Journal of Advances and Scholarly Researches in Allied Education
Vol. 16, Issue No. 6, May-2019, ISSN 2230-7540

Combating Terrorism in India through Prediction of Risk Associated With an Attack Using Investigative Data Mining Techniques

Sanjay Dwivedi[1]* Dr. Prabhat Pandey[2]

[1] Research Scholar

[2] OSD, Additional Directorate of Higher Education Department, Rewa Division, Rewa (MP)

Abstract – After "9/11" terrorist attacks, more advanced information technologies have been developed to counter-terrorism domain to enhance the performance of early warning system. Machine learning-based data mining is applied to predict terrorist activities hidden in terrorist incidents. Data mining classification techniques are mostly used to handle the problem of terrorism in India. An Ensemble Model building approach to classify worldwide terrorist attacks for the prediction of the fatality-risk associated with the terrorist attacks by utilizing different classifiers such as Decision Tree, Naïve Bayes, Lazy classifier IBK (k-NN) and Decision Table is used in the study. These algorithms are implemented in this study using WEKA a data mining tool and have attained fair accuracies in the perspective of classifiers' performance.

Keywords: Data Mining, Combating-Terrorism, Classification Algorithms, Ensemble Modelling, WEKA.

----------X----------

1. INTRODUCTION

Terrorist attacks are prevalent, leading to political instability and social insecurity across the nations. Terrorism is defined by the United Nations as "any action with a political ambition that is intentional to cause death or serious physical harm to civilians."

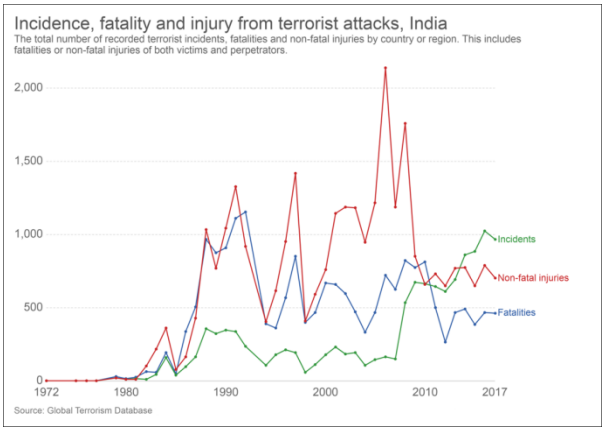

Figure 1.1: Terrorism Occurrences in India

Intelligent Data Analysis: An Automated Technique

The process of applying or using those patterns to analyze data and make predictions is not data mining. A more accurate term for those analytical applications is "automated data analysis", which can include analysis based on pattern queries, where the patterns can be developed from data mining or by methods other than data mining.

DATA MINING Vs KDD

Broadly conceived, data mining is a field of computer science that can be described as concerned with 'the extraction of hidden, previously unknown and potentially useful information from data' or 'extracting useful information from large datasets or databases'. Whereas the term Knowledge Discovery in Databases (KDD) commonly is used to cover the whole trajectory from data preparation up to implementation, the term 'data mining' tends to be restricted to the actual extraction process itself.

To improve counter-terrorism, several research works are developing capable and specific systems; data mining is not an exception. Immense data is floating in our lives, despite the fact that scarce availability of genuine terrorist attack data openly makes it complicated to fight terrorism. The thesis focuses on investigative data mining techniques and discusses the function of machine leaning in the counter-terrorism task.

2. LITERATURE REVIEW

PGIS defined terrorism as events involving "the threatened or actual use of unlawful power and violence by non-state actors to achieve a monetary, political, religious or communal goal through cruelty, fear or intimidation." [1]

M. DeRosa stated that investigative data mining techniques can be valuable tools for counterterrorism in many ways. One preliminary benefit of the data mining process is to aid in the essential task of accurate identification of the piece of data. [2]

B. Thuraisingham expressed that data mining grown as a new discipline for more than a few reasons. On the other hand data mining in combating terrorism is used due to the increase in the speed of computers processing power and the declining cost of technology as well. [3]

Galar M. et al. mentioned that the accuracy of the classifiers can be even more improved by means of the ensemble classifiers. Ensemble method is a machine learning approach where the predictive performance of multiple algorithms is better when matched up to the individual algorithms alone. Ensemble methods enhance the accuracy of a single classifier by training diverse classifiers and incorporating the decision to produce a single class label. [4]

Sanjay Dwivedi et al. said that the increasing open source terrorism activity databases have demanded for more logical analysis to be done on the nature of terrorism and terrorist activity. Though, a most important drawback of all these databases is that they have usually excluded attacks by the domestic criminals or organizations [5].

Muhammad and Kazi devoted towards analysing a GTD data set containing terrorism data that is specific to Pakistan from the year 1970 to 2014 by using a supervised machine learning techniques comprising an ensemble technique; a method of Meta machine learning using the Bayesian classifier and the decision tree classifier [6].

3. MATERIALS AND METHODS

Global Terrorism Database (GTD): An Overview

The GTD is an open-source and freely available database that compiles information on global terrorist activities. It is administered by the National Consortium for the Study of Terrorism and Responses to Terrorism (START), which falls under the University of Maryland and the U.S. Department of internal Security. GTD maintains a comprehensive database on terrorist incident occurred in period of 1970 to 2015. With a global outlook, scope of the GTD is extremely large.

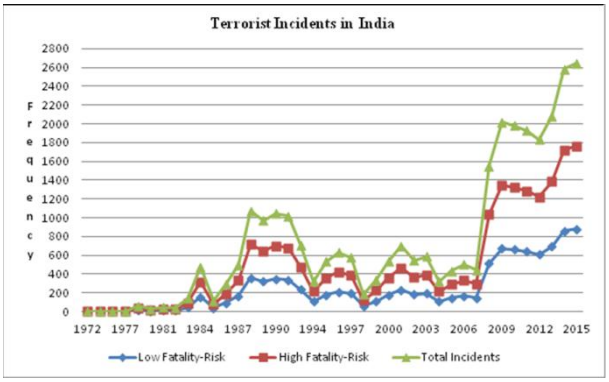

Figure 3.1: Showing terrorism frequency in India

Dataset Description:

Table : 3.1 Data Description

S. No.	Attribute Name	Data Type
1.	iyear	Numeric
2.	imonth	Numeric
3.	iday	Numeric
4.	extended	Numeric
5.	provstate	Nominal
6.	city	Nominal
7.	attacktype1_txt	Nominal
8.	targtype1_txt	Nominal
9.	weaptype1_txt	Nominal
10.	gname	Nominal
11.	nperps	Numeric
12.	fatality_risk	Nominal

Machine Learning Algorithms

Machine learning is the science that explores how algorithms can be constructed so that they can learn from data and make future predictions. These algorithms build a mathematical model based on some example inputs and use the model to make some predictions or decisions. Using the model any input can be mapped to a range of outputs. The flow of creating the model from the input data and the processes of mapping new inputs to expected outputs is demonstrated in Figure 3.2.

Machine Learning Algorithms

Machine learning is the science that explores how algorithms can be constructed so that they can learn from data and make future predictions. These algorithms build a mathematical model based on some example inputs and use the model to make some predictions or decisions. Using the model any input can be mapped to a range of outputs. The flow of creating the model from the input data and the processes of mapping new inputs to expected outputs is demonstrated in Figure 3.2

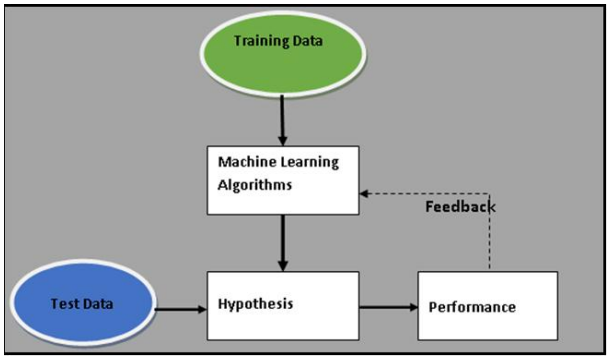

Figure 3.2: Machine Learning Workflow

Supervised learning:

The objective of supervised machine learning is to develop a model that makes predictions based on facts in the presence of uncertainty.

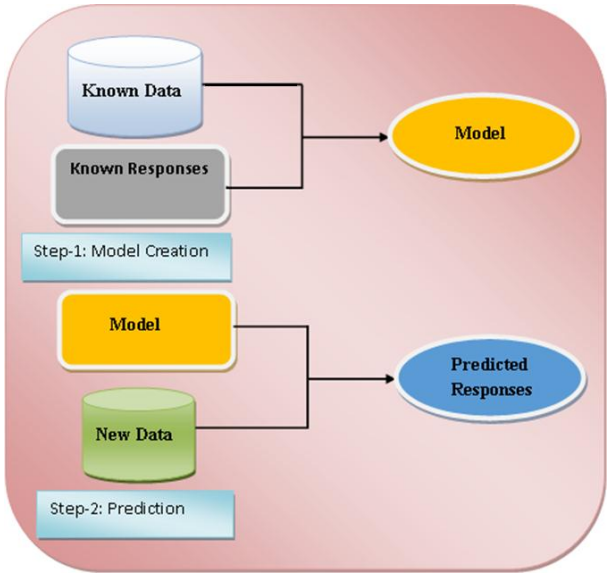

Figure 3.2: Supervised Machine Learning (Step-1 and Step-2)

4. MODEL CREATION AND EVALUATION

Development of the ensemble Model is two-phased processes. The first phase is all about the construction and evaluation of the base classifiers. In other words, in this phase best performing modelling algorithms are selected for further improvement by using ensemble modelling techniques in the next phase. There are many ensemble techniques that can be used in the development of improved efficient Model in the combination of selected base classifiers that perform exceptionally.

Building Ensemble Model

In statistics and machine learning, Ensemble Methods use Meta Machine Learning algorithms which are a combination of several base learners. Ensemble methods are used in various data mining applications in order to build an efficient Model that gives better predictive performance. In this research work the obtained accuracy of classifiers can be even further increased by means of the Ensemble classifiers. The classification models are constructed for all the datasets using famous machine learning classification algorithms such as J48, Naïve Bayes, IBK, Decision Tree and Decision Table. Individual classification Models are constructed for each algorithm and for the two datasets altogether.

As all the predictions are based on the very same dataset and they are certain to be prejudiced by a similar set of information, it is not guaranteed that confidence level can be achieved more than 99% at all. However, five experiments are performed, one for each different machine learning classification model. The ensemble model is built by combining the top four performing classifiers for better predictive performances and outcomes. The Ensemble model is built subsequently after all the four classifiers are trained in parallel and outcomes of all these are combined using the vote Meta classifier for giving the final results. The received Ensemble model is evaluated using 10-fold Cross-validation test mode to check the prediction capability using various metrics of performance that are already used for the base classifiers.

Table 4.1: Classification Accuracy of Ensemble Model on dataset GTD2, using Test Mode: 10-fold Cross-validation

Ensemble technique	Correctly Classified Instances		Incorrectly Classified Instances		Kappa-Statistics	Time taken to build Model (in seconds)
	number	%	number	%		
Vote	4654	82.18	1009	17.82	0.58	5.29

Here, from the table 4.1 it is clear that Ensemble model building approach works better than the individually developed models to make predictions using these base learners trained on the same dataset. Here, from all the developed models some good performing models are selected and combined to get better classification accuracy and other statistics. From the table 4.1 a conclusion can be drawn that the ***Vote*** Ensemble Model: a meta machine learning algorithm has shown better statistical figures and performed better in comparison to earlier developed an individual models in this research work.

The ***Vote*** model has achieved the 82.18% of accuracy with 4654 correctly classified instances in the dataset with 5663 instances in all. This ensemble model has scored approximately 0.6 as a value for kappa statistics, which is recorded highest among all other developed models.

Table 4.2: Confusion Matrix for the *Vote* Ensemble-Model on dataset GTD1

Total number of instances **(n) = 5663** **Test Mode:** 10-Folds Cross-validation		Fatality-Risk **(Predicted Class)**	
		High	**Low**
Fatality-Risk **(Actual Class)**	**High**	3471	429
	Low	580	1183

The Confusion Matrix in **table 4.2** shows that there are 3471 instances truthfully identified as positive (True Positive), 1183 instances are truthfully identified as negative (True Negative). While 580 instances are falsely identified as positive (False Positive) and 429 instances are falsely identified as negative (False Negative).

5. INTERPRETATION OF RESULTS ON PERFORMANCE METRICS

In this section, the performances of various prediction models developed during this research work and already described earlier in the previous chapter are analysed. It is already stated that for evaluating the performance of classifiers, the metrics- Accuracy, Precision, Recall, F-measure and ROC-Area have been taken in to consideration. The individual classification results are given for the dataset in the tables with the illustration as charts.

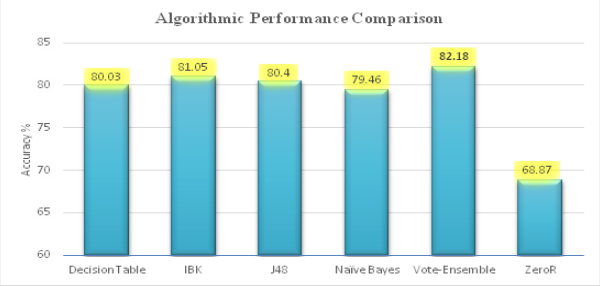

Figure 5.1: Accuracy of all Models

Assessment of Results

The data set was divided into two parts: one for training the evaluation model and the other for testing the model. To train and test the performance of the fatality-risk assessment model, a 10-fold cross-validation method was used. Hence, the data set divided into ten parts and takes nine of them as training data and one of them as test data. The sample data in each test set got a score between 0 and 1, which is verified by taking the threshold value from 0.1 to 0.9, and the evaluation index selected is Accuracy, Precision, R*ecall*, and *ROC-Area*. 10-fold cross-validations test mode conducted ten and sought the average value as an estimate of the final model accuracy, as shown in figure 5.1. It can be thought of as a measure of classifiers completeness. A low recall value indicates many false positives entries in the confusion matrix. In this study, it is clear from the table 5.4 that Decision Table Model is the top-performing algorithm and has shown the outstanding Recall value. In considering ROC value, it is noted that vote ensemble model has performed well on this metric and scored 0.867, which is recorded as highest among all other models developed.

6. CONCLUSIONS OF THE RESEARCH WORK

We took India as the research area and combined the best performing classification algorithm using ensemble technique through WEKA tool. The Model developed in this research work has attained results with fair accuracies ranging from 80-85 %. The results have given some idea that the northernmost parts of the India are at high-risk for cross-border terrorist activities. In fact, these areas have experienced a high incidence of terrorist attacks in recent years. The Lashkar-e-Taiba (LeT) and Hizbul Muzahiddeen (HM) separatist organizations have increasingly become religious extremist organizations, and the capacity of extremist organizations for transnational activities has increased. The penetration of the “Islamic State” in India especially in northern regions indicates that India may become a major battlefield in the international fight against terrorism. The results also show that Chhattisgarh, Jharkhand, Bihar, and coastal areas like Orissa, Kerala and Maharashtra are in High-risk areas of domestic terrorist attacks by CPI-Maoists. Thus, the next step in the prevention of domestic terrorism should focus more on these areas.

REFERENCES:

1. “The United Nations in the Fight against Terrorism,” United Nations, available at http:// www.un.org.

2. B. Thuraisingham (2003). Data mining, national security, privacy and civil liberties. SIGKDD Explorations.

3. Ghada M. Tolan and Omar S. Soliman (2015). An Experimental Study of Classification Algorithms for Terrorism Prediction. International Journal of Knowledge Engineering, Vol. 1, pp. 107-112. DOI: 10.7763/IJKE.2015.V1.18.

4. Sanjay Dwivedi and Prabhat Pandey (2018). “Analysis on Investigative Data Mining in the Counter-Terrorism, Technology and Transparency”, in Journal of Advances and Scholarly Researches in Allied Education [JASRAE] (Vol:15/ Issue: 5) DOI: 10.29070/2230-7540:167-172.

5. F. Bolz et. al. (2001). The Counterterrorism Handbook: Tactics, Procedures, and Techniques, CRC Press.

6. Muhammad, H.; Kazi, H. (2016). Use of Predictive Modeling for Prediction of Future Terrorist Attacks in Pakistan. *Int. J. Comput. Appl.*, *179*, pp. 8–16

7. The United Nations in the Fight against Terrorism," United Nations, available at http:// www.un.org.

8. J. Han: Data Mining Concepts and Techniques, second edition, Pg. No. 310 - 312.

Corresponding Author

Sanjay Dwivedi*

Research Scholar