International Conference on Inter Disciplinary Research in Engineering and Technology [ICIDRET]

ISBN	978-81-929742-5-5	**Vol**	I
Website	www.icidret.in	**eMail**	icidret@asdf.res.in
Received	14 - February - 2015	**Accepted**	25 - March - 2015
Article ID	ICIDRET042	**eAID**	ICIDRET.2015.042

Big Data On Terrorist Attacks: An Analysis Using The Ensemble Classifier Approach

R. Sivaraman[1], Dr. S. Srinivasan[2], Dr. R M. Chandrasekeran[3]
[1]Assistant Professor, Department of Computer Science and Engineering,
Anna University, Regional Office, Madurai, Tamil Nadu, India.
[2]Professor, Department of Computer Science and Engineering,
Anna University, Regional Office, Madurai, Tamil Nadu, India.
[3]Professor, Department of Computer Science and Engineering,
Annamalai University, Chidambaram, Tamil Nadu, India.

ABSTRACT - *Terrorism has virtually invaded our day to day lives. We can't imagine of passing a day without a terrorist attack in any part of the country that has brought in irreparable loss to mankind and also invaluable material destruction. The knowledge and information we collect about the terrorists' operations are highly voluminous and is increasingly becoming multidimensional, thereby pushing the analysis of Big Data into new frontiers. This data when combined with counter-intelligence inputs brings in a new perspective on the efforts to combat terrorism. As new terror outfits spring up consistently, applying suitable data mining techniques on such Big Data has a great impact on the counter terrorism measures and understanding the pattern of attacks. In this research we have analyzed the performance of classifiers like decision tree and ensemble classifier on the Global Terrorism Database and the results have shown that the ensemble method outperforms for the given dataset.*

Keywords: Big Data, Global Terrorism Database, Decision Tree, Ensemble, Weka

I INTRODUCTION

Today we live in this era of Big Data where the amount of information gathered and stored on data storage systems are several trillion times more than the population of the World. In fact every one of us have several zetta bytes of data about our own individual information from birth to death that includes telephone call records, emails, messages conveyed on various social messaging platforms, CCTV footages, financial transactions etc., In this age of Information Technology, we leave a footstep of data wherever we move and it is inseparable like our own shadows. This Big Data when used effectively and efficiently holds the key for several unanswerable questions, pattern recognitions and predictions.

With the effect of Terrorism and its frequent threats on us we have been forced to live in an environment which resembles our age old days of forest life, fearing and fighting with wild animals for our mere existence. History repeats again only with new flavors and colors and nothing has changed considerably. But today we are equipped with modern weaponry to fight back than that of the stone tools we used long back. Forget the weapons of steel. We have something that is much stronger than that of all, which is Big Data. This structured data when designed to form conceptual frameworks can reveal us several hidden phenomena and can guide with intuitionistic relations. One such dataset is the Global Terrorism Database [1] referred as GTD, which is an open source collection of terrorism events across the globe from 1970s to 2013. This unclassified database is a comprehensive collection of over 125,000 terrorist attacks with detailed records mentioning the country, type of attacks, targets – civilian, military, business, weapon type etc.,.We have utilized the Global Terrorism Database for this research and have focused on the extraction of information using decision tree and ensemble classifier. The experimental results show that the ensemble classifier can identify the incident types with better accuracy than that of the decision tree classifier.

Cite this article as: Sivaraman R, Srinivasan S, Chandrasekeran, R M. "Big Data On Terrorist Attacks: An Analysis Using The Ensemble Classifier Approach." *International Conference on Inter Disciplinary Research in Engineering and Technology* (2015): 255-261. Print.

II RELATED WORK

Nitin et al. [2] have used the J48 Decision tree algorithm in the classification of criminal records and predicting a crime suspect and for overall analysis of crime data. The data pertaining to various types of crime like traffic violations, theft, fraud, drug offenses etc., were collected by field work. The results of the J48 algorithm are then verified with the correctly classified instances, FP rate, TP rate, confusion matrix, recall, MCC and F_Measure. The classification method would then suggest about the suspect is innocent or not. An email dataset was used by Sarwat Nizamani [3] for detecting email with suspicious content. Their research has focused on the evaluation of machine learning algorithms such as decision tree (ID3), logistic regression, Naïve Bayes (NB) and Support Vector Machine (SVM). The findings prove that the decision tree algorithm (ID3) did well when compared with that of the other classifiers. Upon application of suitable feature selection strategy, an increase in performance was witnessed by the logistic regression algorithm along with the decision tree algorithm. Terror incidents in India was analysed by Borooah et al. [4] used the GTD database to analyse the fatality rates during 1998-2004. Their research separates the influence on the number of attack type and attack group and used the Atkinson's concept of equality-adjusted income to terrorism to arrive at the concept of equality-adjusted deaths from terrorist incidents.

The impact of terrorism on investor's sentiment related to Hospitality stock was studied by Chang et al. [5], and their research has proved that the there was a fall of 10 to 15 percent every year due to terrorist attacks. However once the threat of terror has been withdrawn or subdued the markets recovered after the initial negative reaction and yield better returns up to four times more than the average event and this research used the GTD database to arrive at a logical conclusion. A method to segregate the GTD database by transnational and domestic incidents was devised by Enders et al. [6]. They analysed the impact of transnational terrorism and found that it had a greater negative impact on the economic growth of a country than that of domestic terrorism. The results have shown that cross correlation exists between the domestic and the transnational terrorist events. The findings also suggested that the domestic terrorism can expand to transnational terrorism and hence the target countries cannot turn a blind eye to domestic terrorism in neighboring countries and may have to put an end to the homegrown terrorism.

Young et al. in their research work Veto Players and Terror [7] used the Tsbelis's veto player's theory to analyse why certain democratic countries foster terrorism and a majority of other countries are curbing it effectively. When the terror outfits wanted for a shift in the government policies, then more number of veto players will lead into a deadlock which will tend to generate more number of terror events. The results discussed that with the inability of the societal actors to change the policies of the government through non-violent and institutional participation, the homegrown terrorism cannot be tackled. Nizamani et al. [8] have extensively analyzed the news summaries from the global terrorism dataset using machine learning techniques. They have adopted different learning algorithms including Naive Bayes, decision tree and support vector machine. The findings suggest that the decision tree learning algorithm has high accuracy for detecting the type of the terror incidents. Though the SVM attained high accuracy, the longer execution time is encountered when the dataset is large. The Bayes scored a faster running time at the cost of lower accuracy.

In this research we used the GTD database which holds the comprehensive collection of all terrorist events occurred across the globe between 1978 and 2013 and we propose to extract useful information from this dataset and experimentally prove that the classification techniques like decision tree and ensemble classifier can learn from the dataset to detect the attack_type in the given GTD. The next section gives a detailed insight into the classification algorithms.

III CLASSIFICATION ALGORITHMS

3.1 Decision Tree

The decision tree algorithm generates the tree structure by considering the values of one attribute at a time. Initially the algorithm sorts the dataset based on the value of the attribute. Then it proceeds further looking for the regions that possess single class and identifies them as leaves. For the rest of the regions that contain more number of classes, the decision tree choose another attribute and the branching process is continued till the all the leaves have been identified or there is no attribute capable of producing one or more leaves.

3.2 Ensemble Classifier

Ensemble learning techniques have been shown to increase machine learning accuracy by combining arrays of specialized learners. These specialized learners are trained as separate classifiers using various subsets of the training data and then combined to form a network of learners that has a higher accuracy than any single component. Ensemble techniques increase classification accuracy with the trade-off of increasing computation time. Training a large number of learners can be time-consuming, especially when the dimensionality of the training data is high. Ensemble approaches are best suited to domains where computational complexity is relatively unimportant or where the highest possible classification accuracy is desired.

Cite this article as: Sivaraman R, Srinivasan S, Chandrasekeran, R M. "Big Data On Terrorist Attacks: An Analysis Using The Ensemble Classifier Approach." *International Conference on Inter Disciplinary Research in Engineering and Technology* (2015): 255-261. Print.

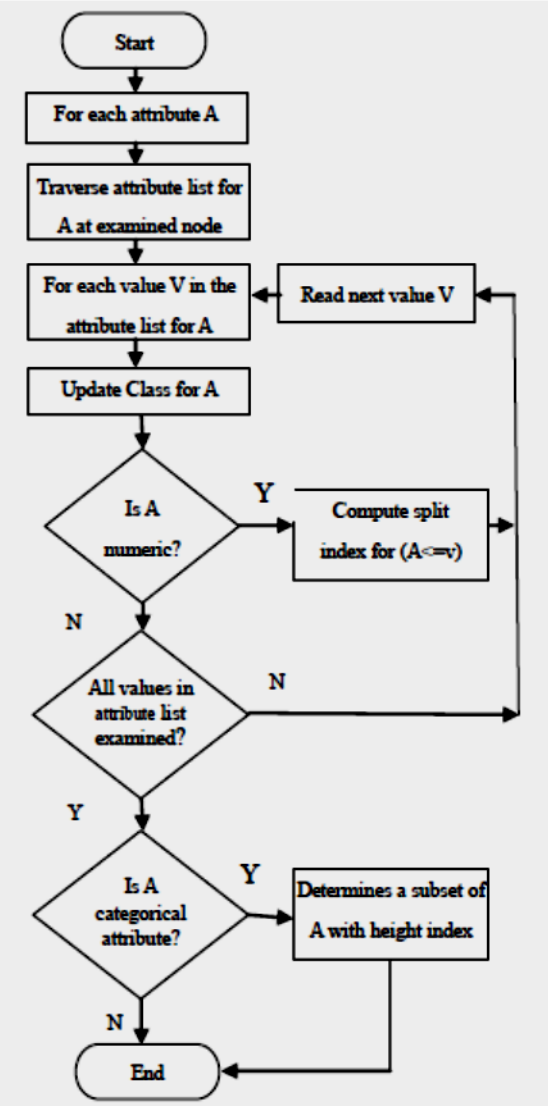

Fig.1. Flow Chart for Decision Tree

Input: Training sets $T=(x_i,y_i)$,i=1 to n: Integer n (iteration number).

Output: Classifier H(x).

For each iteration i= 1 to n

{

Select a subset T_i, of size N form the original training examples T.

The size of T_i is the same with the T where some instances may not appear in T_i, while other appear more than ones.

Generate a classifier $H_i(x)$ from the T_i

}

Fig.2. Pseudocode for bagging

IV PREPROCESSING OF DATA

In this research we used the Global Terrorism Database created by the START: A Center of Excellence of the U.S. Department of Homeland Security and University of Maryland containing terrorist attack data from 1970 to 2013. The original dataset is in the Microsoft Excel format and they have been converted into the ARFF format (Attribute Relation File Format) which is accepted by the Weka tool. From the various fields available in the GTD we have used the year of occurrence, month, day, country, city, attack type, target type, terrorist group name, weapon type, hostage situation and ransom were utilized.

V EXPERIMENTAL ANALYSIS

The dataset under study comprised of 125088 records spanning over the years 1970 to 2013. Each terrorist attack instance is mapped with 17 attributes. The month-wise description of datasets pertaining to terrorist attacks is shown in the Table 1, different attack types in Table 2, weapon types in Table 3 and performance classifiers in Table 4. The experiments are conducted using two well acclaimed classification algorithms, viz., Decision tree J48 which is the WEKA's implementation of C4.5 and Ensemble Classifier.

Cite this article as: Sivaraman R, Srinivasan S, Chandrasekeran, R M. "Big Data On Terrorist Attacks: An Analysis Using The Ensemble Classifier Approach." *International Conference on Inter Disciplinary Research in Engineering and Technology* (2015): 255-261. Print.

Nos	Month	Count
1	JAN	10017
2	FEB	9152
3	MAR	10453
4	APR	10401
5	MAY	11451
6	JUN	10541
7	JUL	11127
8	AUG	11005
9	SEP	9822
10	OCT	10999
11	NOV	10565
12	DEC	9554

Table 1. Summary of Month-wise occurrences

Nos	Attack Type	Count
1	Assassination	15740
2	Unknown	6515
3	Kidnapping	59558
4	Armed Assault	7420
5	Hijacking	30100
6	Barricade Incident	472
7	Infrastructure	3896
8	Unarmed Assault	694
9	Bombing Explosion	692

Table 2. Summary of attacktype-wise occurrences

Nos.	Weapon Type	Count
1	UNKNOWN	9570
2	EXPLOSIVES BOMBS DYNAMITES	61155
3	INCENDIARY	8519
4	FIREARMS	42898
5	CHEMICAL	206
6	FAKE WEAPONS	31
7	MELEE	2417
8	SABOTAGE EQUIPMENT	113
9	VEHICLE	58
10	RADIOLOGICAL	13
11	OTHER	72
12	BIOLOGICAL	35

Table 3. Summary of weapon type-wise occurrences

	Decision Tree		Decision Tree Ensemble	
Attack type	Class recall	Class precision	Class recall	Class precision
Assassination	87.50%	76.13%	89.02%	76.24%
Barricade Incident	90.67%	93.15%	92.00%	94.52%
Kidnapping	95.26%	95.93%	95.26%	95.40%
Infrastructure	96.91%	83.93%	97.42%	82.53%

Cite this article as: Sivaraman R, Srinivasan S, Chandrasekeran, R M. "Big Data On Terrorist Attacks: An Analysis Using The Ensemble Classifier Approach." *International Conference on Inter Disciplinary Research in Engineering and Technology* (2015): 255-261. Print.

Unknown	51.03%	83.90%	50.00%	83.62%
Armed Assault	83.33%	71.43%	66.67%	100.00%
Bombing Explosion	55.56%	38.46%	22.22%	66.67%
Unarmed Assault	48.13%	79. 90%	50.00%	82.02%
Hijacking	63.64%	53.85%	63.64%	53.85%

Table 4: Performance of classifiers

Evaluation measures for determining Accuracy, Precision and Recall were done using the following calculation methodologies,

Accuracy = $(T_p+T_n) / (T_p+T_n+F_p+F_n)$

Precision = $T_p / (T_p + F_p)$

Recall = $T_p / (T_p + F_n)$

T_p represents the number of terror incidences correctly classified for a particular class, F_p denotes the number of occurrences which is incorrectly classified as particular class, T_n depicts the number of incidence that were correctly classified as other class and F_n represents the total number of incidents that were incorrectly classified as another class. The bagging model is employed using Weka tool. Decision tree is used as base classifier and number of iterations used is 5.other parameters for meta classifier and base learner use the default values available in the tool. Ten fold cross validation is used. From the results it is evident that, the ensemble method performed well and often outperformed the single model in terms of precision &recall. Thus, a bagging ensemble can be used with the reasonable assumption that it will not affect performance on s datasets. If time and computational resources are not an issue or the highest possible classification accuracy is desired, then the bagging ensemble model seems to be the best choice.

Cite this article as: Sivaraman R, Srinivasan S, Chandrasekeran, R M. "Big Data On Terrorist Attacks: An Analysis Using The Ensemble Classifier Approach." *International Conference on Inter Disciplinary Research in Engineering and Technology* (2015): 255-261. Print.

```
Bagging (prediction model for label attacktype1)
Number of inner models: 10

Embedded model #0:
ishostkid > 0.500
|   property > 0.500
|   |   iyear > 1972.500: 5 {1=0, 6=4, 3=0, 7=1, 2=12, 4=1, 9=1,
8=0, 5=26}
|   |   iyear ≤ 1972.500
|   |   |   iday > 6.500
|   |   |   |   imonth > 5.500: 5 {1=0, 6=2, 3=0, 7=0, 2=0, 4=0,
9=0, 8=0, 5=2}
|   |   |   |   imonth ≤ 5.500: 2 {1=0, 6=0, 3=1, 7=0, 2=2, 4=0,
9=0, 8=0, 5=0}
|   |   |   iday ≤ 6.500: 4 {1=1, 6=0, 3=0, 7=0, 2=0, 4=4, 9=0, 8=0,
5=0}
|   property ≤ 0.500
|   |   weaptype1 > 11
|   |   |   targtype1 > 5
|   |   |   |   targtype1 > 6.500
|   |   |   |   |   iday > 30: 6 {1=0, 6=3, 3=0, 7=0, 2=0, 4=0, 9=1,
8=0, 5=0}
|   |   |   |   |   iday ≤ 30
|   |   |   |   |   |   imonth > 10.500
|   |   |   |   |   |   |   iyear > 1972.500: 4 {1=0, 6=1, 3=0, 7=0,
2=0, 4=1, 9=0, 8=0, 5=0}
|   |   |   |   |   |   |   iyear ≤ 1972.500: 6 {1=0, 6=5, 3=0, 7=0,
2=0, 4=0, 9=0, 8=0, 5=0}
|   |   |   |   |   |   imonth ≤ 10.500: 6 {1=0, 6=67, 3=0, 7=0,
2=0, 4=0, 9=1, 8=0, 5=0}
|   |   |   |   targtype1 ≤ 6.500: 4 {1=0, 6=1, 3=0, 7=0, 2=0, 4=15,
9=0, 8=0, 5=0}
|   |   |   targtype1 ≤ 5: 6 {1=0, 6=125, 3=0, 7=0, 2=0, 4=0, 9=0,
8=0, 5=0}
|   |   weaptype1 ≤ 11
|   |   |   iday > 2.500
|   |   |   |   targtype1 > 6.500
|   |   |   |   |   iyear > 1976.500: 2 {1=0, 6=0, 3=0, 7=0, 2=2,
4=0, 9=0, 8=0, 5=0}
|   |   |   |   |   iyear ≤ 1976.500
|   |   |   |   |   |   weaptype1 > 7: 1 {1=1, 6=1, 3=0, 7=0, 2=0,
4=0, 9=0, 8=0, 5=0}
|   |   |   |   |   |   weaptype1 ≤ 7
|   |   |   |   |   |   |   iday > 27: 6 {1=0, 6=1, 3=0, 7=0, 2=1,
4=0, 9=0, 8=0, 5=0}
|   |   |   |   |   |   |   iday ≤ 27
|   |   |   |   |   |   |   |   imonth > 2.500: 6 {1=0, 6=9, 3=0,
7=0, 2=0, 4=0, 9=0, 8=0, 5=0}
|   |   |   |   |   |   |   |   imonth ≤ 2.500: 2 {1=0, 6=1, 3=0,
7=0, 2=1, 4=0, 9=0, 8=0, 5=0}
|   |   |   |   targtype1 ≤ 6.500
|   |   |   |   |   targtype1 > 5: 4 {1=0, 6=0, 3=0, 7=0, 2=0, 4=14,
```

Fig 3 Decision Tree Ensemble prediction model (WEKA)

Cite this article as: Sivaraman R, Srinivasan S, Chandrasekeran, R M. "Big Data On Terrorist Attacks: An Analysis Using The Ensemble Classifier Approach." *International Conference on Inter Disciplinary Research in Engineering and Technology* (2015): 255-261. Print.

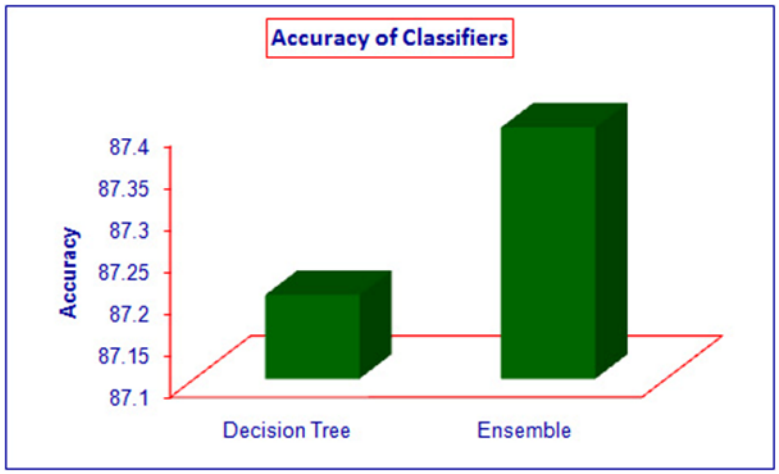

Fig 4. Accuracy of classifiers

VI CONCLUSION AND FUTURE SCOPE

Emerging technology advancements in the Big Data science has opened up new frontiers for research in this arena. However Big the data may look, it is always Small when approached with suitable methods and procedures. It is only in the hands of the researchers for devising tactical strategies to pull out meaning full output from the presented data. Though this research has been limited only decision algorithm and ensemble classifier, this can be extended to incorporate multiple methods or combination of methods. Our future steps include designing a recommender based conceptual framework to analyze the Big Data in real time using Cloud platform.

REFERENCES

[1] http://www.start.umd.edu/gtd/
[2] Nitin Nandkumar Sakhare, Swati Atul Joshi, Classification of Criminal Data Using J48-Decision Tree Algorithm, IFRSA International Journal of Data Warehousing & Mining |Vol 4|issue3|August 2014
[3] Sarwat Nizamani, Nasrullah Memon, Uffe Kock Wiil, Panagiotis Karampelas, Modeling Suspicious Email Detection using Enhanced Feature Selection, International Journal of Modeling and Optimization, 2010-3697
[4] Borooah, Vani K. 2009. "Terrorist Incidents in India, 1998-2004: A Quantitative Analysis of Fatality Rates." Terrorism & Political Violence 21:476-498.
[5] Chang, Charles and Ying Ying Zeng. 2011. "Impact of Terrorism on Hospitality Stocks and the Role of Investor Sentiment." Cornell Hospitality Quarterly 52:165-175.
[6] Enders, Walter, Todd Sandler, and Khrsrav Gaibulloev. 2011. "Domestic Versus Transnational Terrorism: Data, Decomposition, and Dynamics." Journal of Peace Research 48:319-337.
[7] Young, Joseph K., and Laura Dugan. 2011. "Veto Players and Terror." Journal of Peace Research 48:19-33.
[8] Sarwat Nizamani, Nasrullah Memon, Analyzing News Summaries for Identification of Terrorism Incident Type, Educational Research International Vol. 3(4) August 2014

Cite this article as: Sivaraman R, Srinivasan S, Chandrasekeran, R M. "Big Data On Terrorist Attacks: An Analysis Using The Ensemble Classifier Approach." *International Conference on Inter Disciplinary Research in Engineering and Technology* (2015): 255-261. Print.