to the actual results as possible. The explanatory nature of data mining is more concerned with the link between the input p dimensional variables, or the prediction of the algorithm is the main effect of which several variables. In another word, what is the basis of the prediction results? Should we choose prediction or data mining? An important principle is to see if we have the ability to change the data source or redesign the input

data. The data source is irrevocable in the phone bill analysis. Under the given conditions, we need to predict the results accurately, so we choose the prediction to design the algorithm, analyze the abnormal in the communication record and when the suspect to abscond.

2.2 Application of grey forecasting model in the judgment of the abnormal in the communication record

If the suspect's communications with the key contacts become unusually frequent, or the special time period of communication increased, means that something may happen. Conversely, the suspect breaks the long-term communication habits with a substantial reduction in communication frequency, or the suspect reduces the communication frequency with some key contacts. This is often the time that the suspect is alarmed and ready to abscond. The gray prediction model is established for the daily communication times of the key contact person and the number of the communication times of the next day is predicted. Compares the number of times with the actual number of times, if the difference is greater than a threshold, it is considered that things happen. This saves the time for the investigators to win the decision.

The steps of establishing the gray forecast model for the number of the suspects and the key contacts are as follows.

Assuming the primary sequence is $x^{(0)} = (x^{(0)}(1), x^{(0)}(2), \cdots, x^{(0)}(n))$, then the rank of sequence is as follows.

$$\lambda(i) = \frac{x^{(0)}(i-1)}{x^{(0)}(i)}, i = 2, 3, \cdots, n \qquad (4)$$

If all the levels are in the interval $(e^{\frac{-2}{n+1}}, e^{\frac{2}{n+1}})$, data sequence $x^{(0)}$ gets GM（1，1）model. Otherwise, the data is processed properly. Formula of GM（1，1）model is as follows.

$$y^{(0)}(i) = x^{(0)}(i) + c, i = 1,2, \cdots, n. \qquad x^{(0)}(i) + az^{(1)}(i) = b \qquad (5)$$

Estimate values of *a* and *b* are obtained by regression analysis, and corresponding differential equation is $\frac{dx^{(1)}(t)}{dt} + ax^{(1)}(t) = b$. （6）

Variable a is the ash number and *b* is the endogenous control ash. We obtain the following formula by the least square method. Finally, the following prediction values are obtained.

$$\hat{x}^{(1)}(i+1) = (x^{(0)}(1) - \frac{b}{a})e^{-ai} + \frac{b}{a} \qquad (7)$$

We select 10 key suspects associated with a suspect. By comparing predicted and actual values of the number of communications between suspects and key contacts, we focus on the time interval with the difference greater than the threshold.

Table 2 Prediction of whether something happens

contact	March 2nd(actual values / predicted）	March 3rd (actual values / predicted）	March 4th (actual values / predicted）	March 5th (actual values / predicted）	threshold	Whether something happens
1876***0516	3/3	2/3	4/3	1/3	0.75	F
1336***5856	5/6	7/6	5/6	1/6	1.5	True
1307***7958	2/1	1/1	2/1	1/1	0.25	F
1567***1752	8/8	6/8	7/8	1/8	2	True
1307***5976	6/5	5/5	2/5	0/5	1.25	True
1388***2675	3/4	4/4	3/4	2/4	1	F
1573***4859	2/2	1/2	2/2	2/2	0.5	F
1345***7836	4/4	3/4	3/4	2/4	1	F
1568***6579	3/3	3/3	2/3	2/3	0.75	F
1316***7285	3/3	4/3	2/3	0/3	0.75	True

We set the threshold to $T = x^{(1)}(k)/4$. Namely, the threshold is 1/4 of the prediction value. Through the above analysis, the number of communication times of the suspect and most of the main contacts is lower than the expected threshold in March 5th. Changes like these are signs that something's going on. Through the comparison with the actual situation, the prediction results are totally accord with the actual situation at that time.

3 Conclusions

The conclusions of application with grey correlation degree method and prediction method keep agreement with professional investigators. After screening, a few numbers with the greatest relevance is judged as key contact number from more than 500 contacts .This conclusion is drawn on the basis of the comprehensive analysis of several key factors. Besides, a more scientific and objective analysis of the results can also according to the actual weighted properly.

After obtaining the key contact, gray forecasting model is established based on the number of communication between the suspect and the key contact and the "vanishing point" can be accurately judged, which can be proved by experiment. Compared with the average method, the number of calls obtained by the gray prediction is more close to the real value. This is due to the telephone conversation between the rules often have stages, which can be reflected by the use of gray forecasting method while the mean method cannot do this.

This paper discusses the factors such as the number of calls, the number of messages, special time, etc. with the fixed weights which are obtained by the statistics of large number of users. Because different suspect have different communication habits, the focus of future research will be on the communication habit of suspects to dynamically determine the weights of influencing factors to achieve higher prediction accuracy.

4 Acknowledgments

This work is supported by the undergraduate applied talents training program for professional development, the project of Shandong province higher educational science and technology program (grant No. J12LN53) and the project supported by youth foundation of Zaozhuang university (grant No. 2011QN43).

References

[1]Zheng Zhiyong, Zhang Guanghua. Deformation analysis and prediction based on GM (1,1) model [J]. Mine Surveying and Mapping. 2012.28(4):98-112.

[2] Ouyang Lian. The research and application of GM (1,1) -Logistic subgrade settlement combination forecast model [J]. Railway Science and Engineering Journal. 2010.7 (4): 73-76.

[3]Gong Chenglin, Guo Aimin, et al. The application of gray related degree analysis and analytic hierarchy process in the evaluation of grape quality [J]. Southwest China Journal of Agricultural Sciences.2002.15 (1).

[4] Deng J L.The foundation of gray system[M].Wuhan: Huazhong University of Science and Technology Press,2002:1-496.

[5] Zeng B,Liu S F. Prediction model of interval grey number based on its geometrical characteristics[J]. Journal of Systems Engineering,2011,26(2):122-126.

[6] WANG Da peng,Wang Bing wen,LI Rui fan.Improved prediction model of interval grey number based on the characteristics of grey degree of compound grey number[J].Systems Engineering and Electronics,2013,35(5):1013-1017.

[7] Zeng Bo. Prediction model of interval grey number based on kernel and gray [J]. System Engineering and Electronic Technology.2011.4 (33): 821-824.