Terrorism Analytics: Learning to Predict the Perpetrator

Disha Talreja, Jeevan Nagaraj, Varsha NJ and Kavi Mahesh
KAnOE - Centre for Knowledge Analytics and Ontological Engineering
Department of Computer Science and Engineering
PES University, Bangalore, 560085 India

drkavimahesh@gmail.com

Abstract
Data about terrorist attacks in India was analysed. Several machine learning algorithms were trained on the Indian subset of the Global Terrorism Database to learn to predict the perpetrator of a terrorist attack, given data about the types of attack, target and weapon in addition to the location, year and other attributes of the event. It was found that Support Vector Machine technique gave accuracy higher than 75% in predicting the perpetrators. This approach has the potential to aid investigating agencies and carries significant implications for national and international security.

Keywords
terrorism data; predictive analytics; support vector machine; India; perpetrator

I. Introduction

Irrespective of where a terror attack occurs across the globe, it brings out disgust, shock, fright, and uncertainty in people everywhere. The most prevalent aftermath of a terror attack is uncertainty with regard to things such as how they went about planning a major attack undetected, was the terror act an isolated instance or the first of a series, and finally, who were the perpetrators [1]. The aim of our project is to reduce this feeling of uncertainty as far as possible using machine learning and data analytics. Given some information related to the attack such as the targeted group, weapons used, type of attack, property destroyed and so on, we predict the perpetrator of the attack.

We build and test several models on the Global Terrorism Database (GTD) [2], maintained by the National Consortium for the Study of Terrorism and Responses to Terrorism (START). We first establish the principle attributes in the dataset and use them to predict the terrorist group that carried out the attack. Of all the models that we built, Support Vector Machine gave an accuracy higher than 75%.

The perpetrators of some of the deadliest terror attacks in India still remain unknown although a variety of information is available about the attack itself. Table 1 shows the available details for one attack. Our aim is to find out whether any of these details can help us predict the most likely perpetrator of the attack and, if so, who or which group is the perpetrator.

II. Problem Statement

Predicting a terrorist group responsible for an attack is the first step to counter terrorism. Once we find the group responsible for an attack, efficient strategies can be devised to apprehend the culprit. Commonly used techniques to discover the terrorist group responsible for an attack include email tracking, telephone signal information, social network analytics, etc. The focus of this research was to use GTD along with appropriate classifier algorithms to predict the terrorist group responsible for an attack.

III. Background

A. About the Dataset

The Global Terrorism Database (GTD) maintained by the National Consortium for the Study of Terrorism and Responses to Terrorism (START) at University of Maryland, College Park, MD, USA, has data about incidents from 1970 to 2014. The most recent release of GTD includes data for 2015 as well. The rich set of attributes of each terrorist event is documented in the Code Book released along with the GTD. Table 1 shows some of the important attributes of GTD data. It may be noted that numerous sub-attributes and copious details are available in GTD for several of the main attributes shown in the table.

TABLE 1. GTD ATTRIBUTES

GTD	Main attribute	Sub-attribute
1	GTD ID and Date	GTD ID
2		Year, Month, Day
3	Incident Information	Incident Summary
4		Inclusion Criteria
5		Related incidents
6	Incident Location	Country
7		Region, Province/State, City, Vicinity, Location description, Latitude, Longitude
8	Attack Information	Attack type, Successful attack
9	Weapon information	Weapon type, subtype and details
10	Target/Victim information	Target/Victim type and subtype
11		Name of entity
12		Nationality
13	Perpetrator information	Group name and subgroup name
14		Number, Number captured, Claimed responsibility, Motive
15	Casualties and Consequences	Total number of fatalities, perpetrator fatalities and injured
16		Details of property damage
17		Hostages and Kidnapping details
18	Additional Information and Sources	

978-1-5090-6367-3/17/$31.00 ©2017 IEEE

In order to deal with the high dimensionality of the data, we performed Factor Analysis of Mixed Data on the dataset to find those few attributes that potentially contribute most to predicting the perpetrators. These attributes are the following twelve:

- *iyear* – indicates year in which the incident occurred
- *attacktype1* – this attribute captures the general method of attack and reflects the broad class of tactics used, such as assassination, hijacking, kidnapping, bombing, etc.
- *targtype1* – this attribute captures the general type of victim such as business, government, police, military, airports and aircraft, etc.
- *targsubtype1* – a categorical variable that captures the more specific target category and provides the next level of designation for each target type.
- *weaptype1* – this attribute records the general type of weapon used in the incident such as chemical, radiological, nuclear, firearms, bombs/explosives, etc.
- *Latitude and longitude of the location of attack*
- *natlty1* – this attribute indicates the nationality of the target that was attacked, and is not necessarily the same as the country in with the incident occurred.
- property – categorical variable which indicates the existence of evidence of property damaged in the incident
- *INT_ANY* – a variable that describes if the attack was international on either logistic, ideological or miscellaneous dimensions
- *multiple* – cases where several attacks were connected, but where the various actions did not constitute a single incident ('Yes' denotes that the particular attack was a part of a 'multiple' attack)
- *crti3* – a variable that indicates if the attack is outside the context of legitimate warfare activities, insofar as it targets non-combatants (i,e., the act outside international humanitarian law)

A typical record for a terrorist incident in India is shown in Table 2. Table 3 shows the total number of records in GTD.

TABLE 2. A TYPICAL RECORD IN GTD

iyear	2015
latitude	32.91442
longitude	75.14174 (near Jammu)
attacktype	5 (Hostage taking)
targtype1	4 (Military)
targsubtype	29 (Military Unit/Patrol/Convoy)
weapontype	6 (Explosives/Bombs/ Dynamite)
nationality	92 (India)
int_any	1 (International incident)
groupname	Lashkar-e-Taiba (LeT)

TABLE 3. NUMBER OF RECORDS IN GTD

Number of records	156,772
Number of Indian records	6,411

Our focus in this work is on incidents in India [3]. GTD has data available from 1970 to 2015. We used the data up to 2014 for training the machine learning models and tested the models on the data for 2015 as shown in Table 4.

TABLE 4. TRAINING AND TEST DATA SETS

Set	Number of Records
Training Set (1990-2014)	5,131
Test Set (2015)	578

B. *Visualization of Dataset*

As a preliminary step, the frequency of terrorist attacks in India from 1970 to 2015 was plotted on a map of India to visualize the regions most prone to terror attacks [4]. Figure 1 shows the results.

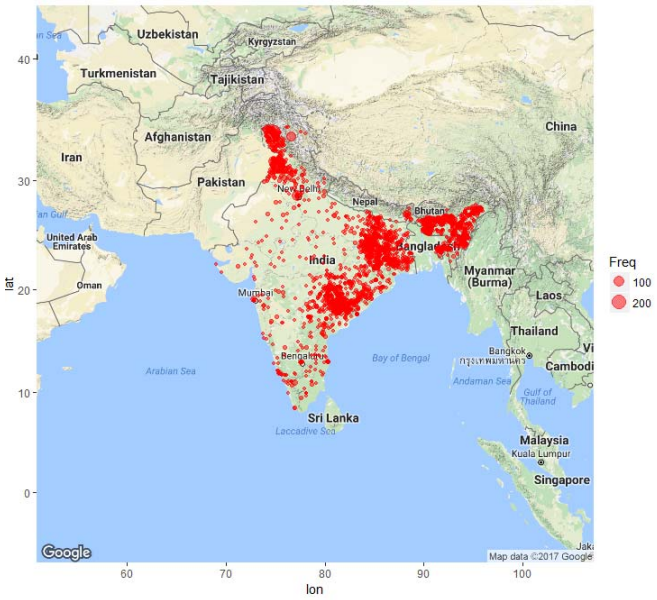

Fig. 1. Regional Distribution of Terror Attacks in India

IV. Literature Survey

Terrorist Group Prediction Model (TGPM) [5] builds a model that learns similarities of terrorist attacks from various terrorist incidents to predict the group responsible for recent attacks using concepts such as Crime Prediction Model, Group Detection Model (GDM) and Offender Group Detection Model (OGDM). The percentage of attacks by each group is calculated. Additionally, each input parameter such as attack type, location, hostage/kidnapping, suicide attack, weapon

type, target type, etc., are assigned weights based on their impact on the incident. The group weight is calculated using the percentage of attacks of each group and the parameter weights. Using these, a number of clusters are created. Association of the input data with the formed clusters is calculated and the highest value is chosen. The group name corresponding to the cluster with the highest association value is predicted to be the most probable terrorist group responsible for the attack. The system gives an accuracy as high as 80.41% but has its own drawbacks. It fails to detect attacks by smaller groups when its pattern of attack matches that of a larger group. The system also fails in cases of a terrorist group using different attack patterns each time.

Another system based on classification algorithms [6] predicts the terrorist group responsible for an attack in the Egyptian context, using data of attacks between 1970 and 2013 from GTD. Two techniques are used to handle missing values in the dataset, namely Litwise-Deletion (data removal) and Mode-Imputation which result in two distinct datasets. Classification algorithms are applied to each of the datasets obtained and the accuracies obtained are used to compare these techniques. The classification algorithms used include Naïve Bayes classifier, K-Nearest Neighbour, Tree Induction (C4.5), Iterative Dichotomiser (ID3), and Support Vector Machine. Experiments are performed with the help of Waikato Environment for Knowledge Analysis (Weka) and the final evaluation is based on four performance measures, namely classification accuracy, precision, recall and F-measure. The training-test data split is obtained using two different techniques - random sampling with 66% as training data and the remaining as test data and 10 fold cross-validation. K-Nearest Neighbour was more accurate compared to other classifiers when litwise deletion approach is used, whereas, with mode imputation Support Vector Machines clearly outperforms the other classifiers. Hence, in our work, we used these classifiers for the Indian context.

V. Machine Learning Models

In this project, we built several machine learning and analytics models on the Indian subset of the Global Terrorism Database, including K-means clustering, Boruta Analysis, and C4.5. We also tried to make predictions based on several individual attributes like target type and weapon type. We outline below the most successful approach in terms of the accuracy of prediction.

The first step was to clean the dataset. Structured data about the terrorist incidents in India were extracted from the dataset (which is in gdb format) including the latitude and longitude of the attack location. On this data, we tried to partition the dataset using *k-modes clustering* as the dataset was categorical [7]. Though clustering did not help us much with the prediction of the terrorist group behind an attack, it gave us an idea of the most promising attributes in the dataset. '*attacktype1*', '*targettype1*' and '*targetsubtype1*' were determined to be some of the promising attributes.

We performed Factor Analysis of Mixed Data (FAMD) on our dataset to extract the most contributing features in the dataset. From the results obtained, up to 83 dimensions were preserved which contributed to a variance of 90.001. The attributes that contributed to the construction of these 83 dimensions were extracted and preserved as the most prominent input features of the dataset. Hence, this helped us infer that *iyear, attacktype1, targtype1, targsubtype1, weaptype1, region_id, natlty1, property, INT_ANY, multiple* and *crit3* were the most promising attributes of the dataset that can be used to predict the terrorist group behind an activity.

A. Decision Tree Algorithms

As evident from the dataset, there exists a non-linear and complex relationship between the independent parameters and the dependent variable, namely the perpetrator. Hence, we decided to use tree-based models to model the data and predict the perpetrator of the attack. In addition, decision tree focuses on the relationship between various events and hence, it replicates the natural course of events, and as such, remains robust. Therefore, we used decision trees to model the given dataset. Additionally, alternative algorithms were used to build decision trees. A decision tree was built using the C4.5 algorithm.

We trained a decision tree classifier [8] on our data to predict the perpetrator. In these algorithms, data is partitioned into sub-groups recursively. An attribute is selected and a logical test is formulated on this attribute. Later, based on the outcome of the test, the tree is branched, the subset of the data satisfying that outcome is moved to the corresponding child node and the tests are run recursively on the child node. We used a decision tree algorithm called C4.5 to classify data and used it for prediction.

C4.5 is a statistical classifier based on ID3 algorithm that works on the concept of information entropy. The training data is seen as a set of classified samples having p-dimensional vectors defining the attributes of the sample. It generates a decision tree where each node splits the classes based on the gain of information. The attribute with the highest normalized information gain is used as the splitting criterion. This approach gave us a prediction accuracy of 65% which is good but not good enough for predicting the perpetrator of a terrorist attack.

B. Random Forest Algorithm

The Random Forests algorithm is one of the best classification algorithms that is capable of classifying large amounts of data accurately. Random Forests are a

collaborative learning method for classification and regression. The technique constructs a number of decision trees during training and outputs the class that is the mode of the classes output by individual trees. The basic principle is that a group of "weak learners" can come together to form a "strong learner". Random Forests gave us a prediction accuracy of 58.8% which was much lesser than accuracy gained by C4.5 algorithm.

The results of these two experiments made it clear that decision tree algorithms are not suitable for the type of categorical data with which we were dealing.

C. Support Vector Machine

SVM performs non-linear classification by mapping the inputs into high-dimensional feature spaces. Since our dataset has independent features and a fairly large training sample, the linear kernel is used to prevent overfitting. Given labelled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples [9]. With an SVM Model built on the dataset to predict the terrorist group behind an attack, the following accuracies where obtained:

- On using the dataset of terror attacks post 1990, an accuracy of 73.24% was obtained.
- On using dataset of terror attacks from 1996 and above, the mean accuracy of prediction obtained was 75.6%.
- On removing certain obscure terrorist group names manually and testing for data up to 2014 only, we were able to obtain accuracies of 78.3%.

Support Vector Machines work exceedingly well on classification problems, especially when the dataset is noiseless and balanced. In addition, SVM avoids overfitting of data. This becomes primarily important for a dataset like the GTD, as overfitting of data in these kind of datasets would result in inaccurate results. For this particular dataset, linear kernel (default option) gave a higher accuracy over other non-linear kernels as the number of features is large and hence, mapping the data to a higher dimensional space would not have any substantial effect.

Table 5 summarizes the accuracy of the three main methods we tried.

TABLE 5. ACCURACY OF PREDICTION

	Algorithm	Accuracy
1	C4.5	60.0%
2	Random Forest	58.5%
3	SVM	73.2%

VI. Discussion

SVM gave us the highest accuracy of nearly 80% on the older version of the GTD data set when we tested it on incidents prior to the end of 2014. The new release of GTD includes about 265 distinct terrorist groups. We observe that in our tests, most of the major terrorist attacks have their perpetrator identified correctly. Certain obscure groups which are highly localized in their attacks (in terms of the regions of India) are not being predicted correctly. Further, a number of new groups have appeared since 2012 and hence, there is a reduction in the accuracy from the previously obtained 80%.

VII. Conclusion

Terrorism continues to be a menace across the globe. Data analytics and machine learning offer a promising approach to aid the investigators in quickly determining the most likely perpetrator of a terrorist attack. We believe that this project has demonstrated how a method like Support Vector Machine can predict the perpetrator correctly four out of five times, thereby enabling the investigating agencies to narrow down the possibilities and act quickly to catch the real perpetrators. We further intend to try other methods such as ensemble classifiers and deep learning to further improve the accuracy of prediction.

REFERENCES

[1] E. Picardo. (2016). **Don't Hide From The Reality Of How Terrorism Affects The Economy**. [Online] Investopedia. Available at: http://www.investopedia.com/articles/investing/030215/how-terrorism-affects-markets-and-economy.asp. [Accessed: 12 Aug 2017].

[2] National Consortium for the Study of Terrorism and Responses to Terrorism (START). (2016). **Global Terrorism Database** [Data file]. Retrieved from https://www.start.umd.edu/gtd

[3] **List of terrorist incidents in India**. (2017, March 30). *In Wikipedia, The Free Encyclopedia*. Retrieved 12:30, May 19, 2017, from https://en.wikipedia.org/w/index.php?title=List_of_terrorist_incidents_in_India&oldid=772979144

[4] Lavanya Hegde, Narella Srilakshmi and Kavi Mahesh. "Visual Analytics of Terrorism Data", In Proc. *IEEE CCEM 2016 International Conference on Cloud Computing in Emerging Markets*, October 19-21, Bangalore, India.

[5] A. Sachan and D. Roy. "TGPM: Terrorist Group Prediction Model for Counter Terrorism", *International Journal of Computer Applications*, vol. 44, no. 10, pp. 49-52, 2012.

[6] G. Tolan and O. Soliman. "An Experimental Study of Classification Algorithms for Terrorism Prediction", *International Journal of Knowledge Engineering-IACSIT*, vol. 1, no. 2, pp. 107-112, 2015.

[7] Kaushik, S. (2017). **An Introduction to Clustering & different methods of clustering**. [Online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/ [Accessed 19 May 2017].

[8] Analytics Vidhya Content Team. (2017). **A Complete Tutorial on Tree Based Modeling from Scratch (in R & Python)**. [Online] Available at: https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/ [Accessed 19 May 2017].

[9] Ray, S. (2017). **Understanding Support Vector Machine algorithm from examples**. [Online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2015/10/understaing-support-vector-machine-example-code/ [Accessed 19 May 2017].