ISBN 978-981-14-1455-8
Proceedings of 2019 the 9th International Workshop on Computer Science and Engineering
WCSE_2019_SPRING
Yangon, Myanmar, February 27-March 1, 2019, pp. 63-67
doi:10.18178/wcse.2019.03.011

Quantitative Analysis of Terrorist Attack Data Based on Weighted Clustering

Lei Juchao [1] and Wang Yaming [2+]

School of Computer Science and Engineering
Xi'an Technological University, Xi'an 710021, China

Abstract. Based on the terrorist attacks of 1998-2017 in GTD, a global terrorism database, this paper makes a deep analysis of the data related to terrorist attacks. By preprocessing and quantifying the data, it mainly uses the theoretical model of factor analysis. Classifying terrorist attacks according to many factors of harmfulness, then analyzing the data between multiple terrorist attacks, using the information of known perpetrators, according to the hazardous screening of suspects, using the characteristics of terrorist attacks, The correlation analysis model is used to determine the unknown perpetrators in terrorist attacks, and the similarity measurement model is used to determine the suspect degree of the perpetrators. The size of The details are as follows:

Through pre-analysis of the data, we can find out the main determinants of the harmfulness of terrorist attacks, including casualties, economic losses, timing, location, and so on. Through the analysis of the degree of influence of the factors, A model of clustering weight hazard classification based on factor analysis theory is established. Factor analysis weight method is used to solve the characteristic weight vector in data analysis and processing. Firstly, clustering is used to classify the terrorist attack events with similar harmfulness assessment, then the overall harmfulness of in-class events is quantitatively evaluated, and the harm of terrorist attacks is divided from low to high into one to five grades. The rating of future terrorist attacks is based on the distance between the event characteristics and the endoplasmic center. The new event is classified to the nearest centroid class, and the model automatically updates the centroid. According to the hierarchical model, the ten most harmful terrorist attacks in the past twenty years can be screened. In a sense, it can be very close to the comprehensive subjective division of many people.

The theory of cosine similarity is introduced on the basis of the previous experiment, and the correlation model based on multidimensional vector cosine similarity is established. First of all, by using the information of the known perpetrators, we exclude the data from which the perpetrators have been arrested, screen out terrorist organizations or individuals according to the magnitude of harmfulness, further analyze the characteristics of terrorist organizations that cause terrorist attacks, and generate feature vectors. At the same time, it analyzes the characteristics of the attack events that no organization or individual claims to be responsible for, and generates the feature vectors. Finally, the multi-dimensional vector cosine similarity algorithm is used to solve the suspect degree of the event suspect organization. Screen out the most likely perpetrators of each incident.

Keywords: Quantitative Analysis, Factor Analysis weight method, clustering, CoSine similarity

1. Introduction

In order to obtain the detailed information of terrorist attacks, the University of Maryland in the United States collected and constructed a global terrorism database (GTD), database, to collect and review the terrorist attacks of nearly 50 years. In the face of massive data, the effective application of data mining technology in the field of public security becomes more and more important [1]. Through the experience and lessons learned in the fight against terrorism, some countries have realized the urgency of adopting data mining technology to enhance the ability to acquire terrorist intelligence and combat terrorism.

[+] Corresponding author. Tel.: +86 17795836239
E-mail address: 616419803@qq.com.

Based on the above background, the collected counter-terrorism intelligence is mined in depth to effectively predict the occurrence of terrorist attacks. Therefore, more and more national governments and anti-terrorism agencies begin to attach great importance to how to use data mining technology in the war on terror to obtain accurate and effective information.

2. Classification Model of Clustering Weight Hazard Based on Factor Analysis Theory

2.1. Theoretical Analysis and Preparation

According to the classification of terrorist attacks according to harmfulness, the usual classification generally adopts the subjective method, mainly divided the casualties and economic losses as the main influencing factors. However, the harmfulness of terrorist attacks depends not only on the casualties and economic losses, but also on many factors, such as the timing, the region and the target, etc., so it is difficult to form a unified standard by using the above classification method. Therefore, it is necessary to establish a quantitative hierarchical model based on data analysis with mathematical modeling method. According to the degree of harm, the records of events in the global terrorism database are classified from high to low to one to five levels, and the ranking model is used to determine the ranking of harmfulness. The top 10 terrorist attacks, and give a list of the damage level of events.

According to the data in the global terrorism database and according to the degree of harm, the classification strategy of each terrorist event is drawn up objectively. Because some specific fields in the data table can determine the harmfulness of the event to a higher degree, it affects the classification of the harm degree of the whole event. In data processing, because there are a lot of missing values in the data table, the irrelevant data is detected, the missing data is processed by interpolation method, and the blank data field is removed to improve the data quality. First, a large number of nominal attributes are divided into two categories, one is the frequency conversion by value (for example, the targtype field), and the other is the effect of the attribute object. The force is transformed into numerical data (such as country field). After calculating the weight of (FAM) with factor analysis method, each event record is evaluated. The K-means clustering algorithm is used to cluster into five categories, and the results of inter-class evaluation are sorted according to the in-class comprehensive evaluation. A rating for assessing the dangers of terrorist activities. According to the results of single terrorist attack evaluation, the top ten terrorist attacks were selected, and the corresponding ID corresponding event grade labels in the cluster result query table were used.

2.2. Model Establishment and Conclusion Analysis

Considering the existence of a large number of fields with low impact factors in the GTD dataset, according to the hazard classification analysis, the fields needed to be used are assumed to be used in the model to be established, and these fields are preprocessed. The higher features of extended, success, suicide, propextent, country, region, city, attacktype, targtype, weapontype, kill and wound are extracted from the twelve influential factors (renouncing the ordinal number of eventid), and the remaining features are divided into two categories. One is the frequency conversion according to the value, which includes the following features: the transformation rule of attacktype, targtype, weapontype, is a certain class of the feature. The number of occurrences in all valid events, the percentage of all types of characteristics, and the conversion to numerical data by the influence of the attribute object. The following features are included: country, region, city, 's transformation rules are based on a ranking of the world's regional influence issued by authoritative bodies, dividing each region into three levels, 3 for high-profile regions, and 1 for low-profile regions. In addition, the kill is characterized by the difference between the total number of casualties and the number of terrorist casualties.

The following data processing is divided into two steps: the first step is data cleaning, which is used to remove noise from the data and correct inconsistencies, including deleting irrelevant data in the original data set, repeating data, smoothing noise data, Filter out data independent of the hierarchical model and handle missing values and outliers. In the process of data cleaning, the processing of missing value is very important. The processing of missing value includes two steps, that is, the recognition of missing data and the processing of missing value [2]. After judging whether there is a missing value, use interpolation to process the data. If there are individual fields missing in the entire record, fill the blank field with 0. The quantity ensures the original integrity of the data, which is helpful to improve the efficiency of the model.After programming to calculate the weights, the weights are shown in Table I:

Table I: Weight of each feature

ID	1	2	3	4	5	6	7	8	9	10	11	12
field name	extended	success	suicide	propextent	country	region	city	attacktype	targtype	weapontype	kill	wound
weight	0.1832	0.0105	0.1425	0.1336	0.0296	0.0344	0.1016	0.0547	0.0663	0.0753	0.0867	0.0811

The next step is to evaluate each record in the data table after getting the weight. Considering that there is no corresponding grade label for each record, the K-Means clustering algorithm is used to classify all cases. The basic idea of clustering algorithm is to divide a data object into a subset, each subset is a cluster, so that the objects in the cluster are similar to each other, but not similar to the objects in other clusters [3]. There are many clustering algorithms, this time we use K-Means clustering method, K-Means is one of the most commonly used clustering algorithm, the biggest characteristic of the algorithm is that the calculation is simple and fast [4].

The idea of using K-Means clustering algorithm is as follows:

(a) first select some classes / groups and initialize their respective centers randomly. The center point is the same position as the vector length of each data point. This requires anticipating the number of classes (that is, the number of centers).

(b) the distance from each data point to the center point is calculated.

(c) calculate each kind of center point as the new center point.

(d) repeat the above steps until each type of center changes little after each iteration. You can also randomly initialize the center point multiple times, and then select the one with the best results.

Achieve the centroid of five categories, as shown in Table II:

Table II: The centroid of mass gathered into five categories

category	centroid
0	[0.0609 0.8723 0.0512 1.8772 2.2650 1.3249 1.1996 1.2691 1.2735 0.3446 0.1479 0.4066]
1	[0.0000 1.0000 1.0000 1.3835 3.1905 1.0000 3.0210 3.0020 1.0000 2.8287 2.7392 4.4585]
2	[0.0000 1.0000 1.0000 2.2400 2.4000 6.0000 1.0000 1.0000 3.0000 5.2595 1.1893 5.6089]
3	[1.3953 10000 5.1162 1.8534 4.3955 2.6046 1.4186 1.3953 2.0465 4.1910 1.3573 4.6783]
4	[4.5280 9.8214 4.6683 3.6522 6.8439 2.3360 1.1498 1.1549 1.5255 4.4151 1.6235 5.0283]

A comprehensive assessment of in-class hazards is obtained, according to which criteria for each level of hazard are obtained.The results are as shown in Table III:

Table III: Hazard classification basis

Level ↓	A	B	C	D	E
category	1	2	4	3	0
Comprehensive assessment of in-class hazards	784.5819655	344.6685599	51.771248	9.34352609	0.781254298

The hazard level in the table ranges from high to low to level A to E, and each event is marked with a hazard level based on the closest distance to which centroid [5]. If a new terrorist attack event occurs in the future, the corresponding characteristics of the event are input into the model, and the distance from each cluster centroid is calculated and compared, which is classified as the closest category, and the damage level of the event is obtained according to the above table.

In the process of establishing a clustering weight hazard classification model based on factor analysis theory, the key to obtain hazard classification is to calculate the weight of each feature that leads to the occurrence of the event. According to the weight and each characteristic value of each event, the hazard assessment of the event is calculated. It is not difficult to see that the comprehensive assessment index of in-class hazards is actually the average value of the harmfulness assessment of all events in the class. Experiment one attempts to list the top 10 most dangerous terrorist attacks in the world, and the problem is translated into a top 10 record for assessing the damage of all events. The evaluation scores and ratings of each event have been marked. The selection method is to first select the top 10 according to the evaluation score in the level A event, and select the insufficient ones in the level B event, and so on. The top 10 most dangerous terrorist attacks in the world, is shown in Table IV:

Table IV: Ten most dangerous terrorist attacks in the world

No.	1	2	3	4	5	6	7	8	9	10
eventid	200109110004	200109110005	199808070002	201406150063	201603080001	200802010006	200409010002	201408090071	200607120001	200708150005

If it is required to rank any specified target event in the database table, it is only necessary to query the records according to the eventid characteristics in the table and obtain the corresponding hazard level.

3. Correlation Model Based on Multidimensional Vector CoSine similarity

3.1 Theoretical Preparation and Analysis

Screening through the database reveals that multiple terrorist attacks have yet to be identified [6]. Through the analysis and integration of existing complete records, a number of cases that may have been committed by the same terrorist organization or individual at different times and in different places have been connected together, and cosine similarity has been used [7]. According to the characteristics of unknown terrorist attacks and those of known criminal organizations, the cases with similar characteristics are classified into one category. This facilitates the identification of suspects in terrorist attacks of unknown perpetrators and the degree of suspicion of these suspects.

In the first step, according to the established quantitative classification model of terrorist attack harmfulness, all known criminal organizations or individuals should be graded quantitatively, and then the quantitative assessment of the harm of organizations or individuals should be carried out indirectly. Based on this idea, through the average harmfulness score of a certain organization or individual's K times of attack, the quantitative score of each known organization or individual's attack event is determined.

The second step is to analyze the characteristics of the first five most harmful organizations or individuals which are selected in the first step, and at the same time, to analyze the characteristics of the attacks of the undetermined organizations, and to associate the attacks with similar characteristics to carry out "serial case" analysis. The cosine similarity algorithm is used to match the first five attacks and the rank of a suspected organization or individual is obtained according to the similarity from high to low.

3.2 Model Establishment and Conclusion analysis

Through the organization or the individual has committed the crime the harmfulness rating, indirectly analyzes the organization or the personal harm magnitude. The solution of the latter two problems is based on cosine similarity algorithm, comparing the similarity between the organizational or individual crime characteristics and the characteristics of the unknown perpetrator event, and classifying all the most similar events into a class.

(a) screening of data for all events committed by each known organization or individual ID.

(b) matching with the event ID that has been rated by the hazard classification model in experiment 1, to determine the event harmfulness grade.

(c) to evaluate the distribution of the harmfulness grade of the organization or individual, and to determine the five organizations or individuals with the highest harmfulness.The results show that the top five organizations or individuals are as shown in the table V.

Table V: Ranking of organisations or individuals responsible for terrorist attacks

Terror level ranking	Organization or individual
1	Kata' ib Hezbollah
2	Popular Front for the Renaissance of the Central African Republic (FPRC）
3	United Front for Democracy Against Dictatorship
4	Jundallah (Pakistan)
5	Ansar al-Din Front

For the second experiment, the model based on cosine similarity algorithm is used to solve the problem. As the name implies, the similarity is evaluated by calculating the angle cosine value of two vectors. The cosine similarity is reflected in two-dimensional space by the coordinate value of the vector [8]. The solution steps are as follows:

(a) screening out the characteristics of all attacks carried out by each known organization or individual

(b) summarize the characteristics of all attacks, assess the comprehensive characteristics of each organization or individual, and generate feature vectors

(c) screening out all the unknown perpetrators or not claiming responsibility for the attacks, and generating a feature vector for each attack

(d) the similarity between the characteristics of unknown perpetrators and that of each organization or individual is calculated by using the formula of multidimensional vector cosine similarity.

The organization or individual that is most similar to the event feature is screened out by ranking the similarity degree.

According to the five most harmful organizations or individuals evaluated in the last step, the characteristics of the five organizations or individuals are analyzed, and the feature vectors are generated. The ID, of attack events is selected randomly in the database to extract the features of these attacks and to generate feature vectors. The feature similarity between event and suspect organization or individual is calculated by using multidimensional vector cosine similarity formula.For three random event samples, the five organizations are ranked as shown in the table VI:

Table VI: Ranking of suspected organizations in three samples of random terrorist attacks

Terrorist attacks ID	Kata' ib Hezbollah	Popular Front for the Renaissance of the Central African Republic (FPRC）	United Front for Democracy Against Dictatorship	Jundallah (Pakistan)	Ansar al-Din Front
201701090031	0.993517	0.892903	0.890335	0.965731	0.884775
201702210037	0.604778	0.551654	0.489784	0.553083	0.341395
201707010028	0.844275	0.966620	0.629141	0.755512	0.629600

In general, under reasonable assumptions, the model has a mathematical background, supported by big data technology and a large amount of data, so it is convenient to verify and improve the model. Model generalization ability, easy to promote, such as the application of natural disaster rating.

4. Acknowledgements

This work was supported by The Education Reform Project of Xi'an Technological University (No.18jgz01).

5. References

[1] Li Guohui. Temporal and Spatial Evolution and risk Analysis of Global terrorist attacks [J]. University of Science and Technology of China. 2014 (10)(in Chinese)

[2] Zhu Kai. Data preprocessing and feature Analysis of Real time data flow [J]. Wuhan University of Science and Technology. 2010 (05)(in Chinese)

[3] Wu Wenshuai. Evaluation and Application of data Mining method based on Multi-objective decision [J]. University of Electronic Science and Technology. 2015 (03)(in Chinese)

[4] Peter Bergen. Interface network. Five reasons to understand the root causes of the frequency of terrorist attacks [DB/OL] https://www.jiemian.com/article/769313.html.[2018-9-18](in Chinese)

[5] Zhang Yabing. A study on the causes and characteristics of extremism in Pakistan and the Government's depolarization [J]. A study of South Asia, 2015 (4): 86-98.(in Chinese)

[6] Wang Zhen. A brief discussion on the Global Anti-terrorism situation in the Post-Islamic State era [J]. West Asia Africa 2018 (01)(in Chinese)

[7] Sui Xiaoyan. Study on temporal and Spatial changes and influencing factors of terrorist attacks in China [D]. Dalian University of Technology. 2017(in Chinese)

[8] Su Weihua. Study on the Theory and method of Multi-index Comprehensive Evaluation [J]. Xiamen University. 2000 (01)(in Chinese)