Hazard Grading Model of Terrorist Attack Based on Machine Learning

Yu Jun
School of Computer Science and Technology
Xi'an Technological University
Xi'an 710021, Shaanxi, China
e-mail: yujun@xatu.edu.cn

Hu Zhiyi
Institute of Engineering Design
Army Academy of PLA
Beijing, 100042, China

Xian Tong
School of Computer Science and Technology
Xi'an Technological University
Xi'an, 710021, Shaanxi, China

Liu Yutong
Engineering Design Institute
Army Academy of PLA
Beijing, 100042, China

***Abstract*—In this paper, there is no unified grading standard for the harm of terrorist attacks. A classification model of terrorist incidents based on machine learning is proposed. First, the data related to the hazard in the Global Terrorism Database (GTD) is extracted and preprocessed. Secondly, the data is extracted by principal component analysis, and all events are aggregated into 5 by K-means clustering. Again, the entropy method is used to calculate the weighting coefficient of each indicator, and the comprehensive score of the hazard of each type of terrorist attack is calculated. Finally, the scores are divided into 1-5 levels of hazard grading models in order of high to low. The results show that the hazard grading model can scientifically and objectively quantify terrorist attacks.**

Keywords-Terrorist Attacks; Hazard; Hierarchical Model; Principal Component Analysis; K-Means Clustering; Entropy Method

I. INTRODUCTION

A terrorist attack is an aggression committed by an extremist or organization that is not in conformity with international morality and is directed against, but not limited to, civilians and civilian installations. It not only has great destructiveness and destructive power, but also directly causes huge casualties and property losses. It also brings tremendous psychological pressure to people, causing a certain degree of turmoil in society and greatly hindering economic development. Global terrorism is a phenomenon of public interest, and everyone is directly affected by it. Therefore, anti-terrorism work is imminent. Big data is now the main source of counter-terrorism intelligence. The Global Terrorism Database (GTD) is the world's most comprehensive database of non-confidential terrorist attacks, containing more than 180,000 terrorist attacks, each containing at least 45 variables. An in-depth analysis of data related to terrorist attacks will help deepen people's understanding of terrorism and provide valuable information support for opposing terrorism and preventing terrorism. Data collection and preprocessing intelligence are the lifeblood of counter-terrorism work. Keeping reliable information in a timely manner can play an active role in combating terrorism and effectively curb the spread of terrorism[2].

Grading catastrophic events (such as earthquakes, traffic accidents, meteorological disasters, etc.) is an important task of social management. The usual grading generally adopts a subjective method, and the authority stipulates the grading standard. The harmfulness of terrorist attacks depends not only on the two aspects of casualties and economic losses, but also on the timing, geography, and targeted objects. Therefore, it is difficult to fully reflect these factors. The

hazard grading of terrorist incidents can clearly define the future attacks, and different levels of events correspond to different treatments. This will not only help the management of social security, but also avoid unnecessary waste of manpower and property.

Combined with big data processing technology, this paper establishes a hierarchical model based on PCA algorithm, K-meas clustering algorithm and entropy method. First, 14 evaluation indicators related to the hazard of the event were selected to preprocess the

DOI: 10.21307/ijanmc-2019-051

existing data. Secondly, the PCA method was used to reduce the index from 14 dimensions to 4 dimensions, and the reduced dimension vector was obtained by the clustering algorithm. Gather into 5 categories, you can get the category corresponding to each event. Finally, using the entropy method to score the hazard of each event and according to the average hazard score of each class. According to the degree of harm from high to low levels 1 to 5. A hazard grading model of terrorism events is obtained with a hazard rating of 5.

II. DATA PREPROCESSING

In this paper, the hazard grading model of terrorism events data is established from some important fields of the GTD original database. The selected data handling requires missing value processing, conversion of characters to numeric values and numerical processing.

A. Important field selection

The Important field of hierarchical is pointed out by *the World Anti-Terrorism Incident Research. T*he Terrorism Hazard Classification Model Data Table has selected the following 14 fields from GTD, as shown in Table 1.

TABLE I. THE SELECTED FIELD TABLE

Field	Description
extended	Whether it is a continuous event
latitude	latitude
longitude	longitude
success	Successful attack
suicide	Suicide attack
nkill	Total number of deaths
propextent	Degree of property damage
nwound	Total number of injuries
country	country
region	area
city	city
attacktype	Attack type
targtype	Target/victim type
weapontype	Weapon type

B. Missing value processing

In the selected field, Python's function *DataFrame. dropna* can delete rows or columns with null values, and retain all data that is not empty. Then the character field needs to be converted to a numeric field.

C. Converting character fields to numeric fields

The character field that need to be converted is as follows:

1) Eventid: Events in the GTD are numbered with 12 digits. The first 8 digits are recorded in the format "yyyymmdd". The last 4 digits calibrate the serial number of the day, e.g. 0001, etc.

2) Country: According to *the developed economies* assessment standards recognized by the United Nations, 168 countries are divided into developed and underdeveloped countries. Since terrorist attacks are more harmful to developed countries, the relevant assignments are shown in Table 2.1.

3) Region: Count the frequency of terrorist incidents in each region and assign the frequency to regional indicator values.

4) City: The world city is divided into three levels: the capital, the provincial capital, and other cities. Since the terrorist attacks are more harmful to the political and economic centers, the relevant assignments are shown in Table 2.1.

5) Attack type: Counting the frequency of occurrence of 9 types of attacks, and assigning the frequency to the attack type indicator value.

6) Weapon type: Counting the frequency of occurrence of 13 weapon types, and assigning this frequency to the weapon type indicator value.

7) Targtype: Counting the frequency of occurrence of 22 target types, and assigning this frequency to the target type indicator value.

TABLE II. THE STATE AND CITY ASSIGNMENT

Index	assignment
developed countries	2
underdeveloped countries	1
the capital	3
the provincial capital	2
other cities	1

D. Numerical processing

In the original GTD database, the nkill field includes the number of all victims and terrorists who directly caused death from terrorist incidents. We use only requires the number of victims and does not require the death toll of terrorists. Therefore, the number of victims is obtained by subtracting the number of terrorist deaths (nkiller) from the total number of deaths.

III. TERRORIST ATTACK HAZARD CLASSIFICATION MODEL

In this paper, the PCA algorithm, K-means clustering algorithm and entropy method are used to classify the terrorist attacks. The process of building a hierarchical model is divided into four steps:

1) The 14 indicators with greater influence is standardized by PCA algorithm. We construct a 14-dimensional matrix, and then reduce the matrix from 14 dimensions to 4 dimensions.

2) The K-means algorithm is used to cluster all the terrorist events in the matrix into five major categories, i.e. five hazard levels.

3) Using the entropy weight method finds the weights of each of the 14 indicators, and then weighting and summing the 14 indicators of each event to obtain the score of the event. For each hazard level, finding the average score for all events is at that level.

4) Sorting by the average scores of the five hazard levels, We divide them into one to five grades from high to low. The higher score means the greater damage.

A. Using the PCA algorithm for dimensional reduction

Principal Component Analysis (PCA) extracts M-dimensional feature matrices from N-dimensional matrices. First, we calculates eigenvalues and eigenvectors of N-dimensional matrices. According to the order of PCA eigenvalues from large to small, we select the corresponding first M eigenvectors., and then obtain an N*M feature transformation matrix T. In this paper, *N=14*, *M=4.* The dimensionality reduction is completed.[6]

The order of PCA eigenvalues generated by 14 indicators from large to small is shown in Table 3.

TABLE III. CHARACTERISTIC VALUES CORRESPONDING TO THE INDICATORS

Indicators	Characteristic values
nkill	9.82022087e-01
nwound	8.06184462e-02
targtype	7.91122120e-03
country	5.20872985e-02
attacktype	4.84991077e-03
region	4.01240379e-02
suicide	2.66626688e+00
city	2.60031933e-02
longitude	1.84972981e+02
extended	1.63936354e+03
latitude	1.36725606e+03
propextent	1.06560032e-01
success	1.04574700e+02
weapontype	0.00000001e+00

In this paper, 98686 data is reduced by the PCA algorithm, i.e. the original 14-dimensional matrix $X = [x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}]$ is reduced to a 4-dimensional matrix $Y = [y_1, y_2, y_3, y_4]$. The corresponding contribution degrees of the 4-dimensional feature vectors are: 0.49, 0.42, 0.06, 0.03, and the sum is greater than 0.99. Therefore, the dimension-reduced matrix preserves most of the original data and can be directly used for clustering.

B. Using K-Means algorithm for Hazard classification

The main idea of the K-means clustering algorithm is to cluster a number of discrete data points with *k* centroids and divide them into *k* clusters to distinguish data points with less similarity. Sum of the squared error *(SSE)* is the objective function of clustering, and classify data points with similar similarity into one class. The method finally converges to the optimal solution by continuously updating the centroid attribution and centroid position of the data points[1]. The algorithm process is as follows:

1) We select 5 event objects as the initial cluster center.

2) We calculate the Euclidean distance from each event to each cluster center and assign this event to the nearest cluster.

3) After all the event assignments are completed, the five cluster centers are recalculated, and compared with the cluster center obtained in the previous calculation. If the cluster center changes, the Euclidean distance and the assigned category are recalculated.

4) When the cluster center does not change, the clustering result is directly output.

Calculate the cluster center to which each type of event belongs, as shown in Table 4.

TABLE IV. TABLE 4. CLUSTERING CENTER FOR EVENT CLASSIFICATION

type	X1	X2	X3	X4	numbers
0	2.4843	-16.3826	-1.3464	0.3081	63122
1	-3.3968	22.8297	-3.8782	0.0615	37848
2	825.778	873.697	28.9316	-104.59	2
3	13.8411	-127.794	19.7789	-2.7281	3500
4	-9.5985	63.3898	16.7324	-1.2382	9711

The formula for calculating each event category is as shown in Equations (1) to (6).

$$D_1 = \sqrt{(y_1 - 8256.783)^2 + (y_2 - 873.658)^2 + (y_3 - 28.915)^2 + (y_1 + 104.608)^2} \quad (1)$$

$$D_2 = \sqrt{(y_1 - 13.840)^2 + (y_2 + 127.794)^2 + (y_3 - 19.779)^2 + (y_4 + 2.728)^2} \quad (2)$$

$$D_3 = \sqrt{(y_1 - 2.484)^2 + (y_2 + 16.382)^2 + (y_3 + 1.346)^2 + (y_4 - 0.308)^2} \quad (3)$$

$$D_4 = \sqrt{(y_1 + 3.396)^2 + (y_2 - 22.829)^2 + (y_3 + 3.878)^2 + (y_4 - 0.061)^2} \quad (4)$$

$$D_5 = \sqrt{(y_1 + 9.598)^2 + (y_2 - 63.289)^2 + (y_3 - 16.731)^2 + (y_4 + 1.238)^2} \quad (5)$$

$$\min_i = \min\{D_1, D_2, D_3, D_4, D_5\} \quad (6)$$

Among them is $Y = [y_1, y_2, y_3, y_4]$ the feature component vector after dimension reduction by PCA algorithm. D_i is the Euclidean distance between the dimension vector and the five cluster centers. min_i is the minimum Euclidean distance, and i is the final event category.

C. Using entropy method for calculating weight coefficient

The entropy method is a mathematical method used to determine the degree of dispersion of an indicator. With the great degree of dispersion comes great impact of the comprehensive evaluation of the indicator. The entropy value can be used to determine the degree of dispersion of an indicator. The steps of calculating the weight coefficient by the entropy method are as follows:

1) We select 14 indicators of 98686 events, and use x_{ij} to indicate the index value of the i-th indicator in the j-th terrorist attack. $(i = 1,\ldots,98686;\ j = 1,\ldots,14;\ n = 98686;\ m = 14)$

2) Normalization of 14 indicators is Normalized processing. The absolute values of the 14 indicators are conversed into relative values. It has different representative meanings that the positive indicator and the negative indicator value (the higher the positive indicator value is the better), the lower the negative indicator value is the better), as shown in Equation (7) and Equation (8).

$$x'_{ij} = \frac{x_{ij} - \min\{x_{ij}, \ldots x_{nj}\}}{\max\{x_{1j}, \ldots, x_{nj}\} - \min\{x_{1j}, \ldots, x_{nj}\}} \quad (7)$$

$$x'_{ij} = \frac{\max\{x_{1j}, \ldots x_{nj}\} - x_{ij}}{\max\{x_{1j}, \ldots, x_{nj}\} - \min\{x_{1j}, \ldots, x_{nj}\}} \quad (8)$$

3) Calculating the proportion of the i-th event in the j-th index are shown in Equation 9.

$$p_{ij} = \frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}} \quad (9)$$

4) Calculating the entropy value of the j-th indicator, are shown in Equation 10.

$$e_j = -k\sum_{i=1}^{n} p_{ij} \ln(p_{ij}), e_j \geq 0$$
$$k = 1/\ln(n) \quad (10)$$

5) Calculating the information entropy redundancy are shown in Equation 11.

$$d_j = 1 - e_j \quad (11)$$

6) Calculating the weights of each indicator are shown in Equation 12.

$$w_j = \frac{d_j}{\sum_{j=1}^{m} d_j} \quad (12)$$

7) Calculating the hazard weighting value of each event are shown in Equation 13.

$$s_i = \sum_{j=1}^{m} w_j \cdot x_{ij} \tag{13}$$

The weighting factors for each indicator are shown in Table 5.

TABLE V. TABLE 5. WEIGHT COEFFICIENTS OF EACH INDICATOR

indicator	x1	x2	x3	x4	x5	x6	x7
Weight	0.25	0.01	0.26	0.15	0.17	0.08	0.01
indicator	x8	x9	x10	x11	x12	x13	x14
Weight	0.01	0.01	0.01	0.01	0.01	0.01	0.01

D. Hazard grading result

All events can be divided into five hazard levels by PCA and K-Means clustering. The hazard score of each event is obtained by entropy method, and the average value of the hazard score of each type of event is obtained. After sorting the average, the five hazard levels are shown in Table 6.

TABLE VI. TABLE 6. HAZARD GRADING RESULT

Hazard level	**Cluster category**	**Hazard level**
1	2	1766.7104
2	3	3.2596
3	0	0.6239
4	4	-2.6904
5	1	-0.8788

IV. CONCLUSION

In this paper, 14 categories related to hazard are selected from the Global Terrorism Database (GTD) for the hazard grading of terrorist attacks; after pre-processing the data used, through principal component analysis (PCA) The related data is used for feature extraction. The K-means clustering method aggregates all events into five categories. The entropy method calculates the weight coefficient of each indicator, and finally obtains the comprehensive score of the harm of each type of attack. According to the comprehensive scores of the five types of attacks, a graded to five-level classification model was obtained. This model quantifies the relevant data of past terrorist attacks, and the obtained model has objectivity. It is necessary to establish more detailed grading standards.

REFERENCE

[1] Sanjun Nie. Research on Counter-terrorism based on Big Data[A]. IEEE Beijing Section. Proceedings of 2016 IEEE International Conference on Big Data Analysis (ICBDA) [C]. IEEE Beijing Section: IEEE BEIJING SECTION Institute of Electrical Engineers Beijing Branch), 2016: 5.

[2] Strang, Kenneth David & Sun, Zhaohao. (2017). Analyzing Relationships in Terrorism Big Data Using Hadoop and Statistics. Journal of Computer Information Systems. 57. 67-75. 10.1080/08874417.2016.1181497.K. Elissa, "Title of paper if known," unpublished.

[3] Li Wei. Characteristics and Trends of Current International Terror and Anti-Terrorism Struggle [J]. Modern International Relations, 2007 (02): 22-27.

[4] Yu Yihan, Fu Wei, Wu Xiaoping. Privacy data metric and hierarchical model based on Shannon information entropy and BP neural network[J].Journal of Communications,2018,39(12)

[5] He Jing. Research and Analysis of Future Anti-terrorism Situation Based on Big Data[J]. Economic Research Guide, 2019(05): 186-187.

[6] Wang Qi, Li Xiaopei, Dong Xinyan. Classification model of wine grape based on principal component analysis[J]. China High-tech Zone, 2018(05): 218.

[7] Wang Chao, Yao Min, Fu Zhanzhan. Research on Emergency Classification Based on Fuzzy Comprehensive Evaluation[J].Software Guide, 2019(04): 149-15

[8] Lu Ronghui. Terrorism and Counter-Terrorism in the Context of Globalization [D]. Suzhou University, 2005.

[9] Wang Chao, Yao Min, Fu Zhanzhan. Research on Emergency Classification Based on Fuzzy Comprehensive Evaluation[J].Software Guide, 2019,18(04):149-152.

[10] Hou Wenjing, Jiang Xinxin, Wen Hong, Lei Wenxin, Xu Aidong. Terminal Security Level Grading Model of BP Neural Network Based on Edge Side[J].Communication Technology, 2018, 51(10): 2455-2458.

[11] Shi Ya, Wang Xiuhua, Yang Wei, Liu Li, Tan Zhezhen, Ouyang Wei. Study on the grading strategy of comprehensive evaluation model for long-term care of the elderly[J]. Chinese Journal of Nursing, 2018, 53(10): 1237-1243