Quantitative analysis of terrorist attacks combined with multiple models

Qingkuo Li[a], Kun Li, Ziqing Zhang

University of Chinese Academy of Sciences, Beijing, China

[a]liqingkuo@iet.cn

Keywords: Neural network; prediction; k-means; sensitivity analysis; clustering

Abstract: The quantitative analysis of the data recorded in the terrorist attacks will help to improve the pertinence and efficiency of the fight against terrorism. This paper comprehensively uses neural network, PCA, k-means, wavelet analysis and other models to determine the level of terrorist attacks, and predicts and analyzes the future counter-terrorism situation, and proposes a process model for terrorists to choose targets, and has received terrorist attacks. The level of hazard of the incident and the suspicion of unknown terrorists and the pre-judgment of future counter-terrorism.

1. Introduction

Due to the serious consequences of terrorist attacks in various parts of the world, the assessment of terrorist violence, the arrest of terrorists and the prevention and control of violent terrorist attacks have gradually received attention from all countries. However, due to the hidden and sudden nature of terrorist acts, it is difficult to The terrorists carried out rapid arrests and could not promptly and accurately issue warnings of terrorist violence to the masses, which made the people live in great fear. Therefore, they can analyze the uncertain factors in the process of terrorist activities, and timely and accurately reflect the threat assessment of terrorist activities. Necessary.

The problems to be solved in this paper are as follows：

1) Establish a quantitative grading model based on data analysis: the harmfulness of terrorist attacks is not only related to casualties and economic losses, but also has a great relationship with the timing, region, and targeted objects, so subjectively enforced grading, It is difficult to accurately assess the hazards of terrorist attacks. Therefore, based on GTD and other relevant information, combined with information processing technology, we will establish a quantitative grading model to classify the degree of terrorist damage from high to low, and use the evaluation model to identify the hazards of the past 20 years. The ten most serious terrorist attacks.

2) Determining the makers of terrorist attacks: Many terrorist attacks have not identified the perpetrators. If a number of terrorist organizations or individuals can be linked to each other, it will help to identify hidden terrorists as early as possible in 2015-2016. There have been no attacks on organizations and individuals claiming responsibility, applying mathematical modeling to identify event makers, and ranking suspects' suspects.

2. Problem Analysis

1) This paper proposes to establish a quantitative grading model based on data analysis to assess the hazard of terrorist attacks. The quantitative grading model should meet the following requirements: the harm of terrorist attacks should be related to factors such as casualties, economic losses, timing, geographical location, objects, etc., and can be evaluated by quantitative grading. For this problem, we must first process the data, eliminate the default more serious values, and use the correlation to find the relationship between different variables. For the factors that are more important to the influencing factors, the default data should be filled, due to the amount of data. It is relatively large, so the neural network model is used to train the data and fill in the default values. Model test: In order to verify the rationality of filling the data, the test is performed using the data without the default value in the original data table. Finally, a comprehensive evaluation model based

  DOI: 10.25236/mmmce.2019.121

on principal component analysis (PCA) was established to quantitatively evaluate the terrorist attacks.

2) Identify the suspects or organizations involved in the terrorist attack, and then try to find out the terrorist attacks related to them, and classify them accordingly, and on the basis of the first question, the terrorist attacks involving terrorist organizations (or individuals) The event is rated accordingly, and then the different levels are weighted to obtain the level of hazard for different terrorist organizations and individuals. The model established for the problem needs to meet the following requirements: the established model needs to be able to find the possibility of internal relationship for the integration of information of the same organization and individual under different time and space relationships, and by rationally applying the model of the first question, Rating the hazards of the organization or individual. To solve this problem, we must first determine the classification criteria of unknown terrorist organizations, locate their numbers and names, and then accurately assess the magnitude of their hazards.

3. Symbol Description

Table 1. Symbol Descriptio

Symbol	Description
x_{ij}	Expressed as the jth dimension of the ith sample
θ	Neuron domain value
η	Learning rate
n	Number of input units in the neural network
m	Number of output units in a neural network
σ_i	Relative error of the first sample
nk_i	Observations of total deaths from the first sample
nk_i'	Estimated total number of deaths for each sample
pt	Degree of property damage
κ_i	Square sum of squared observations of three variables
$\bar{\sigma}$	Average relative error of all test samples
$\tilde{\alpha}_{ij}$	Normalized value of the first indicator of the first evaluation object
d_{ij}	Euclidean distance from the predicted point to the center of the space

4. Data processing

4.1 Preprocessing of data

For the purpose of this paper, in order to better establish the hierarchical quantitative model, the original data needs to be preprocessed. The main purpose of data preprocessing is twofold:

- Reduce data dimensions and reduce useless or duplicate data.
- Correct the wrong data or missing data to ensure the quality and accuracy of the data.

The data recorded in this article is more comprehensive, but some data mining process does not require some data, and in the process of data collection and recording, there are missing, abnormal, irrelevant data and even conflicting fields. . In applying the cleanup method we define the following principles:

- Remove repetitive information
- Remove the negligible information
- Reasonable selection of related information
- Data conversion (data standardization)

For the preprocessing in this paper, the characteristics of the terrorist attacks that exclude the missing values of the terrorist attacks and the classification predictions are included. Among them, 7217 data records in Annex 1 are discarded because there are too many missing values, so the

research value is finally obtained. There are 106,969 data.

Before data pre-processing, first of all, the research and discussion of the overall data of more than 100,000 terrorist attacks will be carried out to obtain macroscopic overall results, which is conducive to subsequent research.

4.2 Normalization of data

After the data is preprocessed, each terrorist attack is related to its corresponding variable. However, different variables are unitized. The error of subsequent research caused by the inconsistent data size between different dimensions is very large, so the data should be normalized and all the data needed should be used. It is dimensionless and its characteristic values are specified to a specific range. However, it should be noted that the normalization of data should follow the principle that "the internal relative gap of the same indicator is different, the relative gap between different indicators is uncertain, and the normalized maximum is equal", so the positive and negative indicators are used to return the data. In this paper, the MAX-MIN normalization method is adopted. The specific expression is:

$$\begin{cases} \tilde{x}_{i,j} = \dfrac{x_{i,j}}{\max x_j} & \text{Positive} \\ \tilde{x}_{i,j} = 1 - \dfrac{x_{i,j} - \min x_j}{\max x_j} & \text{Reverse} \end{cases} \tag{1}$$

4.3 Feature selection of high dimensional space

After pre-processing, because there are many features in the many feature values that have little relationship with the judgment of the title category, there is information overlap between the feature values, the interference judges the main information, and more importantly, when the sample dimension Too high, the learning process of the data sample points is lengthened, which makes the results difficult to converge. Not only does not obtain important clue relationships, but also increases the computational cost. Therefore, feature selection should be performed on high-dimensional space.

At present, the better high-dimensional feature selection method is the FS model. The specific method is to effectively judge the data collection and discard the variable items that are useless or have less effect on the discrimination.

5. Model establishment and solution

5.1 An important feature variable filling model based on BP neural network

A neural network is a complex network system in which neurons are connected to each other to perform information parallel processing and nonlinear transformation by simulating the transmission of information in the human brain. BP (Back Programming Forward Feedback) neural network is a multi-layer feedforward network, which can learn and store input-output mapping relationships. It does not need to establish mathematical equations. It is a commonly used neural network model. It is the BP neural network.

The n-dimensional input of a neuron can be represented by a column vector X:

$$\mathbf{x} = (x_1; x_2 ...; x_n) = \begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{pmatrix} \tag{2}$$

The neuron output can be represented by a column vector W:

$$\mathbf{w} = (w_1; w_2; ...; w_n) = \begin{pmatrix} w_1 \\ w_2 \\ ... \\ w_n \end{pmatrix} \tag{3}$$

The output is:

$$y = f(\mathbf{x}^T \mathbf{w} - \theta) = f(\sum_i w_i x_i - \theta) \tag{4}$$

$$f = \text{sgn}(x) = \begin{cases} 1, x \geq 0; \\ 0, x < 0; \end{cases} \tag{5}$$

Where θ is the domain value of the neuron.

The weights and thresholds can be obtained by learning. If θ is fixed to -1.0 when the dummy node is input, its weight is set to w_{n+1}. At this time, the learning of weights and thresholds is unified into weight learning. The input vector is:

$$\hat{\mathbf{x}} = (x_1; ...; x_n; -1) = \begin{pmatrix} x_1 \\ \vdots \\ x_n \\ -1 \end{pmatrix} \tag{6}$$

The weight vector is:

$$\hat{\mathbf{w}} = (w_1; ...; w_n; \theta) = \begin{pmatrix} w_1 \\ \vdots \\ w_n \\ \theta \end{pmatrix} \tag{7}$$

The output is:

$$y = f(\hat{\mathbf{x}}^T \hat{\mathbf{w}}) \tag{8}$$

For each iteration, the weight adjustment method is expressed as:

$$\begin{aligned} w_i &\leftarrow w_i + \Delta w_i \\ \Delta w_i &= \eta(y - \hat{y}) x_i \end{aligned} \tag{9}$$

This is called the learning rate. When the learning rate is larger, each iteration will be larger, that is, the step size is larger, which can speed up the convergence, but the convergence may be worse, so the learning rate needs to be comprehensively considered.

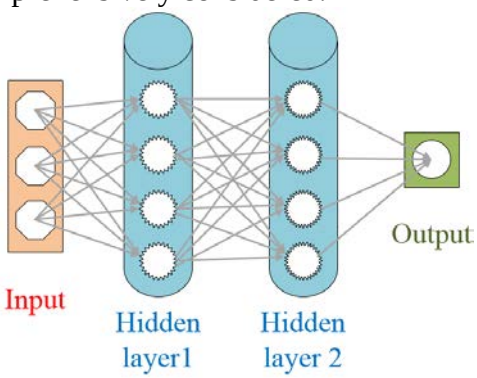

Figure 1 Neural network structure

The neural network structure is shown in Figure 1. The left side is the input layer, the middle two layers are the hidden layers, and finally the output layer. It can be seen that each layer has multiple neurons. Because the network has a strong topology, he can solve the input signal and output signal

through the connection of each neuron without relying on a specific function. The relationship, and because of the complex topology, has a strong fault tolerance. The working process is that when the input signal enters the input layer, the signal propagates forward, and through multiple hidden layers up to the output layer, the output will be compared with the expected value. If the error exceeds our required range, the error will be backpropagated. In this way, the weight coefficient in each layer is corrected to minimize the error of the output signal and the expected value. When the weight is iteratively modified, an ideal output result which is within a limited range from the expected value is obtained, and the training is performed at this time. The model is the model we need.

The method of determining the number of cells in the hidden layer uses the following three methods [2] to select the optimal number of hidden layer units:

$$\sum_{i=0}^{n} C_{n_1}^{i} > k \tag{10}$$

Where k is the number of samples, n1 is the number of hidden layer units, n is the number of input units, and i is a constant between [0, n];

$$n_1 = \sqrt{n+m} + a \tag{11}$$

Where n1 is the number of cells in the hidden layer, n is the number of input cells, m is the number of output cells, and a is a constant between [1, 10];

$$n_1 = \log_2 n \tag{12}$$

5.2 Quantitative evaluation model based on principal component analysis (PCA)

Principal component analysis is to optimize the multivariate cross-section data table under the principle of minimizing the loss of data information, that is, to reduce the dimensionality of high-dimensional variable space. The main purpose of principal component analysis is to use less variables to explain most of the variation in the original data, and to convert many of our highly relevant variables into variables that are independent or unrelated to each other. It is usually the selection of a few new variables that are less than the original variables and can explain the variation in most of the data, the so-called principal components, and a comprehensive indicator for interpreting the data. Thus, principal component analysis is actually a dimensionality reduction method.

- Model establishment

Step 1: Processing raw data

Let the value of the jth indicator of the i-th evaluation object be a_{ij}. The formula for converting each indicator to a standardized value is:

$$\tilde{a}_{ij} = \frac{a_{ij} - u_j}{s_j}, i = 1, 2, ..., n; \; j = 1, 2, ..., m, \tag{13}$$

Among them: $u_j = \frac{1}{n}\sum_{i=1}^{n} a_{ij}, s_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(a_{ij} - u_j)^2}, j = 1, 2, ..., m,$ there are:

$$\tilde{x}_j = \frac{x_j - u_j}{s_j}, \; j = 1, 2, ..., m \tag{14}$$

Step 2: Calculate the correlation coefficient matrix

The correlation coefficient matrix $R = (r_{ij})_{m \times m},$ has:

$$r_{ij} = \frac{\sum_{k=1}^{n} \tilde{\alpha}_{ki} \cdot \tilde{\alpha}_{kj}}{n-1}, i, \; j = 1, 2, ..., m \tag{15}$$

among them:. $r_{ii} = 1, r_{ij} = r_{ji}$

Step 3: Calculate eigenvalues and eigenvectors

The eigenvalue $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0$ of the correlation coefficient matrix R is calculated, and the

corresponding eigenvector $u_1, u_2, ..., u_m,$, wherein $u_j = [u_{1j}, u_{2j}, ..., u_{mj}]^T$ is composed of eigenvectors to form m new indicator variables:

$$\begin{cases} y_1 = u_{11}\tilde{x}_1 + u_{21}\tilde{x}_2 + ... + u_{m1}\tilde{x}_m \\ y_2 = u_{12}\tilde{x}_1 + u_{22}\tilde{x}_2 + ... + u_{m2}\tilde{x}_m \\ ..... \\ y_m = u_{1m}\tilde{x}_1 + u_{2m}\tilde{x}_2 + ... + u_{mm}\tilde{x}_m \end{cases} \quad (16)$$

Where y_1 is the first principal component, y_2 is the second principal component, and y_m is the mth principal component.

Step 4: Calculate the comprehensive evaluation value

The p principal components are selected, and the information contribution rate and the cumulative contribution rate of the feature value $\lambda_j (j = 1, 2, ..., m)$ are first calculated. And the component contribution rate is:

$$b_j = \frac{\sum_{k=1}^{p} \lambda_k}{\sum_{k=1}^{m} \lambda_k}, j = 1, 2, ..., m \quad (17)$$

The formula for calculating the cumulative contribution rate of principal component $y_1, y_2, ..., y_p$:

$$\alpha_p = \frac{\sum_{k=1}^{p} \lambda_k}{\sum_{k=1}^{m} \lambda_k} \quad (18)$$

Step 5: Determine the evaluation model

From the component matrix, we can find five principal components, and then construct a principal component comprehensive evaluation model with the weight of the five principal components, namely:

$$Z = \sum_{j=1}^{p} b_j y_j \quad (19)$$

- Problem solving

According to the above evaluation model, list the top ten terrorist attacks with the greatest degree of harm in the past two decades, and map their corresponding attack cities and latitude and longitude information one by one. The specific data are as follows:

Table 2. Top Ten Terrorist Attacks in the Near Twenty Years

Number	GTD	country	provstate	latitude	longitude
1	200109110004	United States	New York	40.697132	-73.93135
2	200109110005	United States	New York	40.697132	-73.93135
3	199808070002	Kenya	Nairobi	-1.28518	36.821107
4	201710140002	Somalia	Banaadir	2.059819	45.326115
5	200409010002	Russia	North Ossetia-Alania	43.191626	44.541763
6	201404150089	South Sudan	Unity	9.259689	29.800148
7	201406100042	Iraq	Nineveh	36.407394	42.964626
8	200901170021	DROC	Orientale	3.292111	29.168137
9	201310200012	South Sudan	Jonglei	7.181962	32.356095
10	201310200013	South Sudan	Jonglei	6.89375	31.36775

Table 3. Top Ten Terrorist Attacks in the Near Twenty Years

Number	GTD	country	provstate	latitude	longitude
1	200109110004	United States	New York	40.697132	-73.93135
2	199808070002	Kenya	Nairobi	-1.28518	36.821107
3	201710140002	Somalia	Banaadir	2.059819	45.326115
4	200409010002	Russia	North Ossetia-Alania	43.191626	44.541763
5	201404150089	South Sudan	Unity	9.259689	29.800148
6	201406100042	Iraq	Nineveh	36.407394	42.964626
7	200901170021	DROC	Orientale	3.292111	29.168137
8	201310200012	South Sudan	Jonglei	7.181962	32.356095
9	200403210001	Nepal	Central	27.959441	84.895897
10	201712170021	South Sudan	Lol	8.468314	25.679393

As can be seen from table 2, for the same GTD number, look up the original table, the description of the event in the table is basically the same, that is, the same terrorist attack occurred on the same day, if such an event is considered to be the same horror In the case of the attack, then Num1 and Num2 in the above table can be considered as the same terrorist attack. Num9 and Num10 can also be considered as the same terrorist attack, so the arrangement of the top ten terrorist attacks will change slightly, as shown in Table 3.

5.3 Sensitivity analysis

The sensitivity analysis is to see if the obtained model is stable to some disturbances of the input variable. When the given input variable is disturbed, the model can still output the correct result, indicating that the model has good anti-interference ability. Or the solution obtained when the range of these variables changes will not change. This is a key parameter for evaluating the model built.

We conducted a sensitivity analysis of the model of the threat level of event evaluation built in the first question. The analysis process randomly extracts a certain number of the five types of events, and we manually apply different degrees of interference to a certain factor (column) of these events, and see that this interference has multiple effects on the final output. . The results show that even if a large interference is imposed on the strong correlation factor, the impact on the final predicted score is small, and it will not affect the grade evaluation. It will not change the final threat evaluation level of each event, indicating that the model is resistant. The interference performance is very strong.

When the number of deaths increases by a different percentage, the impact on the scores for different threat level events is as follows:

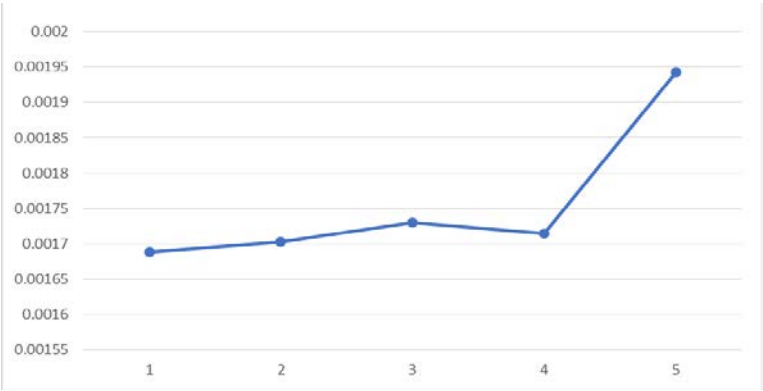

Figure 2 Impact of a 20% increase in deaths on different threat level event scores

As shown in the above figure, when the number of deaths increases by 20%, the impact on the evaluation scores of the five threat levels is very small, all around 0.18%, but the impact on different events is slightly different. The most influential is the five types of events. The impact is the largest with an amplitude of 0.195%. There is almost no evaluation of the rating, and no evaluation of the event has changed. It can be seen that the stability of the model is good at this time.

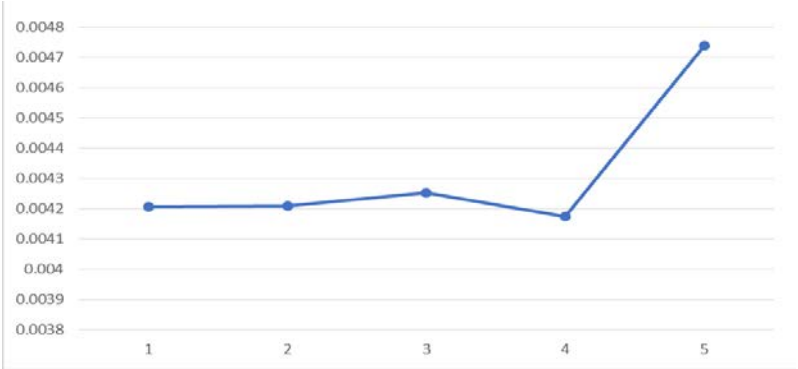

Figure 3 Effect of a 50% increase in deaths on different threat level event scores

It can be seen from the above figure that when the number of deaths increases by 50%, the overall impact on output increases to about 0.43% on average, and the most affected category is still Category 5, which is about 0.475%. In the end, the judgment of the event evaluation level has not changed.

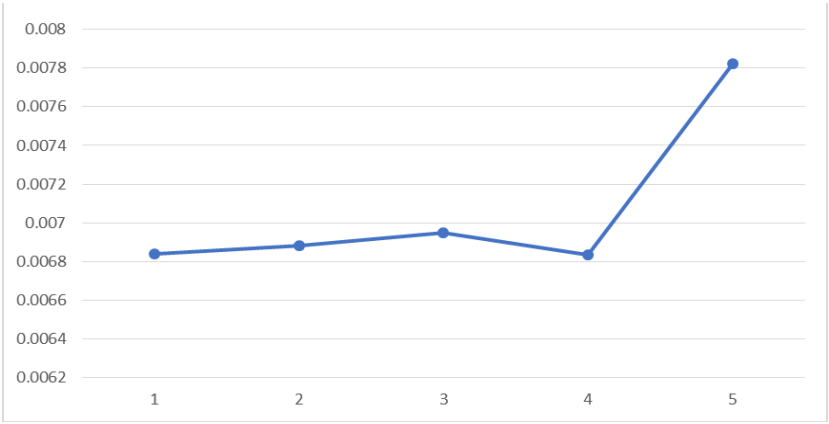

Figure 4 Effect of 80% increase in deaths on different threat level event scores

As can be seen from the above figure, when the number of deaths increased by 80%, the overall impact on output increased to an average of 0.72%, and the most affected category was still Category 5, which was around 0.78%. In the end, the judgment of the event evaluation level has not changed. Therefore, in general, when the number of deaths is very disturbing, the most affected ones are Category 5 events, but the overall situation does not have a big impact on the results.

References

[1] Kaili Zhou, Yaohong Kang. Neural network model and its MATLAB simulation program design [J]. 2005.

[2] Shen Huayu, Wang Zhaoxia, Gao Chengyao, et al. Determination of the number of hidden layer elements in BP neural network[J]. Journal of Tianjin University of Technology, 2008, 24(5): 13-15.

[3] Zhou Aihua, Zheng Yingping, Wang Lingqun. Review of medical data mining[J]. Chinese Journal of Medical Practice, 2005.

[4] Wang Jinghua, Xu Huiming, Li Yong, et al. A power engineering evaluation method based on principal component analysis: CN105631236A[P]. 2016.

[5] Liu Dehai, Zou Huawei, Bao Xueyan. Multi-stage repeated game model of terrorist attacks with long-term characteristics and government anti-terrorism[J]. Journal of University of Electronic Science and Technology of China(Social Sciences Edition, 2016, 18(3):19-23.

[6] Wang Yaonan. Remote Sensing Image Classification Based on Wavelet Neural Network[J]. Journal of Image and Graphics, 1999, 4(5): 368-371.