Proceedings of the 4th International Conference on Computing and Data Science (CONF-CDS 2022)
DOI: 10.54254/2755-2721/2/20220537

Machine Learning based Terrorist Attacks Prediction Algorithm

Wang Yuanyuan

Officers college of PAP, Chengdu, China

409102408@qq.com

Abstract. Terrorist attacks are spreading rapidly all over the world, which has caused heavy casualties and property losses. Therefore, it is very necessary to predict precisely the types of terrorist attacks and provide important information for counter-terrorism work. The existing research on terrorist attacks only analyzes a few characteristics, which leads to the limitations of the research. Therefore, this paper not only adds some continuous and discrete features, but also adds unstructured features based on the current research, which can accurately describe terrorist attacks. This paper proposes a random forest method based on threshold for feature selection because of the added characteristics of terrorist attacks, and then uses XGBoost model to predict the types of terrorist attacks. This method helps to prevent terrorist attacks and reduce the damage caused by terrorist attacks, and provide decision support for the counter-terrorism department to take measures in advance.

Keywords: Feature selection, Random forest, Terrorist attacks.

1. Introduction

Since the September 11th event, terrorist threat of the global has fluctuated in waves in the past 20 years, and terrorism has became the public enemy of all mankind. According to the statistics of global terrorism database, more than 120000 global terrorist attacks were recorded from 2001 to 2019. All countries are facing the threat of terrorist attacks, and human life and social stability have been seriously influence. The analysis and prediction of the types of terrorist attacks will help the counter-terrorism department to find the signs of terrorist attacks in time and take targeted measures in advance for effective prevention, so as to minimize the losses caused by terrorist attacks.

The usual methods of data mining play an important role in the analysis of terrorist attacks. Literatures[1]-[3] used clustering model to study terrorist attacks from the aspects of suspected degree, suspected object, suspected organization, attack mode and attack target risk. With the rapid development of artificial intelligence in recent years, many scholars at home and abroad apply the advanced technology in the field of machine learning to the analysis and prediction of terrorist attacks. Literatures[5]-[10] have studied terrorist attack organizations, suspects, casualties, success probability of the attack, attack trend and so on by introducing various methods of machine learning. However, the existing research has two deficiencies: one is to reflect the characteristics of terrorist attacks less, the other is to predict the types of terrorist attacks less.

DOI: 10.54254/2755-2721/2/20220537

Because of such problems, this paper proposes a terrorist attack type prediction method based on feature selection. This method uses the feature selection method based on random forest to select features firstly, and then uses XGBoost to classify and predict the types of terrorist attacks.

This paper is structured as follows. Section 2 introduces the previous work of researchers in this field. Section 3 describes the terrorist attack type prediction model from three modules: data preprocessing, feature selection and classification prediction, and focuses on the research methods used in the process of model work. Section 4 discusses the research limitations and future research directions.

2. Related works

The research on terrorist attacks events mainly has two aspects at present: one is the traditional data mining method, the other is machine learning. Literature [1] analyzed the terrorist attacks for which no organization claimed responsibility in 2017 adopt clustering. Literature [2] established a clustering model based on improved k-means algorithm, which can judge the suspected object according to the suspected degree of terrorist attack. Literature [3] constructs a risk assessment model of K-means clustering method to quantitatively analyze the risk of attack mode and attack target in terrorist attack cases in civil aviation system. Literature [4] proposed a terrorist attack organization prediction method based on feature selection and hyper parameter optimization, which improved the classification and prediction performance of terrorist attack organizations. Literature [5] uses machine learning method to extract various features of terrorist attacks and predict one or more suspects of terrorist attacks. Literature [6] used random forest algorithm to classify and predict whether terrorist attacks caused deaths and injuries, and then used ridge regression algorithm to predict the specific number of deaths and injuries caused by events. Literature [7] predicts the trend of terrorist attacks around the world by analyzing classification technologies such as lazy tree, multi-layer perceptron and naive bayes. Literature [8] predicts the success probability of terrorist attacks by establishing a deep neural network model. Literature [9] proposed a hybrid prediction algorithm as a decision support tool for terrorism. The feature selection method based on random forest and XGBoost prediction model used in this paper also have relevant research in early warning and decision analysis of other industries. Literature [10] classifies urban greening tree species through random forest feature selection, which improves the classification accuracy of existing similar tree species. Literature [11] uses the feature selection method of random forest to select the optimal variables to establish a model for identifying Chinese spam messages on the basis of natural language processing technology and text classification algorithm. Literature [12] used XGBoost model to predict fan blade icing, and achieved good results in predicting the occurrence of early blade icing fault. Literature [13] analyzed and predicted the online short rent market price by using XGBoost model.

3. Prediction model of terrorist attack types

The terrorist attack type prediction model studied in this paper mainly includes three parts: data preprocessing, feature selection and classifier classification. The data preprocessing module is responsible for data discretization, one-hot coding and mean coding. The processed data are selected by random forest algorithm, and then in the classifier module XGBoost model is used to classify and predict the types of terrorist attacks.

DOI: 10.54254/2755-2721/2/20220537

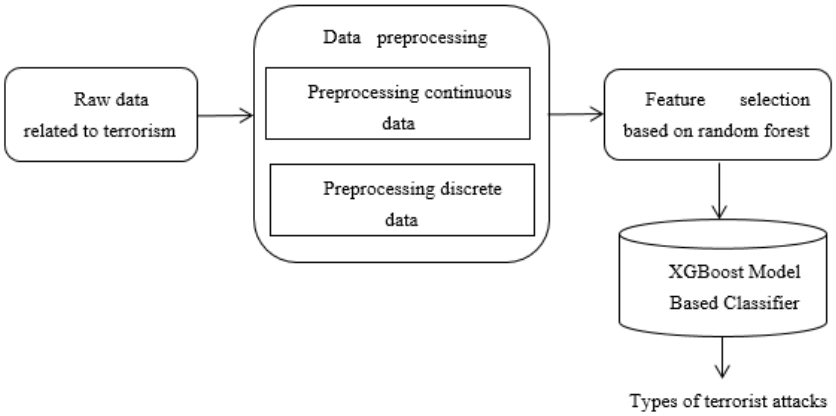

Figure 1. Prediction model of terrorist attack types

3.1 Data preprocessing

The original data describing terrorist attacks in reality are really not standard data that compliant data processing requirements, also include unstructured data and continuous data. These data are preprocessed to obtain structured and standardized data. The main contents of data preprocessing include data cleaning, data integration, data transformation and data reduction [14].

Combined with the real terrorist attack cases published on the Internet and the <<Basic Knowledge of Identifying Religious Extremist Activities (75 specific manifestations)>>, this paper screened 16 attribute features. They are the country, city, region, number of suspects, gender of suspects, age of suspects, height of suspects, facial features of suspects, religious background, whether to kidnap hostages, whether to carry contraband, type of weapons, number of deaths, number of injuries, special signs / clothes and special signs. The suspect's age and height are continuous data. The suspect's facial features, special identification / clothing and special identification products are unstructured data, and the rest are discrete data.

Continuous data Preprocessing. This paper chooses the discretization algorithm based on clustering to discretize the number, age and height of terrorist attack suspects. Take the age discretization of suspects as an example. According to the analysis of terrorist attacks publicly reported by the media, the suspects of terrorist attacks are widely distributed between the ages of 5 and 65. The child are mainly forced by family conditions or lured by religious brainwashing to participate in terrorist attacks, and the type of participation is single. Youth and adults are the main body of terrorist attacks. They participate in many types of terrorist attacks and destructive. One is that the elderly participate in terrorist attacks under family conditions, and the type of participation is single, the other is the senior leaders of terrorist organizations, who organize and plan terrorist attacks and participate in many types of terrorist attacks [15]. Therefore, the density based clustering method can better maintain the distribution information of age attributes. Firstly, several groups are designated according to the suspect's age. In the clustering process, the distribution of the suspect's age in the data space is fully considered, and then the clusters obtained by clustering are processed and merged into the continuous attribute values of one cluster for unified marking. The specific algorithm based on density is as follows:

DOI: 10.54254/2755-2721/2/20220537

Algorithm: DBSCAN

Input: dataset $D=\{x_1, x_2 \mathrm{K}, x_m\}$; neighborhood parameters $(\varepsilon, MinPts)$, sample distance measurement method

Process:

1. initialize core object set $\Omega=\varnothing$
2. for i in D
3. determine the ε-domain subsample set $N_\varepsilon(x_i)$ of sample x_i;
4. If $|N_\varepsilon(x_i)| \geq MinPts$
5. add x_i to the core object sample set Ω
6. While $\Omega \neq \varnothing$
7. randomly select a core object $x_p \in \Omega$
8. density up to sample of $x_p \rightarrow$ generate cluster C_k
9. $\Omega=\Omega / C_k$
10. End while

Output: cluster partition $C=\{C_1, C_2, \Lambda, C_K\}$

Figure 2. DBSCAN

The main process of density based clustering algorithm is to assume neighborhood parameters $(\varepsilon, MinPts)$, the ε-domain $N_\varepsilon(x_i)$ of each sample in the data set $D=\{x_1, x_2, \ldots\ldots, x_m\}$ is found by measuring the sample distance, and the core object set Ω is determined. Then randomly select a core object x_p as the seed, find out all the samples up to its density, generate a cluster C_k, and then remove the core object contained in the cluster C_k from the core object set Ω. After removal, randomly select a core object from the updated set Ω as the seed to generate the next cluster, and repeat the above process until it is empty.

Discrete data Preprocessing. For discrete data, they need to be encoded and converted before they can be used. One-Hot Encoding is a method to convert discrete values into several binary columns. Each discrete value is assigned a unique binary vector composed of 1 and 0. One-Hot Encoding is used to encode the sex of the suspect. Male and female are represented by two 2-bit codes of 10 and 01 respectively. As shown in Figure 3.

DOI: 10.54254/2755-2721/2/20220537

However, when the data set increases, it is obvious that this method not only adds a large number of dimensions to the data set, but also has too much redundant information. The unusual sparse matrix consumes memory and increases the training time, which makes it difficult to deal with the optimization problem. Therefore, other encoding methods need to be considered when dealing with discrete data such as the country and city where the event occurred.

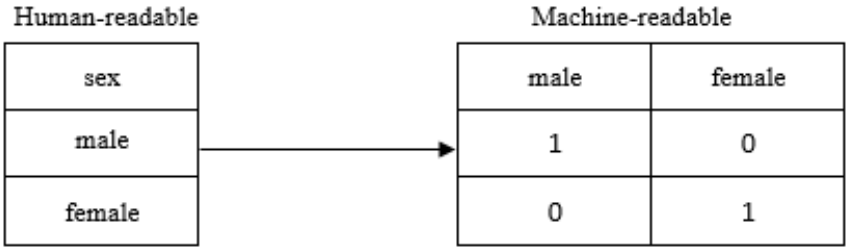

Figure 3. One-Hot Encoding the suspect's gender

The mean coding is an efficient supervised coding method for high cardinality discrete data. For example, the city where the terrorist attack occurred in this paper which is categorical feature. This feature has many values in reality. The mean coding is to express each feature x_i in the feature X (city) as its corresponding target value probability, where y_i is target value probability.

3.2 Feature selection

Feature selection is to select the most suitable features from multiple existing features. While ensuring that important features are not lost, it can not only effectively solve the problem of dimension disaster, but also reduce the difficulty of follow-up learning tasks. The two keys links of feature selection are feature subset search and subset evaluation. Different feature subset search mechanisms and different subset evaluation mechanisms can be combined to obtain different feature selection methods. The feature selection algorithm based on random forest is adopted in this paper. It is an embedded feature selection method. Embedded feature selection is automatic feature selection in the process of learner training, and they are completed in the same optimization process. Random forest is an extended variant of Ensemble Learning Bagging method. It is easy to implement and has good generalization ability. At the same time, it can also be used as a feature selection method. The feature selection process based on random forest is divided into three steps:

Step 1: Constructing random forest

(1) n samples are randomly selected from the original data set through bootstrap sampling, and a total of k samples are sampled to generate k training sets. The samples not sampled in each sampling form the out of bag data (OOB).

(2) Constructing k decision trees based on k training sets, for any decision tree k_i, n ($n < m$) features are randomly selected from m features in each split. The best feature is selected for splitting by calculating the information gain ratio of each feature.

(3) The random forest is composed of generated multiple decision trees. The samples to be tested are input into the random forest, and the classification results are finally determined by voting.

Step 2: Evaluate feature importance

(1) The out of bag data error is calculated for each decision tree in random forest, denoted as OOB_{error1}.

(2)Randomly change the xj（ j=1,2,……,m） characteristic of all samples in the out of bag data, and calculate the out of bag data error again, denoted as OOBerrorj.

(3) The importance of feature xj is

DOI: 10.54254/2755-2721/2/20220537

$$\sum_{j=1}^{m}(OOB_{errorj}-OOB_{error1})/k \tag{1}$$

where k represents the number of decision trees in the random forest.

(4) m features are sorted in descending order of importance.

Step 3: Feature selection

(1) The Sequential Backward Selection algorithm is used to eliminate the last feature in the importance ranking in the feature set.

(2) Successive iteration and calculate the classification accuracy.

(3) The feature set with the least number of features and the highest classification accuracy is obtained as the result of feature selection. The selected characteristics according to a certain threshold are important factors related to the types of terrorist attacks in this paper.

3.3 Classification model

XGBoost is one of the boosting algorithms. It is a gradient lifting tree algorithm based on CART regression tree. Under the same conditions, compared with similar algorithms, XGBoost model greatly improves the calculation speed, and is better at capturing the dependencies between complex data. The objective function of XGBoost includes loss function and regularization item. The mathematical expression is as follows.

$$obj=\sum_{i=1}^{n}L(y_i,\hat{y}_i)+\sum_{k=1}^{K}\Omega(f_k) \tag{2}$$

where the loss function denoted as

$$\sum_{i=1}^{n}L(y_i,\hat{y}_i) \tag{3}$$

y_i is predicted value, $\hat{y}_i$ is real value, f_k is K basis model, Ω is regularization item of model. It is optimized in two aspects: on the one hand, the loss function is a second-order Taylor expansion of the error,

$$\sum_{i=1}^{n}[L(y_i,\hat{y}_i^{(t-1)})+g_i f_t(x_i)+\frac{1}{2}h_i f_t(x_i)+\frac{1}{2}h_i f_t^{2}(x_i)] \tag{4}$$

where g_i and h_i respectively represent the first-order and second-order partial derivatives of the loss function at the first sample point, so the prediction result is more accurate. On the other hand, the regularization item is introduced. The regularization item is a penalty mechanism, which is mainly used to control the complexity of the model and prevent over fitting.

3.4 Implementation issues

This paper uses the feature selection method based on random forest and XGBoost prediction model to predict the types of terrorist attacks. Firstly, the original data is preprocessed, including continuous data, discrete data and unstructured data to obtain structured and standardized data. Then the feature selection method based on random forest is used to select the features of the processed data. The selected features have important correlation with judging the type of terrorist attack. Finally, the selected optimal feature subset is input into the XGBoost model to classify and predict the types of terrorist attacks, and finally the prediction results are obtained.

DOI: 10.54254/2755-2721/2/20220537

4. Conclusion

In order to solve these problems that there is less research on the prediction of the types of terrorist attacks and less reflection of the characteristics of terrorist attacks,combined with the open actual cases, the attribute characteristics reflecting terrorist attacks are summarized and added. Compared with the original, the added attribute features can more significantly describe terrorist attacks, which is more conducive to assist in predicting terrorist attacks. These attributes include continuous data, discrete data and unstructured data. The density based clustering method is used to preprocess the continuous data. One-Hot Encoding is used to encode the low radix discrete data, and mean coding is used to preprocess the high radix discrete data. For the processed attribute features, the random forest feature selection method is used for feature selection, and then the XGBoost model is used to predict the types of terrorist attacks.

As space is limited, this paper mainly studies structured data processing. In reality, terrorist attacks will involve a large amount of unstructured data. Combine with computer vision, natural language processing and other technologies to study and solve the problem of unstructured data in the future work. Besides different feature selection methods and different prediction models will be used for comparative analysis. The combination of machine learning and counter-terrorism action will help to combat terrorist activities and help to carry out counter-terrorism efficiently.

References

[1] Jiang L, Chen Y, Yu L et al. A Data Analysis Method for Anti-terrorism Based on Clustering. Information Research 6(260), 74-77(2019).

[2] Yang Z, Li Y, Zhong Z. Research on Judgment of Suspects of Terrorist Attack Based on Data Mining. Information Research 4(258), 35-40(2019).

[3] Liu M. Risk Assessment of Civil Aviation Terrorism Based on K-means Clustering. Data Analysis and Knowledge Discovery 10(22), 21-26(2018).

[4] Xiao Y, Zhang Y. Terrorist attack organization prediction method based on feature selection and hyperparameter optimization. Journal of Computer Applications 40(8), 2262-2267(2020).

[5] Li H, Zhang N, Cao Z et al. Terrorist Prediction Algorithm Based on Machine Learning. Computer Engineering 46(2), 315-320(2020).

[6] Qiu L, Han X, Hu X. Study on method of consequence prediction for terrorist attacks based on machine learning. Journal of Safety Science and Technology 16(1), 175-181(2020).

[7] Kumar, Vivek, et al. "A Conjoint Application of Data Mining Techniques for Analysis of Global Terrorist Attacks." International Conference in Software Engineering for Defence Applications. Springer, Cham, 2018.

[8] Onyekachi, Uche Stanley, and Tsopze Norbert2& Ebem Deborah Uzoamaka. "Data Mining Approach to Counterterrorism."

[9] Soliman, Ghada MA, and Tarek HM Abou-El-Enien. "Terrorism Prediction Using Artificial Neural Network." Rev. d'Intelligence Artif. 33.2 (2019): 81-87.

[10] Wen X, Zhong A, Hu X. The Classification of Urban Greening Tree Species Based on Feature Selection of Random Forest. Journal of Geo-Information Science 12, 1777-1786(2018).

[11] Zhao Z, Fu X, Jin X et al. Spam Message Recognition Based on Random Forest Feature Selection. Computer and Information Technology 6, 24-26(2018).

[12] Cao Y, Zhu M, Wang X. Wind Turbine Blade Icing Forecast Based on Feature Selection and XBGoost. Electrical Automation 41(3), 31-33,118(2019).

[13] Cao Rui, Liao Bin, Li M et al. Predicting Prices and Analyzing Features of Online Short-Term Rentals Based on XGBoost. Data Analysis and Knowledge Discovery 6: 51-65(2021).

[14] Dong S. Data preprocessing technology in data mining. China Computer&Communication 16,144-145(2018).

[15] Li Y, Mei J, Qin G. Research on Data Preprocessing in the Field of Counter Terrorism Intelligence Analysis. Information Science 35(11), 103-107,113(2017).