Quantitative Analysis of Recorded Data of Terrorist Attacks

Zeng Zhen [1, a], Du Jian[1], Luo Xiao[1]

[1] College of Mechanical and Electrical Engineering, Southwest Petroleum University, 610500, China

[a]sccdlqzz@126.com

Abstract

In recent years, more and more intelligence agencies and anti-terrorism departments have begun to study how to predict the risk of terrorism as early as possible and strangle terrorist attacks at the embryonic stage. Therefore, a scientific quantitative analysis of the data related to terrorist attacks has been carried out to help improve the pertinence and efficiency of the fight against terrorism. Based on the quantitative analysis of the recorded data of terrorist attacks in the world from 1998 to 2017 in the Global Terrorist Database, this paper draws the ten most harmful terrorist attacks in the world in the last two decades.

Keywords

Terrorist attack; Random forest; Machine learning.

1. Introduction

Since the 1990s, especially in the new century, terrorist attacks around the world have brought more and more serious consequences. The accompanying loss of personal and property has led to people losing their sense of security, resulting in a certain degree of social unrest, hindering the normal work and life order, and thus greatly hindering economic development. Terrorism is a common threat to mankind, and the fight against terrorism is the responsibility of every country. 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 counter-terrorism and anti-terrorism.

At present, the classification of the hazards of catastrophic events is usually based on subjective methods. Authoritative organizations or departments have the power to select several important indicators and to enforce the classification criteria. Because of the particularity of terrorist attacks, casualties and economic losses can not completely measure the degree of harm of terrorist attacks. Terrorist attacks can also bring serious social panic compared with other disasters. Relating to many factors, such as the target of attack, the way of attack and the type of weapon used, this paper quantifies the classification of terrorist attacks by establishing a quantitative model based on data analysis, which makes the classification more objective, and lists 10 terrorist attacks with the greatest degree of harm in the past 20 years according to the classification.

2. Model Hypothesis

In the process of mathematical modeling, in order to make the model simple and clear, the following assumptions are established without affecting the meaning and accuracy of the model.

1. Regional impact can be measured by GDP.
2. Events with strong correlation can be regarded as the same event.
3. If the amount of missing data is too large, the feature is not reference.
4. Subclasses can be neglected when the main class can better represent the impact on events.
5. The characteristics of events with little influence can be neglected.

6. Data can be improved by various means.

7. In view of the contradiction between the economic loss and the degree of economic damage in Annex I, the degree of economic damage shall prevail.

3. Data Preprocessing

Through the analysis of the data in the database, we find that the data in the database can be divided into three types: text data, categorized data and numerical data. Some of the data are missing or contradictory. For the text data, we select their corresponding codes and assign the categorized data according to the degree of damage to terrorist attacks caused by the characteristics. For missing data, first count the number of missing data, if the ratio of missing data to total data is less than 10%, then use the method of expectation maximization to complete the data; if the amount of missing data is too large, this feature has no reference. In the remaining features, we have consulted the State Council's "Classification Standards for Particularly Major and Major Public Emergencies", "Australian Air Attack Classification System" and "American Homeland Security Bureau Warning Classification" and so on. We have taken into account the actual damage caused by terrorist attacks, the psychological panic caused to people and the negative impact on society, and have eliminated the obvious impact on society. The characteristics that have little impact on the classification of terrorist attacks, such as additional instructions, have been merged with features that can be merged. In addition, classifying countries and regions according to their numbers is not conducive to reflecting the impact of regional differences on the harmfulness of terrorist attacks. We consider adding regional GDP to solve this problem. We consulted the United Nations Statistical Office's GDP statistics for each region in the past year through the Internet. For those countries or regions that do not exist now, they are replaced by the GDP of new countries or regions that they joined or generated after disappearance.

4. Feature Selection

After data preprocessing, there are inevitably many features that have little relation to the time-harm degree of terrorist attacks. There are information redundancy among the features, even irrelevant to the classification judgment. These features exist in the data space, which will bring negative effects to the accuracy of the model. It is concluded that dimensionality reduction for high-dimensional datasets with large-scale features can not only improve the fitting speed of the algorithm, save time and space, but also avoid the negative impact of irrelevant features on the algorithm fitting. Dimension reduction can be divided into feature extraction and feature extraction.

Among them, feature selection of data sets is essentially a subset of features composed of H-dimensional features, which will be more effective for category judgment according to some filtering conditions, and discards those features which are useless or have little effect on category judgment, and then maps samples to a new feature subspace for representation. That is, a new subspace of R dimension ($r < H$) is selected from the original H dimension space, and then weighted in the new subspace. In order to achieve this goal, two problems need to be solved. The first one is the generation strategy of feature selection, and the second one is how to judge how much the selected new feature subset plays in the determination of categories.

The essence of feature selection is to select a subset of features that best represent the sample representation. Starting from the process of generating and evaluating the new feature subset from the original feature to the feature subset, the feature selection steps can be summarized as Figure 1. Feature selection is essentially a combinatorial optimization problem. Finding the minimum feature subset that meets the requirement is a NP problem.

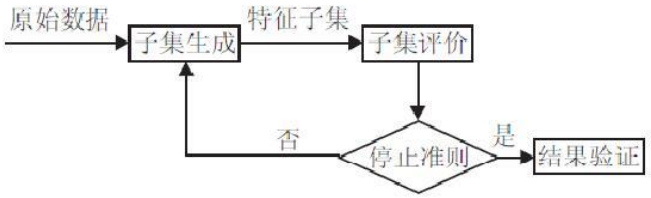

Fig. 1 Feature Selection Process

5. Model Establishment

5.1 Random Forest

Random forest [3] is to build a forest in a random way. Its basic unit is the decision tree. There is no relationship between each decision tree of random forest. Random forest uses bootstrap resampling technique to generate a new training sample set from K samples randomly sampled from the original training sample set N, and then generates K classification trees to form a random forest according to the self-help sample set. The classification results of the new data are determined by the scores formed by the voting of the classification tree. Its essence is to explain the improvement of decision tree algorithm from an intuitive point of view. Each decision tree is a classifier (assuming that it is aimed at classification problem now), then for an input sample, N trees will have N classification results. Random forest integrates all the results of classified voting and assigns the category with the largest number of votes to the final output. This is the simplest Bagging idea. Random forests have the ability to analyze the classification characteristics of complex interactions. They are robust to noise data, high-dimensional data and missing data, and perform well on data sets. The introduction of two randomicities makes it difficult for random forests to fall into over-fitting.

5.2 Sequential Backward Search

The sequential backward search method starts with the original feature set S_n and eliminates one feature at a time. The constraint to be satisfied is the classification error $e_{n-1} > e_n$ generated by the residual feature set S_{n-1}. Since S_n contains n features, all n methods remove one of them from the feature set S_n. For each subset of S_{n-1} features, the corresponding e_{n-1} is calculated. If $e_{n-1} < e_n$, the feature x that maximizes $e_{n-1} - e_n$ is chosen as the optimal feature, and if $e_{n-1} > e_n$ is satisfied, the calculation is stopped.

5.3 Wrapper Feature Selection

Unlike filtered machine learning, which does not consider subsequent learners, wrapped feature selection directly takes the performance of the learner to be used as the evaluation criterion of feature subset. Because the selection method of wrapped features is directly optimized for a given learner, wrapped features generally avoid filtering. LVW is a typical method. Random strategies are used to search feature subsets, and each evaluation of feature subsets requires training learners, which is costly.

5.4 Hazard Degree

The hazard degree of each event is defined as:

$$Z_i = \sum_{i=1}^{114184} \sum_{j=1}^{n} a_j x_{ij} \quad (1)$$

Among them:

Z_i : The harm degree of the first terrorist attack;

a_j : the weight of the j th feature;

x_{ij} : The value of the jth feature of the j th terrorist attack

6. Model Solution

To solve this problem, a Wrapper feature selection method RFFS based on Stochastic Forest is proposed. Random forest algorithm is used to rank the features by measuring the importance of variables. Then a sequence backward search method is used to remove the least important feature (the least importance score) from the feature set at a time, iterate one by one, and calculate the classification accuracy. Finally, the minimum number of variables and points are obtained. The feature set with the highest class accuracy rate is used as the result of feature selection. The flow chart is shown in Figure 2.

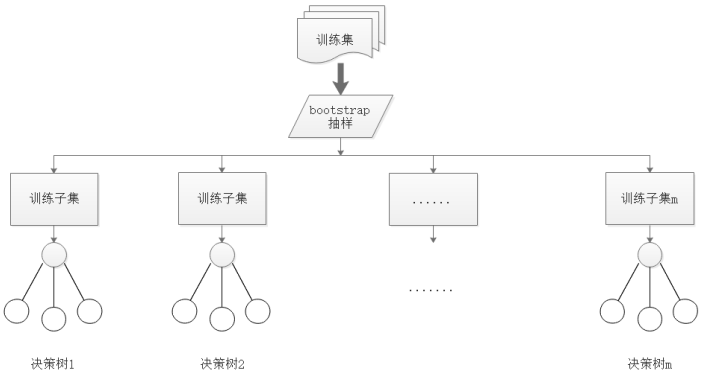

Fig. 2 (a) Stochastic Forest Flow Chart

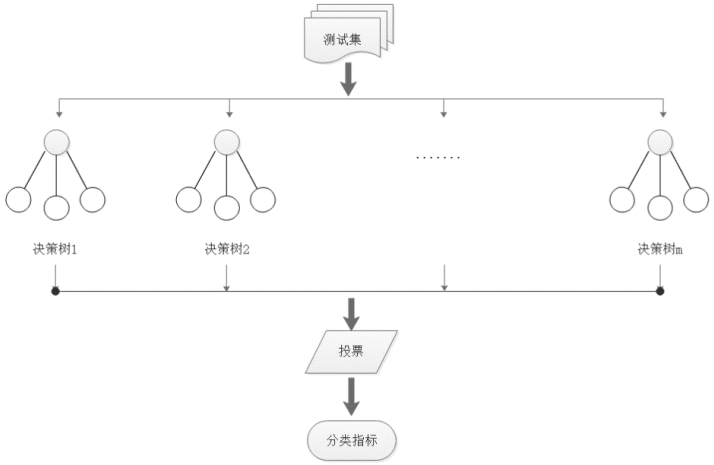

Fig. 2 (b) Stochastic Forest Flow Chart

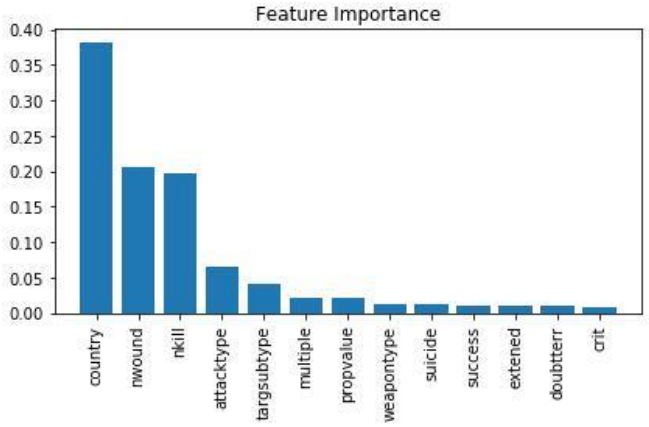

Fig. 3 Feature Extraction Results

After importing the pretreated data, we use Python to implement feature selection based on random forest, and finally get 13 features and their weights as shown in Figure 3 and Table 1.

Table 1 Weights

Features	Weight
Place	0.381470
Selected database standards	0.007650
Suspected terrorist incidents	0.009967
Related events	0.022091
Success	0.011211
Suicide attack	0.012171
Type of attack	0.065892
Types of target victims	0.041678
Weapon type	0.013478
Total number of deaths	0.196067
Total injuries	0.205949
Degree of damage to property	0.021180
Persistent event	0.011191

All terrorist attacks in the past 20 years are quantified by weights (1). The events are classified into five grades. The results are as follows: Table 2

Table2 Grading Results

Hazard level	Hazard value standard
1	>200
2	151-200
3	101-150
4	51-100
5	1-50

According to the level of hazards, the top ten terrorist incidents with the highest degree of hazards in 20 years are listed, such as Table 3.

Table3 Top Ten Terrorist Attacks

Event Number	Event Description	Hazard Value
200109110004	911 Incident in the United States, 3001 people were killed and 16 493 injured	8532
199808070002	US Embassy in Kenya attacked, killing 224 people and wounding 4,000	4086
201603080001	Mustard bomb attacks in Iraq have killed more than 1,500 people and injured more than 1,500 others.	1511
200802010006	Chadian rebels attacked the capital, killing hundreds and injuring more than 1,000 people	1055
201710010018	Terrorists attacked American hotels, killing 58 people and injuring 851.	933
200607120001	A train bombing in Mumbai killed 188 people and injured 817.	922
200409010002	Chechen rebels in Russia hijacked schools, killing 344 people and injuring thousands	860

200708150005	More than 500 people were killed and 1500 injured in a car bomb attack in Iraq	850
201606010046	Abduction of 1500 Iraqi civilians, 49 deaths and more than 900 missing	660 624
200509140001	Bomb attacks in Baghdad killed 160 people and injured 570.	

References

[1] Kan Zhigang, Jinxu. Research and Implementation of Data Preprocessing in Data Mining [J]. Computer Applied Research, (07): 21-28, 2004.

[2] Yao Xu, Wang Xiaodan, Zhang Yuxi, Quan Wen. Overview of feature selection methods [J]. Control and decision-making, 2012.

[3] Liu Kai. Study on adaptive feature selection and parameter optimization algorithm for stochastic forests [D]. Changchun University of Technology, 2018

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

[5] Li Xinhai. Application of Stochastic Forest Model in Classification and Regression Analysis [J]. Journal of Applied Entomology, 2013, 50 (04): 1190-1197

[6] Yao Dengju, Yang Jing, Zhan Xiaojuan. Feature selection algorithm based on random forest [J]. Journal of Jilin University (Engineering Edition), 2014, 44 (01): 137-141.

[7] Hu Jie.Summary of research on feature dimension reduction of high-dimensional data[J].Computer Applied Research, 2008 (09): 2601-2606.