Data Aggregation and Terror Group Prediction using Machine Learning Algorithms

S P Maniraj, Deep Chaudhary, Vankayala Hari Deep, Vishesh Pratap Singh

Abstract : This paper is about to introduce a proposed system that examines growth or decay of the terrorist groups by the time, active locations, types of attack they carry out, motive targets, Weapon mastery and availability and many parameters to analyze the patterns and hidden structures in their activity and to predict the occasion and type of their future attack. We have done a detailed analysis of data we get from different sources and we also performed different classification algorithms on the available data to find the chances of probable attack on different regions.Based on results finding which of the algorithms works with highest accuracy.

Key Words: Analysis, Classification, prediction, terrorism, Data aggregation, Terror Group.

I. INTRODUCTION

Over the past many years the world has been a witness for the remarkable number of terrorist events because of the terrorist event, the main victim is people the terrorist event which is happening in the world is not in the random order each terrorist event is interlinked with other in the world. We used to follow a particular pattern that initiates the terrorist activity; our task is to predict the event before they initiate it. To obtain the real-time data of the past terrorist events in the world and implement the clustering and data aggregation with the algorithms like logistic regression, SVM, K-NN we analyze clusters related to four combinations terrorism - event terrorism – target , terrorism target-terrorism agencies , terrorism agencies-terrorism attack type ,terrorism attack type-terrorism method, terrorism method-victim location. Through dataset, intuition is to annually analyzing the number of content of the cluster, effectively from 1980 to till date.

Based on the clustering and data classification mechanism work is to predict the terrorist group responsible based time duration, event place, agencies, target, and attack type from the input and predicting terrorist group as output..According to the Global Terrorism Data report there is certain criterion to be chalked out to decide the objective and find the solution

II. RELATED WORKS

1.The paper published by Dr.Tariq Mahmood and Mr.Khadija Rohail has suggested worked on applying the use of data minining methods on terror activities in Pakistan. The study was focused on the cluster analysis which is used to group all raw data together in form of clusters as they possess common property, so it was found out to be ClOPE as a proficient algorithm to analyze clusters on the raw data to analyze the terror groups or active organizations of any area as their study was specifically based on 4 provinces of Pakistan it provides the clustered data in form of target-type, target party, locations and several other factors that distinguishes them.

2. The paper on Effectiveness of Terrorism policies by Chaomin Lou and yang li suggests that their work is used to make the assumptions based on theory evaluation and then it test the effectiveness of US counter terrorism policies by implementing a time series interrupted model. This model is used to improve as it categorizes the policies based on their proactiveness it simplifies a time series equation that give results on basis of some mannequin parameters that are used to rate the policies in three categories as proactive, policy that works on the root of terrorism and also insignificant policies it worked well as time series functionality improves with the increase in the number of training data it was a commendable works as it helps to determine active policies and their model was tested on insurance policies too which gave them some excellent results.

3. Most related work was to form a predictive modeling model that gives warnings about the future assaults on Pakistan by two scholars Hina Muhammad Ismail and Abdullah Kazi their journal suggests the use of classification modeling. The study was to analyze the previous incident records which were available in GTD database. This study was focused specifically to Pakistan it takes certain common parameters which helps in identifying the threats to type of targets and locations that helps security agencies to work and make decision based on collective feed by the previous records to aid them for keeping security measures.

III. PROCESSING

A.ARCHITECTURE MODEL

Based upon the proposed system it is essential to depict the significance of each step for elaborating the method or techniques used for drawing out inference based on the given dataset Model is,

Manuscript published on November 30, 2019.
* Correspondence Author

S P Maniraj, Assistant Professor (Senior Grade), Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai (Tamil Nadu), India.

Deep Chaudhary, Student B.Tech, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai (Tamil Nadu), India.

Vankayala Hari Deep, Student B.Tech, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai (Tamil Nadu), India.

Vishesh Pratap Singh, Student B.Tech, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai (Tamil Nadu), India

collecting data is used to capture records of past events to analyze data and find the recurring patterns to draw out inference that which data is useful and which data is of no use. Data is collected from GTD (Global Terrorism Database).

B.DATA COLLECTION

In this phase it is necessary that the data we collect is gathered from relevant sources. The quantity and quality of data tells how accurate the On terrorism which is gathered from relevant sources. Architecture model represents the structure which is made to explain the aspects of work or architectural design.

C.DATA PRE-PROCESSING

It is the most important phase in the building of model because we check the data before analysis & prediction of the data because the collected data is present in the raw format which needs to be cleaned and make it ready for further processes.

D.DATA CLEANING AND PREPARATION MISSINGVALUES

If data contains a lot of missing values then it means that it's incomplete to handle such problem it has two ways either to delete the entire row or column of data or to fill the place of missing data with some value. Generally if there is not so much missing values it is preferred not to delete the entire row or column because it might contain useful information that could help in getting insights of data that's why data is being replaced with mean in place of missing data and some columns are dropped which has majority of NAN values.

E.DATA TRANSFORMATION

Data Transformation is method in which the data is being transformed according to needs of user such that the data mining procedures has to be made more accurate, efficient and easy to understand. The methods generally used are standardization, normalization, construction and concept hierarchy generation.In this construction is used as the GTD dataset has some features which are mixed up like there is single column for (city, village, town, target_sites) thus the values need to be split so attributes are split and new features are made in separate columns.

F.DATA ENCODING

Some of GTD's data is present in the categorical format. In case there is a nominal and ordinal value the values are treated as dummy variables these variables can't be used to predict the results thus we use label encoding and one hot encoding in order to give discrete value to the dummy variable which is used to make the data useful to apply prediction methods.In GTD's data Encoding is used in case of city attribute as it has values with repentance which is used in case of analyzing values for a particular continent or country's data then it is used to categorize them into separate attributes and dummy variable trap is avoided because one dummy variable is used to predict Other dummy variable so to avoid this trap one dummy variable should not be selected.

G.CROSS VALIDATION

It is used to split data into training and test sets where training set is used to train data on hypothesis and test set is used for checking the outcome on other data to check the accuracy of the hypothesis.In the dataset K-fold cross validation is applied because we need to test data unbiased of the values present in training and test sets that's why it is applied as it ensures each fold is used as testing set at some point. After removing all redundant features and data preprocessing data is being divided into 16 attributes to analyze and perform predictions.

Original Data set	Filtered Dataset
136 variables	16 variables
Terrorist Incidents 345228	Terrorist Incident 31560

H.PLOTTING AND CHARTING

Data is analyzed by plotting and charting the data it helped to draw insights from data by plotting histograms and some bar charts to analyze data.

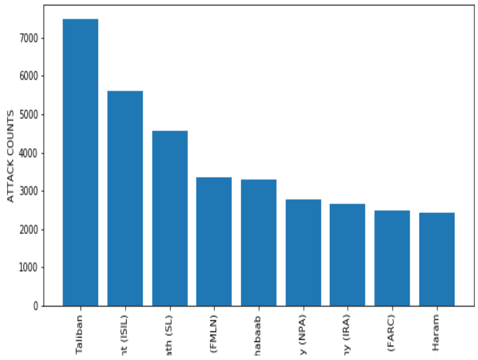
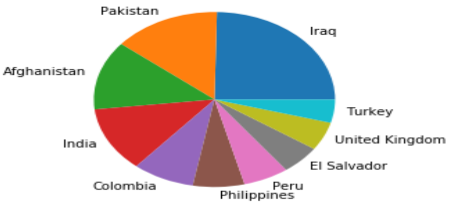


CHART-1:

Above bar chart denotes Attack counts versus Terror groups it describes the ten most effective terror groups.



BAR GRAPH-1:

The given pie chart describes countries which are affected most by terrorist attacks.

I.PREDICTION

Classification is being used for prediction of the values. Classification is used when the result is of categorical type in classification prediction output variable taken is Suspect it is used to find probably which terrorist group could be responsible for the attack. For prediction 3 algorithms are used they were Logistic Regression, SVM and KNN.For all these a comparison is done to estimate which algorithm is more accurate, there evaluation measures are taken as Accuracy, Precision, Recall, F1 scores and there ROC curves are drawn to find further insight and score which algorithm is better.

IV.EVALUATION MEASURES

The evaluation used in classification problems are used to find which algorithm is perfect to approach it has certain parameters that determines it.

A.ACCURACY

It tests the accuracy of classifier by making a confusion matrix and through that it takes the ratio of sum of true positives and true negatives to the sum of true positive, true negative, false positive and false negative but it's not the only parameter that determines the accurateness of classifier.

B.PRECISION

It gives how much precise is the results it gives out of the predicted values it is calculated by taking ratio of true positive by sum of true and false positive values.

C.RECALL

It gives how much positive value is dig out by classifier it is opposite to precision it is calculated by taking ratio of total positive by sum of total positive and total negative.

To seek balance between Precision and Recall use of **F1 score** is required as it gives of average of both of them and there is large number of actual negatives.

D.ROC /AUC

It is used to depict the diagnostic ability of the class system.It is plotted graph against True positive rate versus false positive rates, based on curve and threshold values ROC curved is rated.

V.DATA OUTPUT

Classifiers are tested based on evaluation parameters it is found that the classifier that runs best in case of finding which algorithm is perfect is SVM classifier with best accuracy followed by KNN and logistic regression for each classifier its accuracy, precision, recall, f1 scores are analyzed and the resultant Classifier best suitable for the model comes out to be SVM.

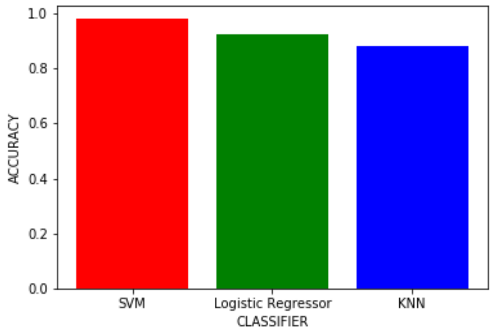


CHART-2:

The above bar chart describes which classifier works best on the model.

VI. CONCLUSION

This paper contains deep analysis of data on terrorism records and it also mentions prediction in which classification is done to find suspect which are responsible for attack or which has more chance to attack it compares the three classification algorithms and checks for the most efficient algorithms it tests algorithms like KNN, logistic regression and SVM in which SVM is found out to be most efficient on the given dataset. Thus its comparison helps to predict the terrorist groups which could be responsible for the attack. This proposed system might help government organizations to fight against terrorism.

REFERENCES

1. Dr.Tariq Mahmood and Mr.Khadija Rohail Analyzing Terrorist Incidents to Support Counterterrorism – Events and Methods
2. AGARWAL, J., NAGPAL, R., & SEHGAL, R. (2013). Crime Analysis using K-Means Clustering. International Journal of Computer Applications
3. Wang X , Miller E , Smarick K , et al. Investigative Visual Analysis of Global Terrorism 2008.
4. NIEVES, S., & CRUZ, A. (2011) Finding Patterns of Terrorist Groups in Iraq.
5. ZHAO, Y. (2012) R and Data Mining: Examples and Case Studies, Academic Press, Elsevier, Australia, 164 pp.
6. Hina Muhammad Ismail and Hameedullah Kazi suggests using classification for predictive analysis.

AUTHORS PROFILE

S P MANIRAJ- Assistant Professor (Senior Grade),Completed B.E.(Computer Science and Engineering) in Anand Institute of Higher Technology in 2004-2008,Completed M.Tech (Embedded System Technologies) in Anna University Tirunelveli in 2008-2010,Current pursuing PhD in school of computing in the domain of Medical Image Processing,Having working experience of 9 years in Engineering and technology.

Deep Chaudhary, Currently pursing B.tech degree in Computer science in SRM College,certified in python programming,R Language,Html and Css,develop an new type of business model using Machine Learning Algorithms,Ux designing.

Vankayala Hari Deep, Currently pursing B.tech degree in Computer science in SRM College,certified in R Language,Digital Marketing,Search Engine Optimization ,Designing website in Wordpress, Designing Database in sql, Ux & Ui Designing.

Vishesh Pratap Singh, Currently pursing B.tech degree in Computer science in SRM College,certified in python programming,R programming,Html and css,web Devolpment,Ui designing.currently aiming for becoming Data scientist.