Suspect Prediction Based on Naive Bayesian Method

Zixuan Hu*

College of Mathematics and Systems Science, Shandong University of Science and Technology, Qingdao, Shandong, 266590, China

*Corresponding author’s e-mail: 1026343526@qq.com

Abstract. Aiming at the terrorist attacks that have happened in 2015 and 2016, which have not been organized or claimed responsibility by individuals, the suspect prediction algorithm based on naive Bayesian method is used to solve the problem of finding the perpetrators of terrorist attacks. Firstly, we select five terrorist organizations and individuals which are more harmful to terrorist attacks. Then, we use the suspect prediction algorithm based on naive Bayesian method to select the terrorist attacks that occurred in 2015 and 2016, which have not yet been organized or claimed responsibility by individuals. Finally, we use the Sklearn machine learning library of Python to calculate and get all the suspects in each incident. The probability of a suspect is the degree of suspicion.

1. Introduction

Terrorism is a common threat to mankind, and the fight against terrorism is the responsibility of every country. The 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, there are few studies on the law of terrorist organization's behavior at home and abroad [1]. Most of the literature and reference materials are mainly from the perspective of international politics, religion, economy, etc. to study the origin, growth reasons of terrorism, the impact on human beings and so on [2,3]. Few literatures use data-driven method to calculate the behavior of terrorist organizations quantitatively, so this paper uses data analysis method to mine and predict the behavior law of terrorist organizations in the database of terrorist organizations, which has a great complementary role for anti-terrorism research, and the research results of this paper have a certain auxiliary role in combating domestic and foreign terrorist forces [4].

This paper mainly considers the application of Naive Bayesian classification to select terrorist attacks that occurred in 2015 and 2016, which have not yet been organized or claimed responsibility by individuals, to form training sets, and use Python's sklearn machine learning library to calculate. Then data processing is carried out, and terrorists are classified. For the case of multiple perpetrators of the same incident, the method of adding events is adopted, that is, each perpetrator commits a crime independently. In addition, the GaussianNB model in sklearn's only allows float type data, so it is necessary to convert the string into floating-point data and store corresponding tables. The simple Bayesian instance is created by calling the GaussianNB model, and the training model is input into the training set. Then the events to be analyzed are input to obtain the probability of all suspects in each event, namely the suspect degree.
2. Suspect Prediction Based on Naive Bayesian Method

2.1. Bayesian statistical method
The classical statistical method is only the use of two kinds of information, namely model information and sample information, and the biggest advantage of the Bayesian method compared with the classical statistical method is the use of Bayesian formula[5]. The formula is as follows:

\[ P(A_i | B) = \frac{P(B | A_i)P(A_i)}{\sum_{i=1}^{n} P(B | A_i)P(A_i)} \]  

(2.1)

Where it is assumed that event B always occurs at the same time as one of the events \( A_1, A_2, \ldots, A_n \). In the above formula (2.1), \( A_i \) can be regarded as all possible causes for random event B to occur. \( P(A_i) \) is expressed as the prior probability of occurrence of event \( A_i \). If event B has occurred, then the probability of event \( A_i \) can be recalculated. Conditional probability \( P(A_i | B) \) is the re-estimation of the probability of event \( A_i \) after the known occurrence of B, which is called the posterior probability of random event \( A_i \).

In simple terms, the Bayesian statistical method achieves the full use of prior probabilities to estimate posterior probabilities. A detailed description of the Bayesian statistical method is as follows.

Unknown parameters can be regarded as random variables (vectors), denoted as \( \xi \). When \( \xi \) is known, the distribution density function \( p(x_1, x_2, \ldots, x_n; \xi) \) of sample \( x_1, x_2, \ldots, x_n \) can be regarded as the distribution density of \( x_1, x_2, \ldots, x_n \) under the condition of \( \xi \), denoted as \( p(x_1, x_2, \ldots, x_n | \xi) \) or abbreviated as \( p(x | \xi) \).

1) The prior distribution of random variable \( \xi \) is represented by \( \pi(\xi) \). Generally, the prior distribution of parameter \( \xi \) is determined by previous knowledge. If the prior distribution is missing, then \( \pi(\xi) \) should be replaced by uniform distribution, which is in the range of \( \xi \);

2) Through prior distribution \( \pi(\xi) \) and conditional distribution density \( p(x_1, x_2, \ldots, x_n | \xi) \), we can calculate the joint distribution density of \( x_1, x_2, \ldots, x_n \) and \( \xi \) and the distribution of samples \( x_1, x_2, \ldots, x_n \). Then we can use them to calculate the distribution density of \( x_1, x_2, \ldots, x_n \) under the condition of \( x_1, x_2, \ldots, x_n \), that is to say, we can use Bayesian formula to solve the posterior probability density \( p(x_1, x_2, \ldots, x_n | \xi) \);

3) The posterior probability distribution density \( p(x_1, x_2, \ldots, x_n | \xi) \) is used to deduce \( \xi \), and the formula is obtained:

\[ p(\xi | x_1, x_2, \ldots, x_n) = \frac{\pi(\xi)p(x_1, x_2, \ldots, x_n | \xi)}{p(x_1, x_2, \ldots, x_n)} \]  

(2.2)

2.2 Naive Bayes classifier
Bayesian classifier provides a classification method, which is based on Bayesian statistical method and Bayesian theorem. The mark of Bayesian classifier is that it uses not only model information and sample data information, but also prior probability distribution information.

The biggest difference between naive Bayesian method [6,7] and Bayesian network is the correlation between attributes. Bayesian network allows the association between attributes. Naive Bayesian thinks that attributes are independent of each other. Although this attribute independence
hypothesis is not completely consistent with the real situation in many cases, it can greatly reduce the complexity of the model and algorithm. The basic idea of Naive Bayesian Prediction is to compare the posterior probability of each category and consider the category corresponding to the maximum probability (i.e. the most likely classification situation) as the prediction result.

The formal description of Bayesian classification prediction is as follows:

Let \( x = \{a_1, a_2, \ldots, a_n\} \) be a feature to be classified, and each \( a \) is a feature attribute of \( x \);

1. Take out the category set \( C = \{y_1, y_2, \ldots, y_n\} \);
2. Calculation \( p(y_1 \mid x), p(y_2 \mid x), \ldots, p(y_n \mid x) \);
3. Find \( p(y_k \mid x) = \max \{ p(y_1 \mid x), p(y_2 \mid x), \ldots, p(y_n \mid x) \} \), then \( y_k \) is the prediction result.

In the formal description, how to calculate the conditional probability in step (2) is the key to the whole algorithm. According to the Bayesian formula, we get the formula:

\[
p(y_i \mid x) = \frac{p(x \mid y_i)p(y_i)}{p(x)} \tag{2.3}
\]

If it is assumed that each attribute of the item \( x \) to be classified is conditionally independent, there is a formula:

\[
p(x \mid y_i) = p(a_1 \mid y_i)p(a_2 \mid y_i)\cdots p(a_m \mid y_i) \tag{2.4}
\]

Further, the Bayesian formula described in equation is transformed into a formula:

\[
p(y_i \mid x) = \frac{p(y_i) \prod_{j=1}^{m} p(a_j \mid y_i)}{p(x)} \tag{2.5}
\]

This is the core idea of the naive Bayes classifier.

3. Suspects prediction results based on naive Bayesian method

The data set we selected originates from question C of mathematical modelling for Chinese graduate students in 2018. The five terrorist organizations and individuals that are more harmful to terrorist attacks are selected as follows: No. 1 Taliban, No. 2 Islamic State of Iraq and the Levant (ISIL), No. 3 Al-Shabaab, No. 4 Kurdistan Workers’ Party (PKK), No. 5 Boko Haram.

Python programming principle: in order to solve task 2, we consider using naive Bayesian classification, select the terrorist attacks that occurred in 2015 and 2016, which have not yet been organized or claimed responsible by individuals to form a training set, and use Python's sklearns machine learning library to calculate.

Data Processing: Regarding terrorists as categories, in the case of multiple perpetrators of the same incident, the method of adding events is adopted, that is, each perpetrator commits a crime independently. In addition, the GaussianNB model in sklearns only allows float type data, so it is necessary to convert strings into floating-point data and store corresponding tables.

Computation: Input the training set training model, then input the events to be analyzed, and get the probability of all suspects in each event, that is, suspect degree.

The calculation results obtained by Python programming are shown in Table 1.
Table 1 Terrorist’s suspicion of typical events

| Suspect No. 1 | Suspect No. 2 | Suspect No. 3 | Suspect No. 4 | Suspect No. 5 |
|---------------|---------------|---------------|---------------|---------------|
| 201701090031  | 1             |               | 2             |               |
| 201702210037  | 1             | 2             | 3             |               |
| 201703120023  | 3             | 2             |               | 1             |
| 201705050009  | 3             | 2             |               | 1             |
| 201705050010  | 3             | 2             |               | 1             |
| 201707010028  | 3             | 2             | 4             | 1             |
| 201707020006  | 1             |               | 2             |               |
| 201708110018  | 1             | 2             | 3             |               |
| 201711010006  | 1             | 2             |               | 3             |
| 201712010003  | 1             |               | 2             |               |

4. summary
This paper presents a mathematical model based on naive Bayesian method. Researchers can not objectively obtain data in the process of collating and collecting terrorist organization attack data, and some data in reality are very important for terrorism research. The results of the model can help researchers to study terrorist attack data and follow-up researchers to use data for terrorism-related research.

References
[1] Alberto, A. (2005) Poverty, Political Freedom, and the Roots of Terrorism. American Economic Review. 95: 50-56.
[2] Aradau C., van Munster R. (2012) The Time/Space of Preparedness. Space and Culture. 15: 98-109.
[3] Luo C.C., Wu D. (2013) Catastrophe Risk Analysis: A Financial Perspective. Human and Ecological Risk Assessment: An International Journal, 19: 1372-1384.
[4] Monahan J. (2015) The individual risk assessment of terrorism. Social Science Electronic Publishing, 18: 167-205.
[5] Goldszmidt M. (1997) Bayesian Network Classifiers. In: Provan, G., Langley, P., Smyth, P., Machine Learning. Kluwer Academic Publishers., Netherlands. pp. 131-163.
[6] Choudhary S., Koul S., Mishra S. (2016) Collaborative job prediction based on Naive Bayes Classifier using python platform. In: International Conference on Computation System & Information Technology for Sustainable Solutions. Bangalore, pp. 165-178.
[7] Zhang N, Wu L, Yang J. (2018) Naive Bayes Bearing Fault Diagnosis Based on Enhanced Independence of Data. Sensors, 18: 463.