Use of Predictive Modeling for Prediction of Future Terrorist Attacks in Pakistan

Hina Muhammad Ismail
Isra University
Hyderabad, Pakistan

Hameedullah Kazi, PhD
Isra University
Hyderabad, Pakistan

ABSTRACT
Recent years have seen an increased interest in Data mining related to terrorism. Large volumes of terrorism records can be analyzed efficiently using data mining techniques to get solutions for crime investigation by law enforcement agencies. On 2014 edition of Global Terrorism Index (GTI, 2015), which systematically rank and compares 162 countries to the impact of terrorism, Pakistan was ranked as the third most affected country. The area of predicting terrorist incidents in the perspective of Pakistan is not adequately explored by the data mining research community which assert a serious concern. This study is focused on analyzing Incident data set from Global Terrorism Database (GTD) specific to Pakistan from year 1970 to 2014 by using predictive modeling. Prediction of future terrorist attacks according to City, Attack type, Target type, Claim mode, Weapon type and Motive of attack through classification techniques will facilitates the decision making process by security organizations as to learn from the previous stored attack information and then rate the targeted sectors/ areas of Pakistan accordingly for security measures.

Keywords
Predictive Modeling, terrorism

1. INTRODUCTION

1.1 Background
Post 9/11 world opens up new channels for scholars to research and security agencies for investigative operations. Consequently, domain experts realized the need of gathering knowledge about the structure of terrorist organizations and to attain information about their operational behavior which is the key to win war on terror. Security agencies are incapacitated without reliable "Data and Data Analytics tools" (Nieves & Cruz, 2011). There exist noticeable databases which keep track of terrorism history, some of them are Global Terrorism Database (GTD, 2015), RAND Database of Worldwide Terrorism Incidents (RDWTI, 2015) and Suicide Attack Database maintained by Chicago Project on Security and Terrorism (CPOST, 2015). Vast amount of terrorism datasets asserts the need of Data Mining tools and techniques for effective analysis of data and producing knowledge that facilitates the process of decision making in counter terrorism (Bobbitt, 2010).

1.2 Definition and Characteristics of DM
Data mining is the process of uncovering patterns in data in order to gain knowledge from large volumes of data (Witten et al., 2011). Generally, there are two kinds of data mining tasks, Predictive and descriptive. Predictive mining tasks predict the value of a specific attribute by performing induction on the current data. Descriptive mining tasks describe properties of the data in a target data set (Han et al., 2012).

1.3 Crime Analysis through Web Mining
Sharma & Sharma (2012) performed web content mining to anticipate crime trends by using K-means algorithm. The algorithm generated two clusters from the data set namely cluster 0(genuine user) and cluster 1(illegal data detected).

These two clusters were trained using classification techniques. Implementation results showed 94.75% accuracy of clustering and 5.28% as false alarm rate. And three classes were predicted on these clusters such as soft, hard and none to show the intensity of crime such as class soft was predicted on cluster 1 with remark of alert, class hard was predicted on cluster 1 with remark of crime and class none was predicted on cluster 0 with remark of genuine user.

Uthra (2014) analyzed web mining for comprehensive study on classification and clustering techniques for crime mining. Using web mining, the classification of crime type from web comprised of traffic violations, fraud, cybercrime, violent crime, fraud, drug offences and arson. K- Means clustering algorithm was used to get structured data from web such as case no., name, crime type, judgment and location of crime.

1.4 Crime Pattern Analysis through Clustering
Malathi & Baboo (2014) applied Enhanced Missing value algorithm and Clustering algorithm on crime dataset from Indian Police department to help in the process of filling missing values of crime data and identification of crime patterns from historical data. Experimental results proved that Crime theft was in flux, it got decreased in year 2007, it got increased in 2008 and 2009, and once again it got increased in 2010. Theft Robbery kept increasing from 2007 to 2010.

Agarwal et al., (2013) performed Homicide crime analysis on offences records from England police and wales by offence and police force area from 1990 to 2012 that composed of attributes such as year, homicide, attempted murder, child destruction and death caused by careless driving and analysis was performed in Rapid miner tool using K-means clustering technique. Plotting clusters of homicide crime with respect to year concluded that homicide is decreasing from 1990 to 2012. These results helped police force to identify crime trend over years and design precaution strategy for future accordingly.

1.5 Predictive Crime Analysis
Khelifa et al., (2013) in their research on "Predictive surveillance for the detection of suspicious objects" used predictive algorithm, Naïve Bayes classifier to predict and detect the existence of suspicious objects using WEKA. The researches got promising results with 88.89% characterized as correctly classified instance such that reporting to correct target class and 11.11 % classified as incorrect instances.

Deshmukh et al., (2014) proposed a mathematical model for crime investigation in which they compared three data mining techniques J48, Naïve Bayes and JRip against sample criminal dataset. Performance comparison is based on accuracy of classification, precision, recall, True positive rate, FP rate etc. The proposed model then suggests the best performing algorithm for sample crime data set and criminal database for the identification of possible suspect crime.

1.6 Data Mining Applications in Counter Terrorism in Pakistan's context

Hussain (2010) in his research on "Terrorism in Pakistan: Changing incident patterns" described empirically the changing incident patterns of terrorism in Pakistan from 1974 to 2007 and studied the temporal (time) and spatial (location) patterns in terrorist incidents. This paper also compare and contrast these patterns, pre and post U.S led invasions of Afghanistan and Iraq. Khan & Din (2014) in their research study on "Data classification and Text mining in CRIMS in the perspective of Pakistan" analyzed the issues of TM in Urdu language specific to Crime Record Management in Pakistan. In the context of CRIMS, preprocessing of text is explained which is to apply techniques to clean and structure the textual data for analysis such as stop word removal, stemming algorithm and document indexing. Categorization techniques is explained by giving the pros and cons of classifier techniques for text classification such as K-nearest neighbor (KNN), Decision tree, Naïve Bayes algorithm and Artificial neural networks. It is concluded that appropriate algorithm for classification of textual documents depends upon problem domain.

1.7 GTD Review

Global Terrorism Database is a terrorism incident database, maintained by the National Consortium for the Study of Terrorism and Responses of Terrorism. Current GTD incorporates incidents of terrorism from 1970 to 2014. GTD information gathering is directed by START staff at the University of Maryland. GTD is the consequence of many steps of data collection efforts, which incorporate media articles and electronic news documents and secondary materials.

Sixteen relevant variables have been selected for the purpose of this study i.e. Year, Month, City, Incident Hotspot, Successful attacks, Suicide Attacks, Attack type, Target type, Corporate entity, Specific target, Target nationality, Motive, Medium/Mode for Claim of responsibility, Number of perpetrators, Number of perpetrators captured and Weapon type. Among 16 variables selected for this study, following 6 variables are selected as class/predicted variables.

1. **City**: Prediction of potential terrorism hotspot based on spatial distribution of existing data.
2. **Attack type**: Anticipate incident with respect to type of attacks occurred before.
3. **Target type**: Anticipate incident according to entities targeted before.
4. **Claim mode**: Classification of medium of claim of responsible attackers in order to predict future incidents.
5. **Weapon type**: Analyze usage trend of different types of weapon used for terrorist activities previously in order to anticipate incidents.
6. **Motive of attack**: To extract knowledge about particular aim of attack by analyzing previous attack's motives.

1.8 Problem Definition

Large volumes of terrorism records can be analyzed efficiently using data mining techniques to get solutions for crime investigation by law enforcement agencies. The area of predicting terrorist incidents in the perspective of Pakistan is not adequately explored by the data mining research community which assert a serious concern.

Prediction of variables such as City, Attack type, Target type, Motive of attack, Claim mode and Weapon type through classification techniques can be very useful in making prediction of future attacks. Consequently, it will facilitates the decision making process by security organizations as to learn from the previous stored attack information and then rate the targeted sectors/ areas of Pakistan accordingly for security measures.

2. PROPOSED METHODS AND TOOL

2.1 Predictive Modeling

A predictive modeling is the method by which a model is made to anticipate the outcome. It predicts the value of future data trends utilizing current or historical data. In case of categorical outcome, it is known as Classification and for numerical outcome, it is called Regression. As this study is about prediction of future terrorist's attacks in Pakistan, and all class/outcome variables are categorical. Hence, Classification is used to anticipate outcome variables. Classification makes assignment of the data into known or predefined classes. In that sense, it is known as supervised learning technique because classes are predetermined on the basis of data attribute values. It is a two-step procedure, comprised of learning step and classification step. In learning step, classification model is developed using training data set. This learning step is typically referred as supervised learning because class label of very training sample is given. Normally, the learned model is presented in the form of classification rules. These rules are used in classification step for prediction of future data samples.

2.2 Ensemble Classifier

Ensemble Classifier is based on an analogy that advisory group of people often come up with smarter choices than individual specialists. Many machine learning techniques learn an ensemble of models that usually increase predictive performance using combined performance benefits over single model. Natural way to ensemble different learned models in case of classification is to make a vote and for numeric prediction it is to compute the average. To achieve the purpose of this study, a Meta learning algorithm named Vote is implemented using WEKA data mining system, which is a base method for classifiers' combination. Vote algorithm takes base (individual) classifiers and combination rule for classifiers as parameters.

Thus, the classification technique employed in this paper makes use of two base classifiers namely, Bayesian classifier (Bayesian belief Network) and Decision tree (J48) algorithms with combination rule of Average of probabilities to build a learned classification model on training data set which is later used for prediction of future terrorist attacks using test data set.

2.3 Bayesian Classifier

These are statistical classifiers that can anticipate class membership probabilities for example, the likelihood that a given record belongs to specific class. These classifiers

haveshown high speed and accuracy when implemented on large databases.

This study uses BayesNet, a Bayesian Belief Networks algorithm (BN), which is a graphical model that allows joint conditional probability distribution that is dependencies among subsets of attributes. This classifier facilitates model of causal relationship, on which learning can be performed, which is later used for classification. BNs are also known as Belief networks or probabilistic networks. They are composed of two components i.e. directed acyclic graph, where every node shows a random variable and every arc represents probabilistic dependence. Second component of BNs is conditional probability table (CPT) for each variable. Joint probability of any tuple (x1,…....., xn) corresponding to variables (X1, …….., Xn) is estimated by equation 2.3:

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | Parents(Y_i))$$

2.4 Joint Probablity Estimation

Where the values of $P\ (xi \mid (Xi))$ match to the Conditional probability table entries for (Xi). (Han et al., 2012)

2.5 Decision Tree Classifier

Decision tree (DT) is a technique for predictive data modeling used in classification and clustering tasks. DT use divide and conquer approach for splitting the problem search area into subsets. Root node initiates the DT while Leaf nodes represent the successful guess or correct prediction.

ID3 (Iterative Dichotomiser 3) DT algorithm was developed by J. Ross. Quinlan in late 1970s and in early 1980s. This algorithm was later enhanced by Quinlan under the label of C4.5. J4.8 is the java implementation of C4.5 in WEKA which is used in this study in Ensemble technique for training a classifier. The ID3 algorithm uses information theory and seek to lower the number of comparisons. Primary approach used by ID3 is to select splitting attributes with highest information gain. And the amount of information related to attribute depends upon the likelihood of the occurrence. The concept of information gain is associated with the idea of entropy which measures the amount of doubt or uncertainty in the data. Entropy lies between 0 and 1. Entropy is zero when all the data belongs to a single class, hence, there is no uncertainty in the data. Formal Definition of an entropy is given below:

$$H(p_1, p_2, ..., p_s) = \sum_{i=1}^{s} (p_i log(1/p_i))$$

Equation 2.4 (a): Entropy

ID3 selects splitting attribute with higher Information gain, which is the difference among information required to make a correct classification before split and information required after split. Information gain is measured by the following equation:

$$Gain(D, S) = H(D) - \sum_{i=1}^{s} P(D_i) H(D_i)$$

Equation 2.4 (b): Information gain

(Witten et al., 2011)

2.6 WEKA Data Mining System

It is an open source software distributed under the GNU General Public License (WEKA, 2015). Latest version, WEKA 3.7.12 is used in this study. Upon launching the software, WEKA GUI chooser initiates with choice among major WEKA GUI applications and supporting tools. The main GUI applications are Explorer, Experimenter, Knowledge Flow and Command Line Interface. Tools include package manager for installation and maintenance of different packages, ARFF (Attribute file relation format) viewer helps in viewing ARFF files in spread sheet format. Sqlviewer facilitates querying SQL worksheet through JDBC connection. Bayes Net editor offers edition, visualization and learning of Bayes networks.

3. RESULTS

Experimental results are discussed in terms of Preprocessing of data set, Cross Validation, Evaluation measures and Classification results.

3.1 Preprocessing of Data set

Data preprocessing techniques can significantly improve the accuracy and the overall time required for the mining process. In the context of Knowledge discovery from data or KDD, there exists four significant steps of data preprocessing, namely, data cleaning, data integration, data reduction and data transformation. Data cleaning, data reduction and data transformation are discussed in detail because of their applicability for this study.

Data Cleaning

Data cleaning attempts to clean the data by removing outliers from the data, by removing duplicate tuples, by resolving inconsistencies or by filling in missing values. In most data mining applications, missing values are either filled in manually or through probable value of an attribute or by central tendency of an attribute but in case of GTD data set, missing values have been retained to maintain data reliability. Moreover, inaccurate or incorrect values in database have been corrected to make data consistent for mining. GTD dataset contains many typographic errors that leads to an extra value for variable as shown in Table 1 that were removed to clean the data.

Table 1. Data Cleaning

Variables	Records with Typographic Errors, Inconsistencies and Redundant data
City	158
Corporate Entity	631
Specific Target	209

Data Selection

Data investigation and mining on large volumes of data can take a lot of time, making such mining unreasonable or unpractical. In this phase, data relevant to analysis have been retrieved by data reduction techniques. In GTD case,

dimension reduction has been performed on GTD data set. Dimension reduction deals with detection and removal of irrelevant or redundant attributes/variables or instances.

Table 2. Data Selection

Original GTD Data set	**Filtered on Pakistan**
133 Variables	16 Variables
Record of 141967 Global terrorism incident	Record of 11503 terrorism incidents

Data Transformation

In this preprocessing step, the data have been altered or unified to make subsequent data mining procedure more accurate, efficient and easier to understand. Approaches to data transformation comprise of Smoothing, Attribute (feature) construction, Aggregation, Normalization, Discretization and Concept hierarchy generation for nominal data. However, only attribute construction and concept hierarchy generation for nominal data have been discussed in detail due to their relevancy with the GTD dataset.

Attribute (feature) construction is about development of new features from the specified set of attributes to assist in the data mining process. Concerning GTD dataset, a new feature namely Incident hotspot has been constructed from the city attribute. Because authors of GTD database have wrongfully included towns, villages and streets in city attribute that caused superfluous data typically data of repeated nature. Incident hotspot is introduced to reach to specific area where unrest or hostile action took place.

Concept hierarchy generation for nominal data, where attributes with lower level concepts have been generalized to higher level concepts such as streets to city. In case of GTD dataset, City and Motive variables have been transformed using this technique based upon number of distinct values per attribute.

The sample results of applying Concept hierarchy generation on City variable is shown in Table 3 where town, villages, streets or markets are generalized into their respective cities.

Table 3. Application of Concept Hierarchy Generation on City Variable

S/N	**Cities**	**Towns/ Villages/Streets or Markets**
1.	Karachi	Metroville, Kiamari, Memon Goth, Sabzi Mandi, Ghanchi Para, Nagan Chowrangi, Khyber Chowk, North Nazimabad, Korangi, Bin Qasim, Gadap, Ancholi, Shantinagar, Aram Bagh, Usmanabad, Bakra Peeri, Haji Abdulnabi Brohi, Quaidabad, Sherpao Colony, Shamsi, Surjani, Albela Chowk
2.	Peshawar	Maira, Adezai, Aza Khel, Shahi Bagh, Chamkani, Gulbela, Mathra, Hayatabad , Sarband, Garhi Qamardin, Pishta Khara, Shaikhan, Sheikh Muhammadi, Surizai Bala, Hasan Garhi, Larama, Faqirabad, Gulbahar Sarbanan, PushtaBajaur, Gandab, Badaber, Bara Sheikhan

The results of applying Concept hierarchy generation on Motive variable is shown in Table 4 where number of distinct values for motive are grouped under the label of specific motive of attack.

Table 4. Application of Concept Hierarchy Generation on Motive Variable

S/N	**Motive**	**Number of distinct values**
1.	To implement Taliban-inspired Sharia	➢ To prevent men from having their hair or beards trimmed in accordance with Taliban-inspired Sharia ➢ Tehrik-i-Taliban Pakistan (TTP) claimed responsibility for the incident, stating that they targeted Yousafzai because she was critical of the TTP and promoted Western ideals in the region
2.	To trigger sectarian violence in Pakistan	➢ The assassination of Javed Qadri, a leader of the Sunni Tehrik party, was attributed to religious rivalry ➢ Police said they believe the killings were in revenge for the shooting deaths of several Shiite Islamics on the previous day

3.2 Cross Validation

Generally, two-thirds of the data is used as training set while the remainder one-third is used for testing. In cross validation, data is split into fixed number of partitions or folds. For GTD data set, 10 folds of data is used for cross-validation.

3.3 Evaluation Measures

After classification model has been developed, it becomes necessary to evaluate performance of a classifier in order to know how accurate a classifier is at predicting class label of tuples. Predictive Model developed for this study is evaluated on the basis of three Evaluating Measures mentioned below:

1. Accuracy
Accuracy or Recognition rate of a classifier on a specified test set is the percentage of correctly classified test set tuples.
It is given as:

$$\text{Accuracy} = \frac{TP + TN}{P + N}$$

Where:
True Positive (TP) specify the positive tuples that were correctly classified by the classifier.
True Negative (TN) specify the negative tuples that were correctly classified by the classifier.
Positive (P) specify all the positive tuples.

Negative (N) specify all the negative tuples.

2. Error Rate
Error Rate which is also known as Misclassification rate on a specified test set is the percentage of incorrectly classified test set tuples.
It is given as:

$$\text{Error Rate} = \frac{FP + FN}{P + N}$$

Where:
False Positive (FP) specify negative tuples that were misclassified as positive.
False Negative (FN) specify positive tuples that were misclassified as negative.
Positive (P) specify all the positive tuples.
Negative (N) specify all the negative tuples.

3. ROC area / AUC area
ROC curve allows us to visualize the trade-off between the rates at which the model can accurately recognize positive cases versus the rate at which it mistakenly identifies negative cases as positive for different portions of the test set. They plot the TP rate (Sensitivity) on the vertical axis against FP rate (1-Specificity) on the horizontal axis.
True Positive (TP) rate is also referred to as Sensitivity or Recall is the proportion of tuples which were classified as class a, among all those which actually have class a.
It is given as:

$$\text{True Positive rate} = \frac{TP}{TP + FN}$$

False Positive (FP) rate is also referred to as Fall-out is the proportion of tuples that belong to different class but classified as class a, among all those which are not class a.
It is given as:

$$\text{False Positive rate} = \frac{FP}{FP + TN}$$

Each point on the ROC plot represents a sensitivity/specificity pair corresponding to a particular decision threshold. A test with perfect discrimination (no overlap in the two distributions) has a ROC plot that passes through the upper left corner (100% sensitivity, 100% specificity). Therefore the closer the ROC plot is to the upper left corner, the higher the overall accuracy of the test. Accuracy is measured by the area under the ROC curve, which is known as ROC area or Area under curve (AUC). An area of 1 represents a perfect test; an area of .5 represents a worthless test.

Accuracy is classified according to following five categories according to Area under Curve (Tape, T.G. 2015).

.90-1 = Excellent (A)
.80-.90 = Good (B)
.70-.80 = Fair (C)
.60-.70 = Poor (D)
.50-.60 = Fail (F)

3.4 Classification Results

Experimental Results from Classification model developed in WEKA for Six Class/Predicted variables are mentioned in this section.

Variable City
Figure 1 shows Accuracy and Error rate for predicting tuples according to City.

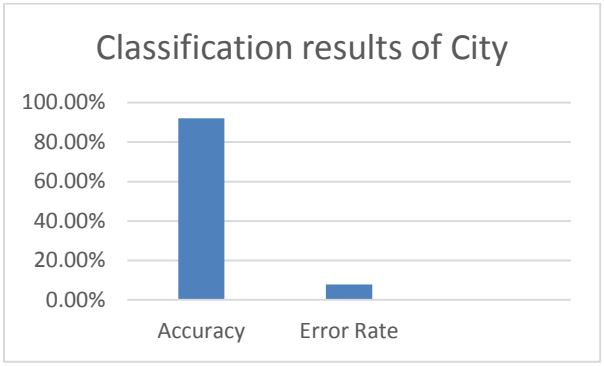

Fig.1 Chart showing Classification results of City

WEKA generates ROC curves for all values of class variable. For City variable, 134 ROC curves are generated according to 134 class values of this variable. To represent sample results of ROC curves for city variable, Roc curve for Karachi city is shown in Fig 2.

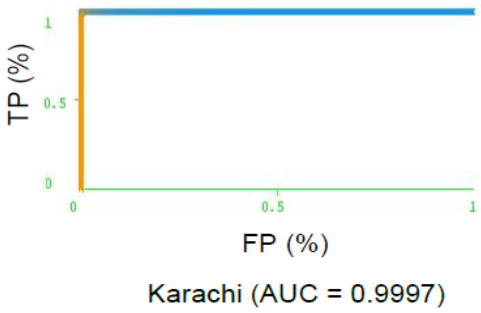

Fig.2 ROC curve for City: Karachi

Performance of a Prediction model according to ROC area/ AUC all targeted cities is shown in Table 5.

Table 5: Evaluation Results of City

Total 134 Cities	**ROC area/AUC**
126 Cities	.90 -1 = Excellent (A)
Cities : Pakpattan and Sujawal	.70-.80 = Fair (C)
Cities : Kohistan	.60- .70 = Poor (D)
Cities: Hunza-Nagar, Tando Muhammad Khan, Khushab, Sherani, Mirpur and Sudhnati	.50-.60 = Fail (F)

Variable Attack Type
Figure 3 shows Accuracy and Error rate for predicting tuples according to Attack type of terrorist activity.

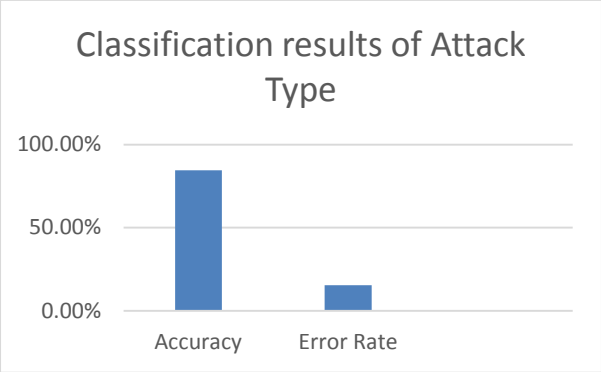

Fig.3 Chart showing Classification results of Attack Type

For Attack type variable, 9 ROC curves are generated according to 9 class values of this variable. To represent sample results of ROC curves for Attack type variable, Roc curve for Assassination attack type is shown in Fig 4.

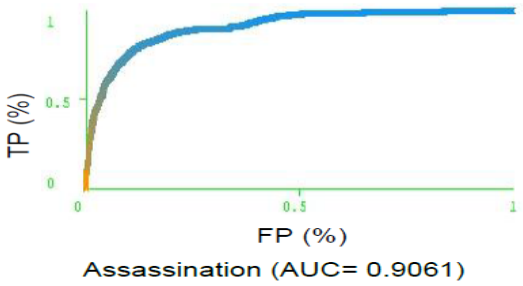

Fig.4 ROC curve for Attack Type: Assassination
Performance of a Prediction model according to ROC area/ AUC w.r.t all Attack types is shown in Table 6.

Table 6: Evaluation Results of Attack type

Total 9 Attack Types	**ROC area/AUC**
7 Attack Types: Assassination, Bombing/Explosion, Facility/Infrastructure, Armed Assault , Unarmed Assault, Unknown, Hostage Taking (Kidnapping)	.90 -1 Excellent (A)
Attack type: Hijacking	.70-.80 = Fair (C)
Attack type: Hostage Taking (Barricade Incident)	.60- .70 = Poor (D)

Variable Target Type
Figure 5 shows Accuracy and Error rate for predicting tuples according to Target type of terrorist activity.

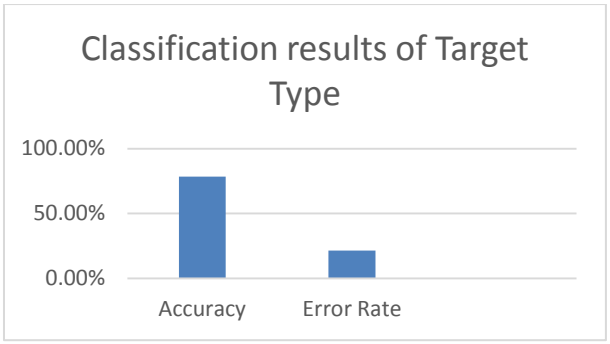

Fig.5 Chart showing Classification results of Target Type

For Target type variable, 9 ROC curves are generated according to 21 class values of this variable. To represent sample results of ROC curves for Target type variable, ROC curve for Target on Religious figures is shown in Fig 6.

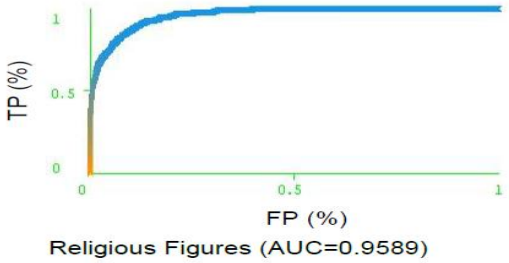

Fig.6 ROC curve for Target Type: Religious Figures

Performance of a Prediction model according to ROC area/ AUC w.r.t all Target types is shown in Table 7.

Table 7: Evaluation Results of Target type

Total 21 Target Types	**ROC area/AUC**
18 Target types: Government (General), Maritime, Government (Diplomatic), Airports and Aircraft, Terrorists/Non-State Militia, Private Citizens and Property, Military, Business, Police, Religious Figures/Institutions, Transportation, Utilities, Journalists and Media, Violent Political Party, Educational Institutions, NGO, Unknown and Telecommunication	.90 -1 = Excellent (A)
3 Target Types : Food or Water supply, Tourists and Other	.80-.90 = Good (B)

Variable Motive of Attack
Figure 7 shows Accuracy and Error rate for predicting tuples according to Motive of attack.

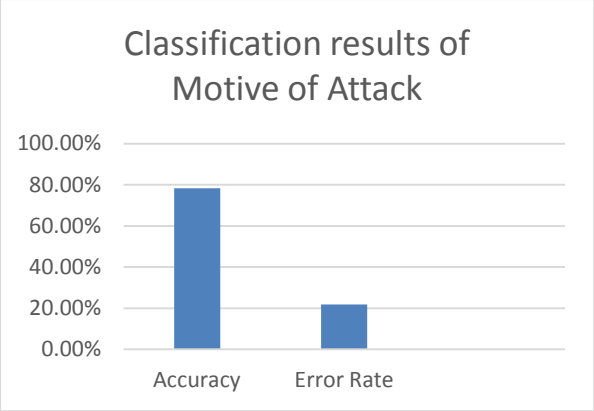

Fig.7 Chart showing Classification results of Motive of Attack

For Motive of attack variable, 72 ROC curves are generated according to 72 class values of this variable. To represent sample results of ROC curves for Motive of attack, ROC curve for Motive: To trigger Sectarian violence in Pakistan is shown in Fig 8.

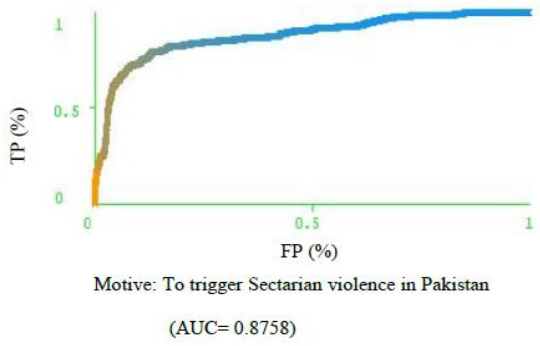

Fig.8 ROC curve for Motive of attack

Performance of a Prediction model according to ROC area/ AUC w.r.t all Motive of Attack is shown in Table 8.

Table 8: Evaluation Results of Motive of Attack

Total 72 Motives	**ROC area/AUC**
8 Motives	.90 -1 = Excellent (A)
12 Motives	.80-.90 = Good (B)
15 Motives	.70- .80 = Fair (C)
14 Motives	.60-.70 = Poor (D)
23 Motives	Below .5 = Fail (E)

Variable Claim Mode

Figure 9 shows Accuracy and Error rate for predicting tuples according to Claim mode of an attack.

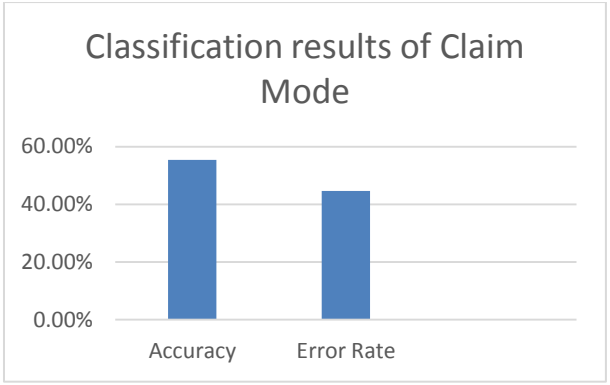

Fig.9 Chart showing Classification results of Claim mode

For Claim mode variable, 10 ROC curves are generated according to 10 class values of this variable. To represent sample results of ROC curves for claim mode, ROC curve for Claim mode: Note left at scene is shown in Fig 9.

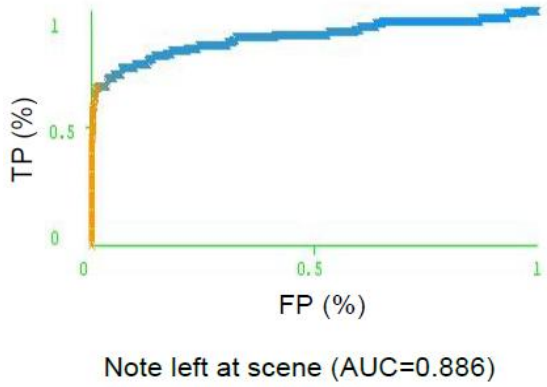

Fig.10 ROC curve for Claim mode

Performance of a Prediction model according to ROC area/ AUC w.r.t all Claim modes is shown in Table 10.

Table 9: Evaluation Results of Claim Mode of Attack

Total 10 Claim Modes	**ROC area/AUC**
3 Claim Modes: E-mail, Posted to website, Blog, etc. and Note left at scene	.80-.90 = Good (B)
Claim mode : Call (post incident)	.70- .80 = Fair (C)
3 Claim modes: Personal claim , video and other	.60-.70 = Poor (D)
3 Claim modes : letter, Call (pre-incident) and unknown	Below .5 = Fail (E)

Variable Weapon Type

Figure 11 shows Accuracy and Error rate for predicting tuples according to Type of Weapon used in attack.

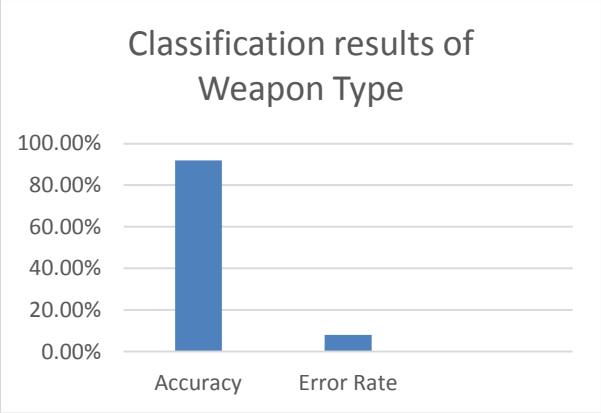

Fig.11 Chart showing Classification results of Weapon Type

For Weapon Type variable, 10 ROC curves are generated according to 10 class values of this variable. To represent sample results of ROC curves for Weapon Type, ROC curve for Weapon Type: Firearms is shown in Fig 12.

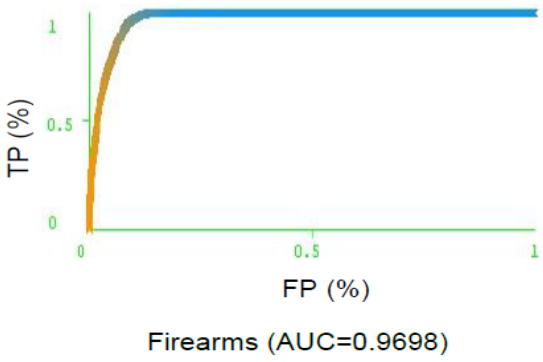

Fig.12 ROC curve for Weapon Type: Firearms

Performance of a Prediction model according to ROC area/ AUC w.r.t all Weapon Types is shown in Table 10.

Table 10: Evaluation Results of Weapon Type

Total 10 Weapon Types	**ROC area/AUC**
7 Weapon Types: Firearms, Explosives, Incendiary, Melee, Unknown, Biological and Chemical	.90 -1 = Excellent (A)
Weapon Type : Vehicle	.70- .80 = Fair (C)
Weapon Type : Sabotage Equipment	.60-.70 = Poor (D)
Weapon Type : Other	Below .5 = Fail (E)

4. DISCUSSION

The developed predictive model was evaluated using three classification evaluating metrics; Accuracy, Error rate and ROC area.

From the evaluation it is seen that:

Classification model developed for City variable achieved Accuracy rate of 92.0977% with 10594 correctly classified instances and Error rate of 7.9023% with 909 incorrectly classified instances. Attack type variable achieved Accuracy rate of 84.5779% with 9729 correctly classified instances and Error rate of 15.4221% with 1774 incorrectly classified instances. Target type variable achieved Accuracy rate of 78.5273% with 9033 correctly classified instances and Error rate of 21.4727% with 2470 incorrectly classified instances. Motive of attack variable achieved Accuracy rate of 78.2157% with 3691 correctly classified instances and Error rate of 21.7843% with 1028 incorrectly classified instances. Claim mode variable achieved Accuracy rate of 55.3897% with 668 correctly classified instances and Error rate of 44.6103% with 538 incorrectly classified instances. Weapon Type variable achieved Accuracy rate of 91.9152% with 10573 correctly classified instances and Error rate of 8.0848% with 930 incorrectly classified instances.

The evaluation results according to ROC area/ Area under curve for all the predicted variables are shown in (Table 5, Table 6, Table 7, Table 8, Table 9 and Table 10). From the evaluation results it is seen that:

1). For City variable, 134 ROC curves are generated according to 134 values of this variable. AUC for 126 cities lies between .90 -1, two cities namely Pakpattan and Sujawal achieved AUC that lies between .70-.80, city Kohistan achieved .60- .70 and six cities namely Hunza-Nagar, Tando Muhammad Khan, Khushab, Sherani, Mirpur and Sudhnati achieved area between .50-.60. This shows the predictive model developed for city provided significantly high results based upon perfect AUC for 126 cities.

2). For Attack type variable, 9 ROC are generated according to 9 values of this variable. AUC for 7 attack types lies between .90-1, Attack type; Hijacking achieved AUC that lies between .70-.80 and Attack type; Hostage Taking (Barricade Incident) achieved area between .60-.70. This shows the predictive model developed for Attack type provided significantly high results based upon perfect AUC for 7 attack types.

3). For Target type variable, 21 ROC are generated according to 21 values of this variable. AUC for 18 target types lies between .90-1, 3 Target types namely Food or Water supply, Tourists and other achieved area that lies between .80-.90. This shows the predictive model developed for Target type provided significantly high results based upon perfect AUC for 18 target types and good AUC for 3 target types.

4). For Motive of attack variable, 72 ROC are generated according to 72 values of this variable. AUC for 8 motives of attack lies between .90-1, AUC for 12 motives of attack lies between .80-.90, AUC for 15 motives of attack lies between .70-.80, AUC for 14 motives of attack lies between .60-.70 and AUC for 23 motives of attack lies below .5 which are of worthless value. This shows the predictive model developed for Motive of attack provided moderate results based upon perfect AUC for 8 motives of attack, good AUC for 12 motives and fair AUC for 15 motives.

5). For Claim mode of attack, 10 ROC curves are generated according to 10 values of this variable. AUC for 3 Claim modes lies between .80-.90, Claim mode namely, call (post incident) achieved area that lies between .70-.80, 3 values of claim mode namely, Personal claim, video and other achieved area that lies between .60-.70 and AUC for 3 values of claim mode namely; letter, Call (pre-incident) and unknown lies

below .5. This shows the predictive model developed for Claim mode provided poor results based upon AUC achieved for mode of claim for responsibility.

6). For Weapon type of attack, 10 ROC curves are generated according to 10 values of this variable. AUC for 7 weapon types lies between .90 -1, Weapon type namely Vehicle achieved area of .70-.80. Only two weapon types namely Sabotage Equipment and other got area of .6 and .2 respectively. This shows the predictive model developed for weapon type provided significantly high results based upon AUC achieved for weapon type of attack.

5. CONCLUSIONS

The focus of this research study was to make prediction of future terrorist attacks in Pakistan based upon stored information of incidents which are available in Global Terrorism Database (GTD) from year 1970 to 2014, specific to Pakistan. Prediction of terrorist attacks is performed according to City, Attack type, Target type, Motive of attack, Claim mode and Type of weapon used in attack.

The results obtained from purposed predictive model reveals that the performance is significantly high with respect to Accuracy and ROC area for 5 predicted variables except for claim mode (mode used by claimants to claim responsibility of attack). This variable got higher Error rate of **44.6103%** with 538 incorrectly classified instances. This variable was introduced in year 1998 therefore claim mode for year 1970 to 1997 contains missing values. These larger number of missing values are the reason for higher error for claim mode.

6. REFERENCES

[1] AGARWAL, J., NAGPAL, R., & SEHGAL, R. (2013). *Crime Analysis using K-Means Clustering*. International Journal of Computer Applications, Vol. 83, No. 4, pp.1-4.

[2] BOBBITT, P. (2010) The New Rules of Engagement: Nine Imperatives for Our Post-9/11 World. Newsweek , Available: http://www.newsweek.com/how-fight-war-terror-better-70899, Date accessed: 2015, 30 April.

[3] Chicago Project on Security and Terrorism: (CPOST), University of Chicago. Available: http://cpostdata.uchicago.edu/search_new.php, Date accessed: 2015, 26 April.

[4] DESHMUKH, S. R., DALVI, A. S., BHALERAO, T. J., DAHALE, A. A., BHARATI, R. S., & KADAM, C. R. (2015) *Crime Investigation using Data Mining.* International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, No. 3, pp. 22-25.

[5] Global Terrorism Index (GTI). Retrieved from: http://economicsandpeace.org/research/iep-indices-data/global-terrorism-index, Date accessed: 2015, 20 March.

[6] HAN, J., & KAMBER, M. & PEI. J. (2012) *Data Mining Concepts and Techniques*, 3rd ed., The Morgan Kaufmann series in data management systems, USA, 740 pp.

[7] HUSSAIN, S. E. (2010) *Terrorism in Pakistan: Incident Patterns, Terrorists' characteristics, And the Impact of Terrorist Arrests on Terrorism*. Ph.D. Dissertation, University of Pennsylvania, Pennsylvania.

[8] KHAN, M., & DIN, F. U. (2014) Data Classification and Text Mining in CRIMS in the Perspective of Pakistan. *Proceedings of the International Conference on Engineering and Emerging Technologie.*

[9] KHELIFA, B., AMINA, B., MADJID, M. (2013) Predictive Surveillance for the Detection of Suspicious Objects, *Proceedings of the 14th Annual Postgraduate Symposium on the Convergence of Telecommunications, Networking and Broadcasting*.

[10] MALATHI, A., & BABOO, S. (2011) *Algorithmic Crime Prediction Model on the Analysis of Crime Clusters*, Global Journal of Computer Science and Technology, Vol. 11, No. 11, pp. 47- 51.

[11] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); the WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

[12] National Consortium for the Study of Terrorism and Responses to Terrorism (START), Global Terrorism Database. Available: http://www.start.umd.edu/gtd, Date accessed: 2015, 14 January.

[13] NIEVES, S., & CRUZ, A. (2011) Finding Patterns of Terrorist Groups in Iraq: A Knowledge Discovery Analysis. *Proceedings of the 9th Latin American and Caribbean Conference for Engineering and Technology*, pp. 3-5.

[14] RAND Database of Worldwide Terrorism Incidents (RDWTI), Available: http://www.rand.org/nsrd/projects/terrorism-incidents.html, Date accessed: 2015, 30 January.

[15] SHARMA, A., & SHARMA, S. (2012). *An Intelligent Analysis of Web Crime Data Using Data Mining.* International Journal of Engineering and Innovative Technology. Vol. 2, No. 3, pp. 203-206.

[16] TAPE, T.G. Interpreting Diagnostic Tests, University of Nebraska Medical Center, Available: http://gim.unmc.edu/dxtests/, Date accessed: 2015, Nov 15.

[17] UTHRA, R. G. (2014) *Data Mining Techniques to Analyze Crime Data* , International Journal for Technological Research in Engineering , Vol.1, No. 9, pp. 882-884.

[18] WITTEN, I. H., & FRANK, E. (2011) *Data Mining Practical Machine Learning Tools and Techniques*, 3rd ed., The Morgan Kaufmann series in data management systems, USA, 665 pp.

[19] ZHAO, Y. (2012) *R and Data Mining: Examples and Case Studies*, Academic Press, Elsevier, Australia, 164 pp.