Data Analysis Method of Terrorist Attacks Based on AHP-DBSCAN Method

Siyuan Wang1,a, Gang Wang2, Jiarui Zhang1
1Graduate College, Air Force Engineering University, 710051 Xi’an Shaanxi, China
2Air and Missile Defense College, Air Force Engineering University, 710051 Xi’an Shaanxi, China
a Corresponding author: 982268878@qq.com

Abstract. Firstly, the risk assessment system of terrorist attacks is established, and the relevant indexes of terrorist events are screened, combined and quantified. This paper gets the index weights by AHP method, and the degree of damage is calculated. Then it uses DBSCAN method to cluster the sample data and optimize the parameters Eps and MinPts. The results show that potential terrorist organizations or individuals are predicted, and this method can be applied to other similar data analysis problems.

1.Introduction
Terrorist attacks refer to attacks manufactured by extremists or organizations against but not limited to civilians and civilian facilities, which are not in line with international morality[1]. An in-depth analysis of the data related to terrorist attacks will help deepen people's understanding of terrorism and provide valuable information support for anti-terrorism[2]. After synthesizing all kinds of factors, the harmfulness degree of terrorist events is evaluated in a unified way. This paper selects, combines and quantifies the relevant indexes, establishes the evaluation system of the harmfulness degree of terrorist attacks, and uses the membership function to solve the harmfulness degree.

The study finds that several cases of the same terrorist organization or individual committing multiple crimes at different times and locations are related to a certain extent. If they are linked together to organize a unified investigation, it will help to improve the efficiency of solving cases, and is conducive to the early detection of new or hidden terrorists. The types, means and harmfulness of related cases are similar to a certain extent[3,4]. Clustering method can be used to quickly classify a large number of incidents and obtain information about potential terrorists. DBSCAN (density-based spatial clustering of applications with noise) is a classical density-based clustering algorithm proposed by Ester Martin et al. It measures the density of a hyperspherical region by the number of data objects contained in it. DBSCAN can detect clusters of arbitrary shape and effectively identify noise, and has strong clustering ability. Based on the hazard assessment system, this paper uses DBSCAN method to cluster the data, and optimizes the key global parameters Eps and MinPts to predict potential terrorist organizations or individuals.

2.Evaluation system and membership function of the degree of damage to terrorist attacks
The data used in this paper are the records of terrorist attacks in the world from 2016 to 2017 collected by an organization in the Global Terrorist Database (GTD). After data cleaning of duplicate and invalid incidents, the remaining 22678 valid incidents are valid. According to the contents of the database, there are more than 50 related indicators of terrorist incidents. First of all, this study should
optimize the selection and integration of indicators. Terrorist incidents have great lethality and destructive power, can directly cause huge casualties and property losses, but also bring tremendous psychological pressure to people, causing a certain degree of social unrest, hindering the normal work and life order, thus greatly hindering economic development\cite{5,6}. Therefore, this paper selects eight key indicators for analysis: area, inclusion criteria, whether it is part of an event, attack type, target/victim information, casualties, property loss, whether it is an international event. The evaluation system of the degree of damage to terrorist attacks is shown in Figure 1.

![Evaluation system of the degree of damage to terrorist attacks](image)

Figure 1. evaluation system of the degree of damage to terrorist attacks.

(1) area

The same terrorist organization or individual often chooses the same area where terrorist incidents occur. Regions are determined by both the country and the city in which the incidents occurred. For countries, according to the regional importance of the countries and regions where the terrorist attacks occurred, they are divided into important countries or regions, non-important countries or regions. Among them, important countries or regions include Canada (38), the United States (217), mainland China (44), Hong Kong (89), Japan (101), Macao (117), South Korea (184), Taipei (201), China (97), Austria (15), Belgium (21), Denmark (55), Finland (68), France (6), 9), Germany (75), Iceland (91), Ireland (96), Italy (98), the Netherlands (142), Norway (151), Portugal (162), Spain (185), Sweden (198), Switzerland (199), Britain (603), Russia (167), Australia (15), New Zealand (144), and others are non-important countries or regions (the numbers in brackets are the numbers of countries or regions in GTD). The quantitative results of national indicators are shown in Table 1.

| Countries                      | Quantized value |
|-------------------------------|-----------------|
| Important countries or regions| 0.7             |
| Non-important countries or regions| 0.3            |

As a political, economic, cultural center and residential gathering place, the danger of terrorist attacks in cities is much higher than that outside cities. The quantitative results of urban indicators are shown in Table 2.

| Whether it occurs in Cities | Quantized value |
|----------------------------|-----------------|
| Occurring in cities        | 0.7             |
| Occurring outside the cities| 0.3             |

Analytic Hierarchy Process (AHP) refers to the decision-making method that decomposes the elements that are always related to decision-making into objectives, criteria, schemes and so on, and makes qualitative and quantitative analysis on this basis. The index weight of the country where the event happened and whether the event happened in the city is (0.75, 0.25) obtained by the AHP method, and the overall quantitative value of the location index can be calculated.
(2) Inclusion criteria
There are three criteria for inclusion of events in the GTD database: 1) political, economic, religious or social objectives; 2) intent to coerce, intimidate or incite more people; 3) beyond the scope of international humanitarian law. AHP method is used to distribute the weights among the three, and the weighted index is (0.2, 0.2, 0.6). The overall quantification method of the selected criteria is obtained by summing the three weighted indices as shown in Table 3.

| Inclusion criteria                      | Quantized value |
|----------------------------------------|-----------------|
| Meeting all three criteria              | 1               |
| Meeting No.1,3 or No.2,3 criteria       | 0.8             |
| Meeting No.1,2 criteria                 | 0.4             |

(3) Whether it is part of an event
The same organization or individual often carries out a series of terrorist activities. Therefore, if both terrorist incidents are part of a series of events, they may be acts of the same organization or individual. In this paper, if an event is part of a series of events, the quantized value is 0.7; otherwise, the quantized value is 0.3.

(4) Attack type
Different types of attacks will directly affect the degree of harm of the incident, such as assassination and other ways are generally political and can cause bad effects. Armed assault and explosion always result in casualties. These methods are more harmful. Unarmed attacks and other means involve small casualties, resulting in a relatively low degree of harm. Therefore, the type of attack is also an important index to evaluate the degree of damage. The quantization method of attack type is shown in Table 4.

| Attack type                      | Quantized value |
|---------------------------------|-----------------|
| Assassination                   | 0.9             |
| Armed attack                    | 0.8             |
| Bombing / explosion             | 0.7             |
| Hijack                          | 0.6             |
| Hostage taking (roadblocks)     | 0.5             |
| Hostage taking (kidnapping)     | 0.4             |
| Facilities/infrastructure attacks| 0.3             |
| Unarmed                         | 0.2             |
| Unknown                         | 0.1             |

(5) Target/victim information
When the murderer targets different targets/victims, the damage degree will be significantly different. In addition, the target/victim information also indicates to a certain extent that the main target of terrorists, the same terrorists often choose the same target. Referring to the data, this method classifies the target/victim information into four categories, and the corresponding quantitative methods and values are shown in Table 5.

| Target/victim information                              | Quantized value |
|--------------------------------------------------------|-----------------|
| Government (general meaning, diplomacy)                | 0.8             |
| Police, military                                       | 0.6             |
| Business, abortion-related                              | 0.4             |
| airports and aircraft, educational institutions         |                 |
| food or water supplies, maritime, non-governmental     |                 |
| organizations, other, religious, telecommunications,   |                 |
| transport, utilities                                   |                 |
(6) Casualties

When the number of casualties in two terrorist incidents is close, the manner of attack, the location chosen and the population targeted by the perpetrators are likely to be the same, that is, the same organizations or individuals. And the number of casualties is also an important index to judge the degree of harm of the event. The more casualties, the greater the degree of harm of the event. The membership function of casualty can be expressed as a partial decreasing semi normal distribution function.

\[ u_{c}(p) = 1 - e^{-k_{p}p} \]  

In the formula, \( p \) is the total number of casualties, and \( k_{p} \) is the related parameter. Here \( k_{p} \) takes 0.00002.

(7) Property loss

In the GTD database, property losses fall into four categories: 1) catastrophic (possibly more than $1 billion), 2) significant (possibly more than $1 million, but less than $1 billion), 3) small (possibly less than $1 million), 4) unknown. The quantitative value of property loss index is shown in Table 6.

| Property loss | Quantized value |
|---------------|-----------------|
| catastrophic (possibly more than $1 billion) | 0.9 |
| significant (possibly more than $1 million, but less than $1 billion) | 0.6 |
| small (possibly less than $1 million) | 0.3 |
| unknown | 0.1 |

Table 6. Quantitative results of Property loss.

(8) Whether it is an international event

International events often have greater influence and harm than non-international events. If an organization or an individual always produces non-international terrorist incidents, then it has a very small probability of creating international incidents. Therefore, the quantitative value of international events is set to 0.7, and the quantified value of non-international events is set to 0.3.

3. Normalization method and weight determination of hazard level evaluation indicators

Assuming the number of terrorist attacks is \( m \), and the event set can be expressed as \( M = \{M_1, M_2, ..., M_n\} \). The number of selected indicators is \( n \), and the set of indicators can be expressed as \( U = \{U_1, U_2, ..., U_n\} \). According to the previous analysis, there are eight hazard assessment indicators selected by the model, which are: area, inclusion criteria, whether it is part of an event, attack type, target/victim information, casualties, property loss, whether it is an international event. The quantitative value of the hazard assessment index \( U_j \) of event \( M_i \) is \( x_{ij} \). The initial decision matrix of events can be expressed as \( X = (x_{ij})_{m \times n} \).

In the initial decision matrix, the dimension and order of magnitude of each index of the event are different. It is impossible to make direct analysis of the next decision. Therefore, this study needs to normalize the initial decision matrix \( X \) and construct the standard decision matrix \( Y = (y_{ij})_{m \times n} \).

For the cost index:

\[ y_{ij} = \frac{\max_j(x_{ij}) - x_{ij}}{\max_j(x_{ij}) - \min_j(x_{ij})} \]  

For the benefit index:

\[ y_{ij} = \frac{x_{ij} - \min_j(x_{ij})}{\max_j(x_{ij}) - \min_j(x_{ij})} \]  

Among them, The cost indexes are the index with smaller value and greater danger and the benefit
index is the index with greater value and greater danger. The eight indexes selected in this paper are changed into benefit indexes after quantifying the above methods. After normalization, the standard decision matrix \( Y \) is obtained, which is the basis of the next step.

The weight value of each index in hazard calculation is assigned by AHP method. Finally, the weights of each index are \((0.0441, 0.0664, 0.0300, 0.1150, 0.1572, 0.2521, 0.2276, 0.1076)\), and the comprehensive hazard metrics of all events can be obtained in turn.

4. Predicting potential terrorists based on DBSCAN method

DBSCAN (density-based spatial clustering of applications with noise) is a density-based spatial data clustering method\(^7,8\). The algorithm takes the region with enough density as the distance center and grows the region continuously. By using the concept of density-based clustering, clusters of arbitrary shape can be found in the noisy spatial database, and adjacent regions with enough density can be joined together to deal with abnormal data effectively. It is mainly used for clustering spatial data. Compared with other clustering methods, DBSCAN does not need to input the number of clusters to be classified\(^9,10\), which is in line with the condition that the number of potential terrorists cannot be known before the prediction. Therefore, DBSCAN method is used to cluster analysis.

4.1 Basic definition

Neighborhood radius (Eps): The search radius of the algorithm at run time. The area within a given object radius Eps is called the Eps field of the object.

Density (MinPts): Point density in the search radius Eps.

Core point: Points in Eps neighborhood whose points are greater than or equal to MinPts.

Boundary point: It is not the core point but belongs to the Eps neighborhood of other core points.

Noise point: The point object that is neither the core nor the boundary point.

Direct density can be as follows: \((p, q) \in \text{Data set } D\), p is the core point, q is in the field of Eps of p, then q is called the point of direct density of p.

As shown in Figure 2, A is the core point, B and C are the boundary points, and N is the noise point.

![Figure 2. Basic principle of DBSCAN method.](image)

4.2 Basic process

(1) DBSCAN searches clusters by checking the Eps neighborhood of each point in the data set. If the Eps neighborhood of point p contains more points than MinPts, a cluster with P as the core object is created.

(2) DBSCAN then iteratively aggregates objects directly densely reachable from these core objects, which may involve the merging of some densely reachable clusters.

(3) The process ends when no new points are added to any cluster.
4.3 Parameter optimization

DBSCAN includes three inputs: 1) data set D, 2) the minimum number of neighborhoods points that a given point becomes a core object in the neighborhood: MinPts; 3) neighborhood radius: Eps. Eps and MinPts need to be set according to the specific circumstances.

The value of Eps can be obtained by drawing Eps-K curve, and the position of obvious inflection point of Eps-K curve is a better parameter. If the parameters are set too small, most of the data can not be clustered; if the parameters are set too large, multiple clusters and most objects will be merged into the same cluster.

The selection of MinPts has a guiding principle, MinPts≥dim+1, where dim represents the dimension of the data to be clustered. It is unreasonable to set MinPts to 1 because each independent point is a cluster. When MinPts<2, the result is similar to the nearest neighbor method for hierarchical distance. MinPts must select a value greater than or equal to 3. If the value is too small, the result in the sparse cluster is considered that the boundary point is not used for further extension of the class because the density is less than MinPts; if the value is too large, the two neighboring clusters with higher density may be merged into the same cluster. Therefore, whether the value is properly set will have a great impact on the clustering results.

4.4 Experimental results

First, the Eps-K curve is obtained by DBSCAN clustering analysis of the normalized data matrix. As shown in Figure 3, when Eps is lower than 0.1, the curve has obvious inflection point. Further observation shows that the value of Eps should be about 0.08.

![Figure 3. Eps-K curve.](image)

When the Eps value is too large, the data can only be divided into very few categories, that is, all events are caused by very few suspects, which does not conform to the characteristics of terrorist attacks. When the Eps value is too small, the number of categories will rise sharply, and eventually form a separate category of each event, the number of events is the number of categories. Obviously, it is not in line with the actual situation. After many experiments, it was found that the data were divided into 892 groups when the field radius Eps=0.09 and the minimum field number MinPts=9. It is found that there are as many as 1000 terrorist organizations on record in the world at present. It is more reasonable to calculate the number of potential terrorists as 892 individual terrorists. Some clustering results are shown in Figure 4.
Figure 4. Clustering results.

According to the evaluation system and membership function, the average hazard degree of each group of events can be calculated separately. Some groups with the greatest hazard degree should be listed as key monitoring and prevention objects. In addition, each group of data centers can be obtained from the above information, and each group is represented as \((y_1, y_2, \ldots, y_n)\). Assuming a new terrorist attack occurs, the index is quantified and normalized according to the same method to obtain \((x_1, x_2, \ldots, x_n)\). The Euclidean distance formula is as follows.

\[
d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}
\]

(4)

The Euclidean distance between the event and all clustering results is obtained from the upper formula. The lowest distance is the group with the highest degree of suspicion, that is, the event is most likely to be committed by the same potential terrorists. Joint investigation will greatly improve the detection rate of terrorist attacks and make a great contribution to the world's anti-terrorism.

5. Conclusion

As a typical method of spatial density clustering, DBSCAN has attracted more and more attention because it can supplement the adaptability of spatial clustering to data sets, the sensitivity of noise points and the deficiency of arbitrary shape clustering recognition, and has been widely used in practical case analysis. On the basis of extracting the data of terrorist attacks in GTD database from 2016 to 2017, this paper firstly makes the data cleaning, establishes the evaluation system of the degree of the event hazard, obtains the weight of each index by AHP method and calculates the degree of the hazard. Then, this paper uses DBSCAN method to cluster the sample data, and optimizes the parameters Eps and MinPts. Experimental results show that this method has good applicability and flexibility in the analysis of terrorist attack data. The results can also be used for further in-depth data mining, and provide great help for the world anti-terrorism cause.

Because the terrorist attack data in the database contains various factors and complex situations, there are still many areas unnoticed for improvement in the analysis of terrorist attack data, for example, the method does not consider the subjective preferences of terrorists in the choice of weapon types. How to select and integrate the indicators more comprehensively and rationally, and how to further optimize the overall parameters will be the next research direction of this topic.

References

[1] G.H. Li. University of Science and Technology of China. D. (2014)
[2] S. Nunn. Journal of Homeland Security and Emergency Management. J. 2(3) (2011)
[3] B. Durodié, S. Wessely. The Lancet. J. 360(9349) (2002)
[4] R.A. Blais, M.D. Henry, S.R. Lilley, J.A. Pan, M. Grimes, Y.Y. Haimes. Systems and Information Engineering Design Symposium. P. SIEDS '09 (2009)
[5] W. Gong, Z. Liu, Y. Wang, J. Xu. Computer Simulation. J. 32(04): 30-33+398 (2015)
[6] D.Y. Sun, W.J. Li, J.L. Fu, B. Li. Journal of Safety and Environment. J. 13(05): 259-264 (2013)
[7] J.Y. Fu, C.F. Jing, M.Y. Du. Science of Surveying and Mapping. J. 1-11[10-24]. (2018)
[8] S.R. Feng, W.J. Xiao. Computer Engineering and Applications. J. (20):216-221 (2007)
[9] H.F. Zhou, P. Wang. Journal of Xi'an University of Technology. J. 28(03):289-292 (2012)
[10] Z.Q. Fu, X.F. Wang. Network Security Technology & Application. J. (08):34-36 (2018)