International Journal of Innovative Research & Growth
A Peer Reviewed Journal
E-ISSN: 2455-1848
homepage: http://www.ijirg.com/ijirg/
Volume: 13 Issue: 2, April 2024

Analyzing Global Terrorism Database for Identification of Terrorist Group

P. L. Verma[1], Sanjay Dwivedi[2*], Shivendra Kumar Dwivedi[3]
[1]Department of Physics, Govt. Vivekanand P.G. College, Maihar-485771, M.P., India
[2]Govt. SGS P.G. College, Sishi-486661, M.P., India
[3]Sharda Mahavidyalaya, Sarlanagar-485114, M. P., India
E-mail: sanjaydwivedi_sidhi@yahoo.co.in
* Corresponding Author

Article Info
Received 21 February 2024
Received in revised form 31 March 2024
Accepted for publication 5 April 2024
DOI: 10.26671/IJIRG.2024.2.13.103
Cited as
Verma, P. L., Dwivedi, S., and Dwivedi, S. K. (2024). Analyzing Global Terrorism Database for Identification of Terrorist Group. *Int J Innovat Res Growth*, 13, 42-48.

Abstract

The objective of this paper is to analyze different classification algorithms in data mining to predict terrorist group involved in an incident. Terrorist groups are highly dynamic and mysterious, which makes it challenging to track their activities and prevent incidents. Prediction of terrorist group using chronological data of terrorist activities has been less explored due to the lack of detailed terrorist data which contain terrorist group's attacks and activities. The study tries to investigate the GTD through classification method for pattern discovery. This research paper proposes a framework for terrorist group prediction that is based on data mining classification techniques. The framework has been validated with the experimental results using WEKA tool.

Keywords: - GTD, Counter Terrorism, Terrorist Group Prediction, Classification, WEKA.

1. Introduction

India had faced lots of challenges such as population growth, poverty, starvation, illiteracy, gender-inequality and many more. However, terrorism is highly dangerous till now affecting the mankind and humanity. Attack at the parliament of India on 13th December 2001 was the black day in Indian history and raised concerns about terrorism among citizens. This terrible incident highlighted the urgent need for collaboration between researchers and intelligence community to counter terrorism at large scale. Terrorist group prediction for an incident is a difficult task due to the lack of clear and precise terrorist past data. Data mining classification techniques are widely used to resolve the issue of terrorism.

In this paper training the models that learn to detect the type of terrorist group involved in the incident in the course of GTD analysis. The experiments are performed on the part of Global terrorism database (GTD) as it is a repository of terrorist activities occurred around the world. It is believed that GTD has various appealing patterns still unseen and the full potential of this resource is still to be revealed.

This study tries to investigate the GTD through classification method for pattern discovery. The given framework trains the classifiers to learn patterns of the terrorism events and classify the new event from the GTD as a specific type of terrorism incident. The classifiers are trained by training set to make predictions from the available data. In this study experiments are performed using three different classifiers i.e. decision tree (J48 WEKA implementation of C4.5), Support Vector Machine and Naive Bayes.

10-fold cross validation method is used for experimental analysis and all used classifiers on the GTD are shown here. In this study Waikato Environment for Knowledge Analysis (WEKA) is used for applying text mining techniques.

1.1 Overview of Global Terrorism Database (GTD)

The Global Terrorism Database is an open-source database that contains information about terrorism that occurred between years 1972-2015 across India. In present research data between periods of 1995-2015 is taken for study [1].

Data on terrorist attacks have special challenges like-

- Coarse measurements; many categorical and qualitative variables.
- Important variables missing: intentions and strategies of the terrorists.

2. Related Work

Global Terrorism Database is a large collection of terrorism incident data in all over the world. It is a good source for counter-terrorism and criminology research. A number of researchers have analyzed the dataset and presented their useful findings in the literature. In this paper some of them are discussed.

Dugan et al. [2] have used GTD for analyzing hijacking incidents before 1986. The authors used continuous time survival analysis to estimate impact of counter-hijacking interventions on the hazard of differently motivated hijacking attempts and logistic regression analysis to model the predictors of successful hijackings. The authors found that the policy interventions examined significantly decreased the likelihood of non-terrorist but not that of terrorist hijackings.

The Author I. Rizwan et al. [3] compare two different classification algorithms namely; Naive bayes and Decision Tree for predicting "Crime Category" for different states in USA. In the experiment, 10-fold cross validation was applied to the input dataset, separately for both Decision Tree and Naive bayes to test the accuracy of the classifiers. It showed that Decision Tree algorithm out performed Naive bayes algorithm and achieved 83.951% accuracy in predicting "crime Category".

The author G. Faryral et al. [4] have proposed a novel ensemble framework for the classification and prediction of terrorist group in Pakistan that consist of four base classifiers namely; NB, KNN, ID3 and Decision Stump (DS). Majority vote-based ensemble technique is used to combine these classifiers. The results of individual base classifiers are compared with the majority vote classifier and it is determined through experiments that the new approach achieves a considerably better level of accuracy and less classification error rate as compared to the individual classifiers.

Greenbaum *et al.* [5] have used the GTD to analyze the impact of terrorism on Italian employment and business during 1985 to 1987. The authors concluded that terrorist attacks reduced the employment following the year of attack. The authors used terrorist attacks data from 1970 to 2004. In the article, the authors have tried to show the characteristics of global terrorism. The authors also included an analysis showing the link between the terrorism and political affairs in the country.

A study by Nizamani et al. [6] presented a semantic based news analysis method using a technique known as semantic role labelling. Paper deals with the analysis of news summary, therefore, we also here discuss some related work in connection to the news analysis The study dissects the new reports in order to highlight important information from the reports.

In this paper, text mining approach is applied to the major variable of the dataset that is the summary of terrorism incident. We try to extract information about type of terrorism incident and associated terrorist group from the GTD. We experimentally show that classification techniques can learn from the available dataset to detect the incident type and involved terrorist group. The next section presents various classification algorithms used in this study.

3. Classification Algorithms

Classification is a kind of supervised machine learning algorithm [7]. It takes training examples as input along with their class labels. It can be defined by following equations:

$$D = \{t_1, t_2, \ldots t_n\} \quad (1)$$
$$t_i = \{a_1, a_1, \ldots, a_m\} \quad (2)$$
$$C = \{c_1, c_2, \ldots, c_k\} \quad (3)$$

Where *D* is a dataset consisting of *n* training examples, t_i is a training example, each a_i is an attribute, m is the total number of attributes and c_i is a class and k is the total number of classes. With respect to our terrorism incident type detection *D* is collection of *22235 terrorism incidents*, each terrorism incident t_i comprises of *5345 attributes* a_i and C is a set of terrorism incident type and total number of incident types *k* is 9.

3.1 Decision Tree

Decision tree is a kind of divide and conquer algorithm. A decision tree consists of finite number of nodes—internal and external nodes. Each internal node corresponds to an attribute selected by some measure of algorithm like information gain or gain ratio that divides the training examples into the parts according to the values of that attribute. For example, if the attribute has three possible values, then there will be three branches going out from that node. The choice of attribute at particular level of hierarchy usually depends on the class distinction ability of that attribute. External nodes in the decision tree contain decisions or the class value. ID3 (Iterative Dichotomiser 3) is a kind of decision tree algorithm by Quinlan [8]. The algorithm suffers from over fitting and also the algorithm can only work on nominal values and discrete values and also ID3 does not deal with missing value. To overcome these issues of the ID3, Quinlan proposed C4.5 algorithm [9]. It uses pruning to overcome over fitting problem, uses discretization at a certain threshold to deal continuous data and ignores missing value attributes while making decisions.

3.2 Naive Bayes (NB)

Naive Bayes [10] is a simple and efficient technique used by data mining community for classification task. It uses Bayes theorem to estimate probabilities for each class to decide the class of an instance. NB assigns the maximum probability class label to a test instance [11].

3.3 Support Vector Machine (SVM)

SVM is considered to be the state of art classification algorithm. SVM is a supervised machine learning technique used for classification. SVM is based on Vapnik's [12] statistical learning theory. SVM has some unique features due to which it is considered as state-of-the-art in classification. It is considered well suitable for the task of text classification and hand written digit recognition. Its unique features for text categorization are: (i) It works well with high dimensional data; (ii) It can make a decision boundary by using only a subset of training examples called support vectors; (iii) It can also work well on non-linearly separable data by transforming the original feature space into a new feature space that is linearly separable by using the kernel trick. Joachims [13] has defined some properties of text classification for which SVM is the ideal choice of solution. SVM has a main limitation that it suffers from long running time when runs on large datasets.

4. Data Preprocessing

In this paper, we have used terrorism incidents occurred in India for a period of period of 1995 to 2015 from GTD. The necessary part of the dataset is transformed into the ARFF file, in which each instance represents an incident from GTD. ARFF is an abbreviation for Attribute Relation File Format used by WEKA [14]. From the GTD we have used only two fields of each incident namely- summary (a text field) which presents the description of the incident, and type of incident which takes a value from one of the types of terrorism incidents. The summary field is further pre-processed using text mining processes because it involves the free text. This further pre-processing is applied using WEKA utility (String to Word Vector). This utility performs text mining steps such as, tokenization, stop word removal and feature weighting, etc.

5. Framework for Prediction of Terrorist Group

Terrorism incident type detection is considered as a text classification problem. We carried out the task using classification algorithm. The process involves the training data, from which the patterns of the incident type are learned by the learning algorithms.

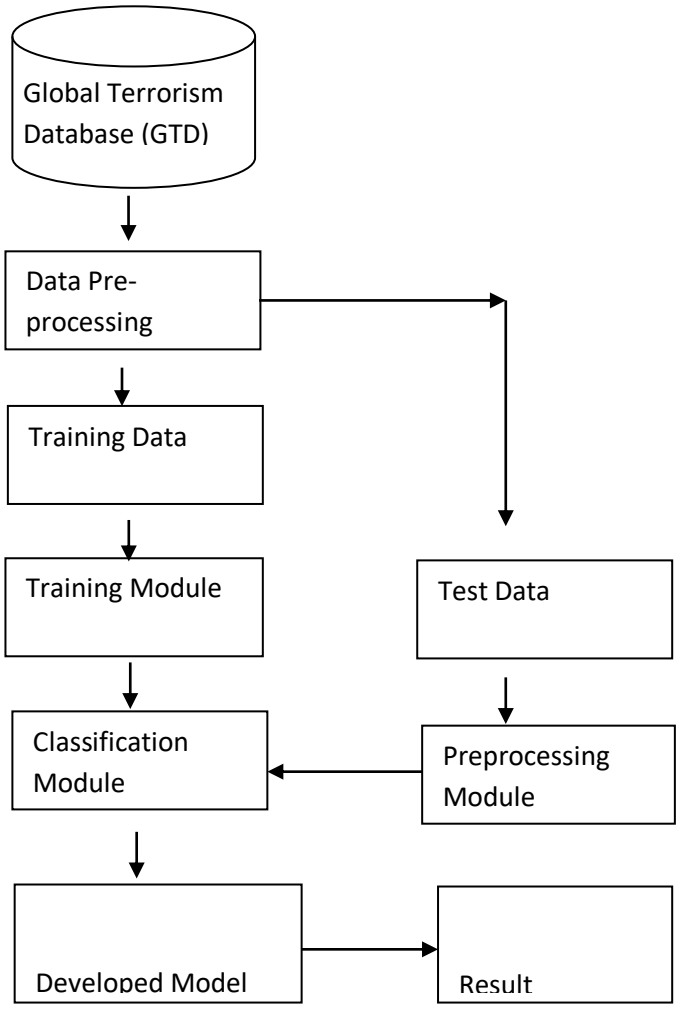

Figure-1 Framework for Identification of Terrorist Group

We provided the training incidents as training data which is comprised of the summary of incidents as well as type of the incident. After the training data is provided, pre-processing is performed, which makes the data in appropriate form for different classification algorithms. Once the classifier is trained on training data, the learning is applied for predicting the type of terrorism incident using only the abstract of incident. The complete process is demonstrated in Figure 1.

6. Experimental Data Analysis

The experiments are conducted using terrorism incident records from GTD between the periods of 1995 to 2015. Each terrorism record is comprised of an incident summary and a number of other features describing terrorism incident including the attack-type, weapon-used. For experiments total 7656 records have been taken. After pre-processing there are total *3345* distinguished features. A short description of the dataset is provided in Table-1. A detail of all the incident types, including the number of incidents of each type is demonstrated in Table-2. The experiments are conducted using three well-known classification algorithms, namely; Decision tree J48 (WEKA's implementation of C4.5), Naïve Bayes (NB) and Support Vector Machine (SVM). These are widely used classification algorithms very famous among research community. The evaluation method and evaluation measures used in the experimentation are described in the following sub-section.

Table-1 General information about dataset used in experiments

Incident Period	1998-2015
Total number of incidents	7656
Total number of features	3345
Total number of incident-type	9

Table-2 Incident type distribution in training data

Type of Incident	**No of Incidents**
Armed Assault	6797
Assassination	1167
Bombing Explosion	10731
Facility Infrastructure Attack	1820
Hijacking	59
Hostage Taking Barricade Incident	134
Hostage Taking Kidnapping	1111
Unarmed Assault	275
Unknown	141
Total	**22235**

Table-3 Terrorist Group Name distribution in training data

Type of Weapon used	**No. of Incidents**
Biological	36
Chemical	230
Radiological	13
Firearms	113

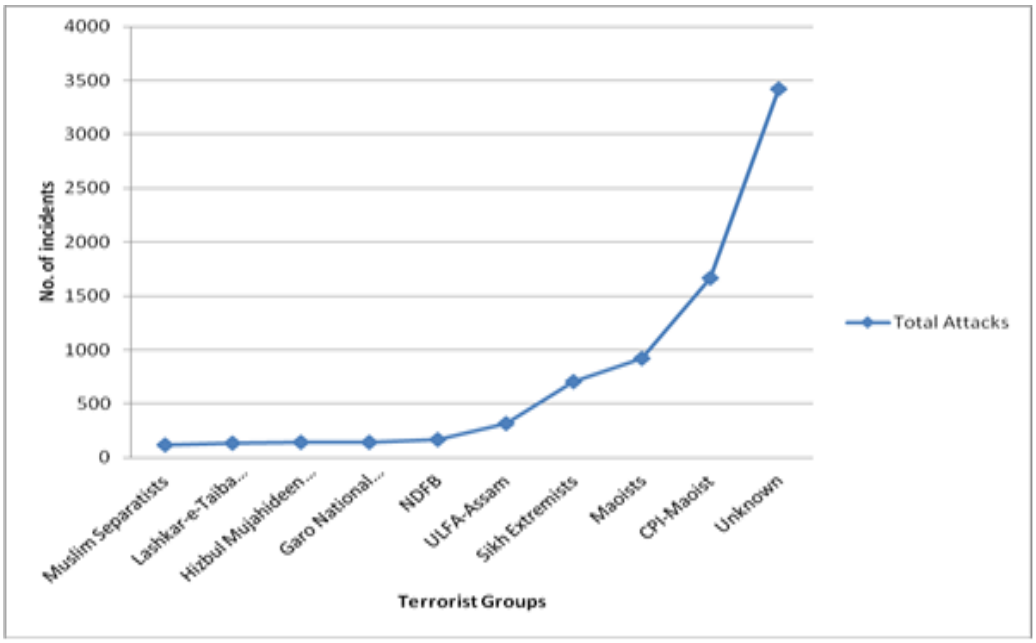

Figure-2 Main terrorist groups and attacks by them.

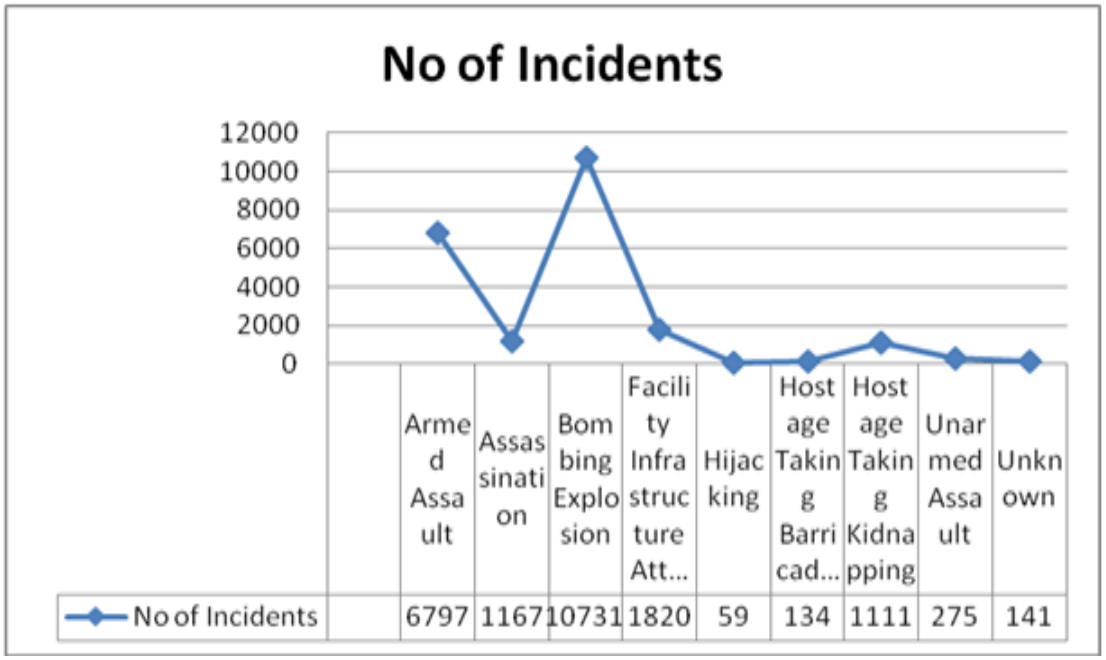

Figure-3 Attack types and number of incidents.

6.1 Evaluation Method

In present study 10-fold cross validation method is used for the purpose of evaluation of results. Tenfold cross validation splits the dataset in 10 subsets. It runs for 10 rounds, in each round 9 subsets are used for training and one of them is used for testing. In each round a new subset is chosen for testing. After 10 rounds the average accuracy of all the rounds is measured.

6.2 Evaluation Measures

The evaluation measures that we have used are accuracy, precision and recall [15]. These measures are calculated as follows:

$$\text{Accuracy} = (Tp+Tn) / (Tp+Tn+Fp+Fn.) \quad \ldots\ldots\ldots\ldots\ldots (4)$$
$$\text{Precision} = Tp/ (Tp + Fp) \quad \ldots\ldots\ldots\ldots\ldots (5)$$
$$\text{Recall} = Tp/(Tp + Fn) \quad \ldots\ldots\ldots\ldots\ldots (6)$$

Where *Tp* is the number of incidents correctly classified as particular class, *Fp* is the number of incidents that were incorrectly classified as a particular class. *Tn* is the number of incidents that were correctly classified as other class and *Fn* is the number of incidents that were incorrectly classified as another class.

7. Result

The experimental result in summarized form is depicted in table-4 and figure-4 and it clearly illustrates that from the GTD we can successfully detect terrorism incident type and associated terrorist group. The classification algorithms can extract this information successfully. It is clearly depicted in the figure that decision tree correctly detects 83% of incidents with a balance of precision and recall.

Table-4 Outcome of performance measures for all classifiers.

Performance Measures → / Classifiers ↓	TP	FP	Precision	Recall
Naive Bayes	0.687	0.233	0.747	0.687
Decision Tree	0.831	0.192	0.812	0.831
SVM	0.790	0.198	0.799	0.790

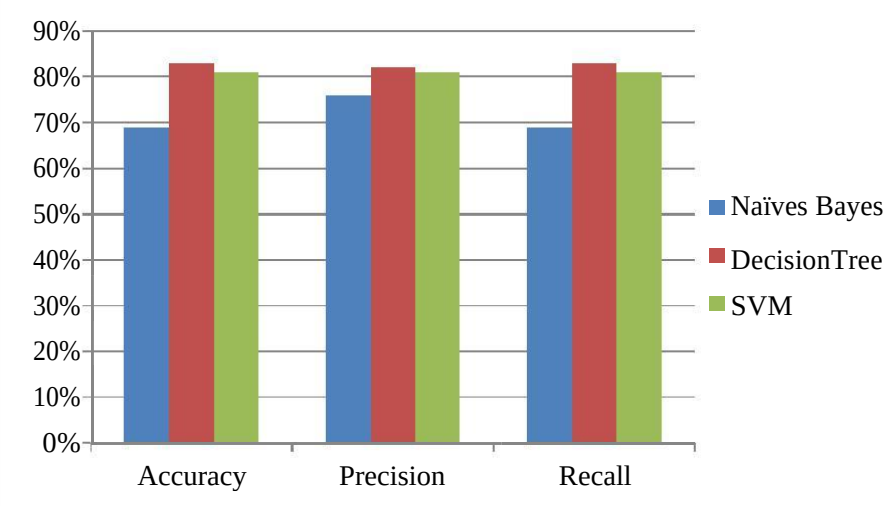

Figure-4 Experimental results.

8. Conclusions and Future Work

In this paper we applied data mining techniques and presented the experimental results for detecting terrorism incident types and associated terrorist group. We performed experiments using classification algorithms, such as decision tree, Naïve Bayes and state of the art SVM. Experimental results illustrate that the task can be successfully carried out using classification algorithms with satisfactory results. The results show that a high accuracy is achieved using J48 (decision tree) algorithm with a balance of precision and recall. SVM also achieved high accuracy, but it takes long running time when there is large dataset. The accuracy achieved using Naïve Bayes is lower comparatively but it runs faster. In future, we intend to incorporate the semantic knowledge which will have positive effect on the accuracy of the task. It is also included in our future plans to make use of spatio-temporal features from the dataset, in order to find the correlations among incident type, time and geo space.

Conflict of Interest

The Authors declares that there is no conflict of interest in this manuscript.

References

[1] http://www.start.umd.edu/gtd/about/
[2] Dugan, L., LaFree, G., and Piquero, A. R. (2005). Testing a rational choice model of air- line hijackings. Criminology, 43, 1031-1065.
[3] Rizwan Iqbal, Masrah, A. A. M. et al. (2013). An Experimental Study of Classification Algorithms for Crime Prediction. IJST, 6, 1-7. DOI: 10.17485/ijst/2013/v6i3.6
[4] Faryral, G., Wasi, B. H., and Usman, Q. (2014). Terrorist Group Prediction Using Data Classification. Proceedings of the International Conferences of Artificial Intelligence and Pattern Recognition, Malaysia.
[5] Greenbaum, R.T., Dugan, L., and LaFree, G. (2007). The impact of terrorism on Italian employment and business activity. Urban Studies, 44, 1093–1108.
[6] Nizamani, S., and Memon, N. (2012). Semantic analysis of FBI news reports. In Neural Information Processing (pp. 322-329). Springer Berlin Heidelberg.
[7] Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing surveys, 34, 1-47.
[8] Quinlan, J. R. (1986). Induction of decision trees. Journal of Machine Learning, 1, 81-106.
[9] Quinlan, J. R. (1993). C4.5: Programs for machine learning. Machine Learning, 16, 235-240.
[10] McCallum, D. J., and Nigam. K. (1998). A Comparison of event models for Naive Bayes text classification. Technical Report. Workshop on learning for text categorization, 41–48.
[11] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H. et al. (2007). Top 10 algorithms in data mining. Knowl Inf Syst,14, 1-37.
[12] Vapnik, V. N. (1995). The nature of statistical learning theory. 1st Edition, New York Springer, ISBN 978-1-4757-2440-0.
[13] Joachims, T. (2001). A statistical learning model of text classification for Support Vector Machines. Proceedings of the 24th annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 128-136. https://doi.org/10.1145/383952.383974
[14] Hall et al. (2009). The WEKA Data mining software: An Update. SIGKDD Explorations, 11.
[15] M. Hossin, Sulaiman,M. N., Mustapha, A., and Mustapha, N. (2011). A Novel Performance Metric for Building an Optimized Classifier. Journal of Computer Science, 7, 582-509.