A Novel Algorithm for Criminal Statistical Processing

LIN Jianhui[1, a *], Chen Jin[2,b]

[1] Department of Information Technology, Hubei University of Police, P.R. China

[2] Adult education office, Hubei University of Police, P.R. China

[a]linjh_eis@126.com, [b]8255@hbpa.edu.cnl

* Corresponding author

Keywords: Social order trend; Criminal statistical data; profiling vector; Anomaly value

Abstract. Social order tendency analysis is studied based on intensive probe to present criminal information research jobs. The phenomenon of "routine fluctuation" and "warning fluctuation" in the field of social public order is extensively studied. An urban crime distribution vector generation algorithm, which based on exponential attenuation, is generated by combining the historic statistic data and with the present one. a Composite Statistic Profiling-Vector Hypothesis-test Algorithm, consist of a system anomaly testing expression and a corresponding referenced anomaly testing expression, is presented to exam the present statistics value set, thus social system anomaly index is obtain. Prototype system and experiment results show its precision and credibility.

Introduction

As the developments and changes of politics, economics and culture in a certain human society, social order tendency fluctuates correspondingly. To be objectively, change is eternal; however the changes of social order surely affect the social stability and social life. Therefore, how to distinguish the warning changes from routine changes is becoming event important for us[1].

Generally, crime tendency can be abstracted from the comprehensive processing toward many relevant factors, which involve social status, population fluctuation, political reformation, crime quantity, crime type, crime structure, etc[2]. In historical researches, time series analysis algorithm, regression analysis, SVM, Bayes analysis and Markov chain algorithm are used to calculate or forecast the trend of social crime[3]. Though many researches on criminal trend have been achieved, we still find that rare algorithm can effectively process the social order trend for decision makers to formulate the comprehensive method to decrease the criminal rate[4]. In this paper we mainly deal with a novel algorithm which analyze social order trend from criminal statistical data.

Attenuation Based Profile Vector

In terms of individual, each occurrence of a crime is certainly a (p, 1-p) binary event, which is a common probability event in reality[5]. In addition, generally if a certain event possesses a characteristic that its occurrence is "rare", a Poisson stream may be used to approximate it[6]. Thus, we can verify that general crime events observe Poisson distribution and urban criminal event observes same distribution, then we may use some values derived from distribution property to depict the urban public order status.

Composite distribution. Considering the great diversity of crime types, we have to aggregate some crime types in order to obtain a better statistics distribution. Since that the sum of a series Poisson distribution also observes a Poisson distribution[7], we attempt to use Poisson distribution means to represent the urban crime occurrence property. Nevertheless, as a profile value which can indicate the implicit rule of urban crime, it must represent the fluctuation rule of present status as well as the affection of historic statistic data[8].Hence, for each component of profile vector, we process it as follows.

At the first time, we collect the mean value of a certain crime. We note it as w and draw a directed segment to represent it. After a specified time span which is defined in advance, we whirl the previous segment counterclockwise with angle a, the project of it on its original direction is:

$$w' = w \cdot \cos(\partial)$$

When second measure comes, rotates the historic vector and projects it, combine with the present statistics data to generate new profile vector.

$$P_i(x) = \sum_{j=1}^{n} \frac{\zeta^{j-1} \cdot w}{(1+\zeta)^j} \tag{1}$$

Within it, ζ is system attenuation factor. On condition that system statistic interval is d (day), we set s (day) as the semi-attenuation circle, then attenuation factor is:

$$\zeta = e^{\frac{d(\ln 0.5)}{s}}$$

Then a new status component is generated. After a certain number of time span passed, we get a profile vector component contain enough historic vector component. Certainly, to the extreme it can represent the property of present statistic measure and rationally consider the effect of historic measure[8].

Finally, we can define：

$$P(x) = (P_1(x), P_2(x), \cdots, P_n(x))$$

as the system profile vector.

Experiments and Results. According the algorithm model above, we design a series of experiment on the urban crime event:

(1) For every 5 days, execute a Chi-square testing[9] on near 30 days statistics date to obtain its distribution mean.

(2)At the basis of 30 days, execute a Chi-square testing on near 30 days statistic date to obtain its distribution mean, every new testing span slide the time window with 10 days.

We make experiments on the profile vector algorithm, experiments results are shown in Fig.1. Certain line 3 can better represent the historic effect and be not sensitive to present statistic data.

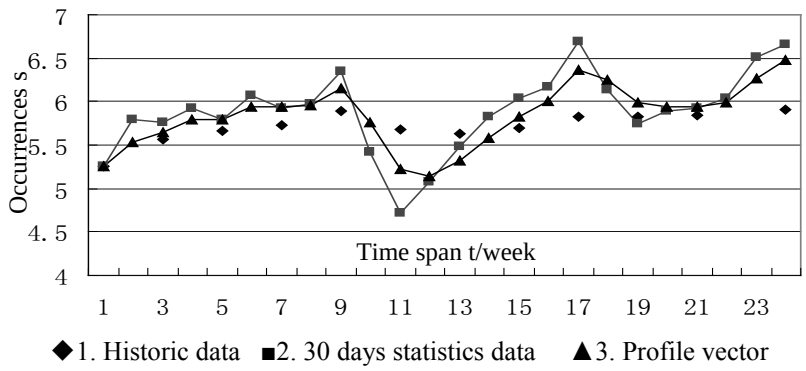

Fig.1 Three classes of means. Line 1 is historic accumulative data, line 2 is 30 days statistics data, line 3 is profile vector.

Consequently, we can draw a conclusion that we can use a series of attenuated historic profile vector and present statistics combine into a new profile vector. The vector can better represent the social public order status and is not sensitive to present statistic data, towards many type of crime we classify raw data into optimized based on Poisson distribution.

System detection algorithm

In this section, how to depict the comprehensive status of social public order is the first problem we have to solve. We propose a flow process as follows:

- System initialization phase

- Anomaly detection phase
- Recovery phase

Derivation of System detection algorithm. Assume a certain complex system C which contains n parameters $M=(m_1,m_2,\cdots,m_n)$ and M can be described by a non-negative integral within each fixed time span t, after an preset enough long time γ:

$$M=(\vec{m}_1,\vec{m}_2,\cdots,\vec{m}_n=\begin{bmatrix} m_1(1) & m_2(1) & \cdots & m_3(1) \\ m_1(2) & m_2(2) & \cdots & m_3(2) \\ \cdots & \cdots & \cdots & \cdots \\ m_1(j) & m_2(j) & \cdots & m_3(j) \end{bmatrix}$$

Within it, $j=\mathrm{int}(\gamma/t)$, $m_i(j)=c_{ij}$.

Assuming matrix $M'=(\vec{m}'_1,\vec{m}'_2,\cdots,\vec{m}'_n)$, its factor $\vec{m}'$ contains q lines, then the relation between M and $\vec{m}'$ can be described as follows:

$$\begin{cases} m'_1(x)=m_{last\_i}(x) \\ m'_q(x)=m_{last\_i-q+1}(x) \end{cases} \tag{2}$$

Provided random variable $\vec{m}_i$ observes Poisson distribution, we can generate N-dimensioned random variable $Mean=(\vec{m}ean_1,\vec{m}ean_2,\cdots,\vec{m}ean_n)$, which generate a new line when M' process an update:

$$mean_{i,j}=E(\vec{m}_j(x)) \tag{3}$$

Def. 1: Vector $profile=(mean'_{last,j}(x))$ is called profile vector of complex system C, if the following requirements are met:

$$mean'_{k,j}(x)=E(\vec{m}'_j(x)) \quad \text{k=1}$$

$$mean'_{k,j}(x)=\frac{mean_{k,j}(x)+\zeta\cdot mean'_{k,j-1}(x)}{1+\zeta} \quad \text{k>1}$$

ζ is a constant, and $0.9\le\zeta<1$, $Mean'$ is call profile set of complex system C.

Def. 2: N-dimensioned random variable D is called anomaly index if following requirements are met:

$$d_i=(0)\ ,if\ i=1$$

$$d_i=mean_i-mean'_{i-1},\ if\ \ i>1$$

Since every column of D is difference value between the mean value of a Poisson distribution variable and the weighted value of which, obviously d_i observes a Normal distribution. We can draw a conclusion that:

$$\bar{d}\sim N(0,1)$$

$$\frac{(n-1)S^2}{\sigma^2}\sim\chi^2(n-1)$$

And they are independent of each other, then

$$\bar{d}/\sqrt{\frac{(n-1)S^2}{\sigma^2(n-1)}}=\frac{\bar{d}}{S/\sqrt{n}}\sim t(n-1)$$

Now making a hypothesis: $H_0\ :\ \mu_d\ =\ 0$

Note sample mean value and sample variance value of $d_1,d_2,\cdots,d_n$ as $\bar{d},S^2$ separately, then:

$$\bar{d}=\frac{1}{n}\sum_{i=1}^{n}d_i$$

$$S^2=\frac{1}{\text{n -1}}\sum_{i=1}^{n}(d_i-\bar{d})^2$$

Now setting a small positive number $\alpha > 0$, we suppose Eq.3.5 as anomaly detecting formula of complex system C and Eq.3.6 as referenced anomaly detecting formula.

$$|t| = \left| \frac{\bar{d}}{s/\sqrt{n}} \right| \geq t_{\partial/2}(n-1) \tag{6}$$

$$p = P(t \geq t_{\partial/2}(n-1) \mid \mu_d = 0) \tag{7}$$

System detection algorithm process. In general, the Composite Statistic Profiling-Vector Hypothesis test Algorithm we supposed can be described as follows:

Composite Statistic Profiling-Vector Hypothesis test Algorithm

Input: attenuation period ζ,

parameter statistical value $measure_i$,

detection segment *ds*, statistical segment *ts*

Output: anomaly index of complex system *C*

Method:

```
while .t.
   line=countline(M);
   insertline(M,measure);   // insert a new line into M
      for each measure
         if line<ts
            window_start=0;
             window_end=line;
         else
            window_start+=1;
             window_end+=1;    //slide windows
         endif
        mean_value=getmean(M,window_start,window_end);    // compute mean value
      endfor
   insertline(mean,mean_value);       // insert a new line
   if first_line_of_mean
         insertline(mean1,mean_value);
     else
        mean1=(mean+ζ·mean1)/(1+ζ) ;
   endif
   for each end_of_ts
      d=mean-mean1 ;
       insertline(D,d);
   end for
   s=variance(d) ;   //count statistical standard variance
   d1=average(d) ;
    t=abs(d1).sqrt(n).s
   insertline(T,t);   //record the testing anomaly value
   if t>threshold_value
         if test_fail     // refuse 0 assumption
          send_abnormal(serious_alert); // serious alert
         else
          send_abnormal(warning);    //warning
```

```
        endif
    else  send_normal();
    endif
    return t;
  endwhile
```

As far as the process of anomaly value is concerned, there are two opinions:

(1) to eliminate the effect system property vector upon system property value to make system property stable;

(2) since the system anomaly value implies the fluctuation of social status, it should be remained to affect and update social order vector in time.

Considering the fluctuation characteristics of social order status against relatively stable status, its value may not be constant but changed slowly. This kind of movement derived from the anomaly status of complex system, as a result the anomaly status should remain. Of course, anomaly rises from statistical error have to be eliminated.

Experiments and conclusions

In experiment environment, we collect the statistical data of criminal cases in a certain city of China in 2009. Considering that the algorithm bases upon historical data, we abstract last 20 weeks data as our detection data, the others is treated as the training data of the model.

Setting ζ=0.9330, experiments results are showed as Fig. 2. Setting alert threshold as 0.70 and system can receive 3 alerts. Comparing with initial data we collect, urban criminal statistics data changes sharply when alert signal received. Consequently we can draw a conclusion that analysis prediction results are scientific and accurate.

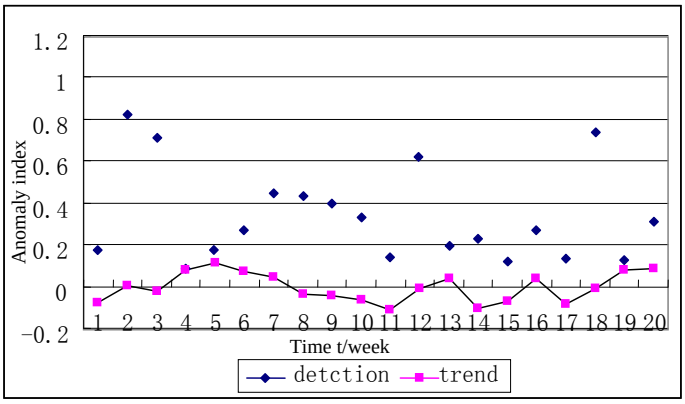

Fig. 2 Experiment results and comparison. Line 1 is the detection result of Algorithm, Line 2 is statistical Criminal data trend.

There are two methods to adjust the accuracy of system. Adjusting the width of statistical span can affect the sensitivity of system. Certainly the sensitivity of warning anomaly detection may slightly decrease. To sum up, in order to obtain more scientific and accurate results, we should appropriately adjust system parameters according historic collecting data and expert advices.

In addition, according to the character of the complex social system, it uses many layers of nonlinear processing units for feature extraction and transformation, the processing algorithms may be supervised or unsupervised and applications include pattern recognition and statistical classification, and it is based on the learning of multiple levels of features or representations of the data, then optimized algorithm based on deep learning may be applied to accurate the detection efficiency and alert correction in the future researching.

Acknowledgments

This Work is partially supported by the humanities and social science research projects of Hubei Province (2012G121) and the universities youth science and technology innovation team project of Hubei Province (T201222). It is also supported by Cooperative Innovation Center of Digital Data Forensics.

References

[1]Lin Jianhui, Huang Tianshu. Analysis and realization of a crime prediction system based on statistics algorithm. Journal of Information and Computational Science. 4(2007), 1045-1051

[2]Chandra B, Gupta M. A multivariate time series clustering approach for crime trends prediction. IEEE International Conference on Systems, Man and Cybernetics. (2008), 892-896

[3]Kelli Crews. Bayesian Network Modeling of Offender Behavior for Criminal Profiling. Proceedings of the 44th IEEE Conference on Decision and Control, and the European Control Conference. (2005),453-459.

[4]Pillai R.K.G. Simulation of Human Criminal Behavior Using Clustering Algorithm. International Conference on Computational Intelligence and Multimedia Applications. (2007), 105-109

[5]Kianmehr, Keivan. Effectiveness of support vector machine for crime hot-spots prediction. Applied Artificial Intelligence. 22(2008), 433-458

[6]Do Kim. Cyber Criminal Activity Analysis Models using Markov Chain for Digital Forensics. International Conference on Information Security and Assurance. 4(2008), 24-26

[7]Oatley, Giles C. Crimes analysis software: 'Pins in maps', clustering and Bayes net prediction. Expert Systems with Applications.25(2003), 569-588

[8]Lin Jianhui. An urban criminal statistic profile vector Algorithm. International Conference on Information Management, Innovation Management and Engineering. (2009) ,213-218

[9]Lin Jianhui, Huang Tianshu. Crime Information Analysis and Complex Social System Tendency Researches. Computer engineering and application.17(2011),246-248.