Global Terrorrism Analysis

Sudarshan R , Pranav Yogi Lodha , Gaurav Pandey

Department of Computer Science and Engineering

PES University, Bangalore, 560085 India

sudarshanmurthy333@gmail.com

Abstract

—Terrorist attacks are on the rise in the last few decades. Global terrrorism,nightmare for many people and for any country. To prosper in every sector this has to be reduced as much as possible and our goal is towards this but first we need to analyse the available data and gain some insights on the same. In this we present the work done in the past and their assumptions and methods/techniques used in the past and later we say how all this can be improved and how our approach is different from those and what makes us stand out of the crowd .

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 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.The terrorism is a boon to mankind and it will always be that way and this is something we dont have control upon and as an effect of this many people are dead and few have lost there home etc...But this has to stopped at any cost and its our duty to do this as a responsible citizen

Now,we have a dataset which has informations regarding the previous attacks and does not meet the mark on data quality because the data is not complete,accurate,consistent and has outliers.So any analysis on this will lead to insignificant outcome and our aim is to get a result which is significant and accurate as possible and as first step to this we can increase the data quality which we will discuss about it later .

The given dataset contains around 200 attributes and many too many rows to keep a count so that's a good start .The Terrorism dataset offers us many too problems to solve for example ,where the terrorism activity has been in peek over the past few decades and why is it only in that place it has occurred many a times and whats makes it so easy to plan there and how many casualties ,people losing there home and what not?.

And again the number of problems has no bound and we try to maximize to solve this as much as possible .

Why do we need to solve these problems?Why is it important?

This is often asked to any analyst.Why is it important?This could be explained by an example.

Consider a scenario where a place is been bombed again and again over the given years,we as an analyst try to gain information about this using various techniques.For example ,if a place is been bombed we try analysis what thing make them to bomb there and how many people will be affected by this and it there any seasonality for this activity and if there is any seasonality ,is there a fixed period ?And by gathering all this info we can fit a model and try to analyse various components and make a valid prediction and forecast our model if it is accurate enough .

Given all these things out next step is to approach to solve these problem and then implement in a proper way for the welfare of mankind

II. Problem statement

The objective of our work is as follows .

1)Why countries/groups attack and where are their attacks based?To find factors like National income,religion, happiness index ,population,mode of governance,level of development etc that influence these attacks.Through this we'll try to figure out the mentality and the motives of the various terrorist groups in various countries.

2)A case study of the various attacks taking place in India. Which groups are responsible for these and why certain areas are targeted? To predict future attacks.

3)To do a case study on Iraq being the country causing maximum attacks and Taliban group being the group causing maximum of these attacks. Finding out various factors which influence these attacks by these groups.

4)Final objective will be to be able to predict which country is most likely to be attacked by which group/country.This kind of analysis will help different defence agencies to be well prepared and equipped for any future attacks.

III. Background

A. About the dataset

The Global Terrorism Database (GTD) is an open-source database including information on terrorist attacks around the world from 1970 through 2017. The GTD includes systematic data on domestic as well as international terrorist incidents that have occurred during this time period and now includes more than 180,000 attacks. The database is maintained by researchers at the National Consortium for the Study of Terrorism and Responses to Terrorism (START), headquartered at the University of Maryland.
For each event a wide range of information is available,including the date,location of the incident, weapons used,nature of the target,casualties .

But the disadvantage of this dataset is ,it is not complete,accurate,and has noisy data.
As our first step we cleaned this dataset and made it more interpretable and believable .

IV. Literature Review

Data Visualization as a Preprocessing Step in Designing of Data Mining Tools Visualizing Time Series Pattern of Rainfall Data(2018)[2].This is about the Visualization of a time series data and how important step it is in analysing the data .Visualization is necessary to perform analysis and visualize patterns, cycles, trends in the dataset.

The ARIMA Model is applied to visually analyze patterns in the dataset and a suitable model is developed for further use in forecasting. As we know that the efficient and effective approach for time series visualization and analysis is the Box-Jenkins “Auto Regressive Integrated Moving Average (ARIMA) model”. And also it explains how the ARIMA model is a better fit for the dataset and plotting ACF and PACF to find optimal AR and MA models.

Machine Learning Approaches to Uncover Terrorism Network in India (2020)[3]. Several attempts were made to exterminate terrorism in the past but there is a need for some alternative promising techniques to eradicate the immense loss caused due to terrorism. Therefore, authors of this paper have devised various approaches by applying different machine learning algorithms and tried to visualize the result through visual analytics that recognize hidden patterns and trends in terrorist events and networks.They start with feature selection and then explains visualization ,correlation analysis and then discusses ML algorithms like K-means clustering and validates the model based on a few criteria .

Prediction of Satellite Time Series Data Based on Long Short Term Memory Autoregressive Integrated Moving Average Model (LSTM-ARIMA) (2019)[4].In this the author explains time series data analysis as a method of predicting future values by observing historical data by using the traditional Autoregressive Integrated Moving Average model (ARIMA) .And he uses LSTM-ARIMA algorithm to predict the time series data of a meteorological satellite telemetry parameter and analyze the error of the prediction data as it has high accuracy and strong reliability prediction than ARIMA alone.And then he checks the correctness of the model by validation test.

Terrorism Analytics: Learning to Predict the Perpetrator(2017)[5].Data about terrorist attacks in India is analysed. Several machine learning algorithms are explained and 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.

They use Support Vector Machine technique in predicting the perpetrators.They first explain the visualization of the dataset (How significant it is)and then go about explaining how ML model is used in predicting and finding its accuracy.

Visual Analytics of Terrorism Data (2016)[6].They used Social Network Analysis(SNA) which offers promising techniques among which visual analytics combines the best of statistical analysis with human ability to visually recognize hidden patterns and trends in terrorist events and networks.The entire paper gives the roadmap for our project starting from collection of data to analysis.

An Outlier Detection Algorithm Based on Cross-Correlation Analysis for Time Series Dataset(2018)[7].This paper emphasizes how important outlier detection is when a time series data is used for an analysis. They propose an Outlier Detection method based on Cross-correlation Analysis (ODCA). ODCA consists of three key parts. They are data preprocessing, outlier analysis, and outlier rank. First, they investigate a linear interpolation method to convert assembled outliers into isolated ones.Second, a detection mechanism based on the cross-correlation analysis is proposed for translating the high-dimensional data sets into 1-D cross-correlation function, according to which the isolated outlier is determined.

Finally, a multilevel Otsu’s method is adopted to help select the rank thresholds adaptively and output the abnormal samples at different levels.

V. METHOD

A. Experimental dataset

we are working with the Global Terrorism Database (GTD) which is an open-source database including information on terrorist attacks around the world from 1970 through 2017.

A few of the important attributes included in the study are:

- Attack type
- city location
- Latitude,Longitude
- Success rate
- Date and Time of incident

B. Data Pre-processing

Pre-processing the data is an important stage in any model building and is required for many reasons.
Before the model is built, it is also essential to ensure the quality of the data for issues such as reliability, completeness, usefulness, accuracy, missing data, and outliers. Reliability of data is a major issue when the data is collected .

1)Data Cleaning

i)Missing data is another frequently observed problem and we have come up with strategies for handling missing data.

- Use central tendency like mean/median
- Use regression

ii)Smoothing the noisy data(error in a measured variable)is another important task in data pre-processing. Techniques such as binning and weighted difference can be applied to this.

iii)Outliers-For any result to be significant we need to remove influential observations/outliers .This can be identified using methods like clustering and box plot analysis of a variable.

2)Data integration is not a problem for our study as we are using a standard available dataset.

3)Descriptive Analytics

It is always a good practice to perform descriptive analytics before moving to predictive analytics model building. Descriptive statistics will help us to understand the variability in the model and visualization of the data through, say, a box plot which will show if there are outliers in the data. Another visualization technique, the scatter plot, may also reveal if there is any obvious relationship between the two variables under consideration.

Correlation analysis is very important as it shows any associative relationships between the attributes.And the correlation type applied depends on the attribute type.For example,for categorical we apply chi-sq test . And when the datatype is interval/ratio we apply Pearson's correlation and so on.And high correlation coefficient does not mean they are highly correlated ,it could even be spurious!.

Skewness test using Kurtosis and Pearson's moment coefficient is also done as we get to understand how the data is distributed . And if skewed how is it advantageous to us?

4)Data Transformation

In this preprocessing step, the data are transformed or consolidated so that the resulting mining process may be more efficient, and the patterns found may be easier to understand.This is just to increase the data quality and to reduce the effect of the measured unit of the attribute.

C. Model Building

After all the visualisations and preliminary analysis is done, we will be building various models like

- Simple/Multi linear regression model
- AR,MA
- ARIMA

These models can be applied to the dataset and get answers for the proposed problem statement.For example, by using simple linear regression we can predict the future value of an dependent variable by using independent variables and we can use the OLS method to get the regression parameters required .

After the model is developed we will do validation tests like R^2 ,t-statistic to find the significance of the explanatory variable on the outcome variable,F-statistic to get the overall significance of the model.

Time series analysis on the dataset using AR ,MA is also a kind of regression and where in AR the future value is predicted/forecasted using regression on the same . And the ACF/PACF is used to fix the number of lags .

D. Validating the model

A major concern in analytics is over-fitting, that is, the model may perform very well in the training data set but may perform badly in the validation data set. It is important to ensure that the model performance is consistent in the validation data set as was in the training data set. In fact, the model may be cross validated using multiple training and test data sets.

Methods used:

- R^2
- T-statistic
- F-statistic
- Residual analysis

In case of time series analysis like AR ,ARIMA

- Check for stationarity
- White noise

These validations are very important for any model built and has to be performed before deploying the model.

VI. SUMMARY

For any analysis to produce significant results ,one has to follow all the steps involved and in this report we have described the literature review and the outline of our task .It provide a rough view about our hold on the dataset and the problem we are trying to solve .

VII. REFERENCES

[1] E. Picardo. (2016). Don't Hide From The Reality Of How Terrorism Affects The Economy. [Online] Investopedia.

[2] Manoj S. Chaudhari , Nitin K.Choudhari.Data Visualization as a Preprocessing Step in Designing of Data Mining Tools Visualizing Time Series Pattern of Rainfall Data.2018 International Conference on System Modeling & Advancement in Research Trends (SMART)

[3] Anchal Hora, Asmita Bari , Sonal Rawat.Machine Learning Approaches to Uncover Terrorism Network in India..2020 International Conference for Emerging Technology (INCET)

[4] Yuwei Chen ; Kaizhi Wang.Prediction of Satellite Time Series Data Based on Long Short Term Memory Autoregressive Integrated Moving Average Model (LSTM-ARIMA) .2019 IEEE 4th International Conference on Signal and Image Processing (ICSIP)

[5] Disha Talreja , Jeevan Nagaraj , N J Varsha , Kavi Mahesh. Terrorism Analytics: Learning to Predict the Perpetrator.2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)

[6] Lavanya Venkatagiri Hegde ,Nerella Sreelakshmi, Kavi Mahesh.Visual Analytics of Terrorism Data.2016 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM)

[7] Hui Lu , Yaxian Liu ,Zongming Fei , Chongchong Guan.An Outlier Detection Algorithm Based on Cross-Correlation Analysis for Time Series Dataset