Improving the Classification of Terrorist Attacks

A Study on Data Pre-processing for Mining the Global Terrorism Database

José V. Pagán
Electrical & Computer Engineering and Computer Science Department
Polytechnic University of Puerto Rico
San Juan, Puerto Rico
e-mail: jvpagan@msn.com

***Abstract*—The objective of this paper is to analyze different data preprocessing techniques for mining the Global Terrorism Database to improve the classification of terrorist attacks by perpetrator in Iraq. Four methods for dealing with missing values (Case Deletion, Mean Imputation, Median Imputation and KNN imputation), three discretization methods (1R, Entropy and Equal Width), and three different classifiers (Linear Discriminant Analysis, K-Nearest Neighbor and Recursive Partitioning) are evaluated using ten-fold cross-validation estimates of the misclassification error. The study concludes (i) that data preprocessing can significantly reduce the classification error rate for this dataset, and (ii) that adding Global Positioning System coordinates for the location of the incidents can further reduce the classification error rate.**

Keywords- classification; data mining; data preprocessing; terrorism.

I. INTRODUCTION

Data is usually far from perfect. A focus on improving the quality of data typically improves the quality of the resulting analysis. Because data quality problems cannot always be prevented, data mining focuses on detection and correction of data problems and on the use of algorithms that can tolerate poor data quality [1].

In this paper, we analyze how different data preprocessing techniques can improve the classification of terrorist attacks by perpetrator in Iraq. Section II discusses the evolution of terrorism, its causes, and its exponential growth in Iraq over the last 30 years. Section III discusses the different techniques and tools available for analyzing terrorism data, including the Global Terrorism Database (GTD) and related tools developed by the Human-Computer Interaction Lab and START at the University of Maryland. Section IV discusses data preprocessing techniques for treating outliers and missing values, and a classification of various discretization techniques. Section V describes the experimental methodology used and the results. Finally, section VI provides a conclusion and a discussion of the results.

II. TERRORISM

Over the last 20 years, terrorism has evolved, expanded and become a significant influencing factor in international politics. The most recent definition of terrorism by the United States State Department, Department of Defense, and Central Intelligence Agency, defines terrorism as “premeditated, politically motivated violence perpetrated against non-combat targets by sub-national groups or clandestine agents, usually intended to influence an audience” [2].

In the past, terrorism events were mostly assassinations intended to influence political changes. Today, terrorist events are choreographed spectacles that reverberate in seconds around the globe, uniting extremists and shocking public audiences. Technological advances in media coverage and globalization have given terrorist events new added meaning, urgency and complexity. Consequently, Terrorism has become less predictable, its motivations less understandable and its logic of violence less clear and restrained [3].

Probably the most contested cause of terrorism is an aggrieved group resorting to violence for nationalist or separatist reasons. Depending on one's point of view, this violence can be considered as resistance against an (external) oppressor. Colonized states nationalism movements commonly turn to terrorism, it being "the resort of an extremist faction of this broader movement" within an ethnic minority [4]. To generalize it further, ethnic conflict arises from a "complex combination" of class, inequality, political opportunity, mobilization resources and "ethnic strength” [5, 6].

According to David Rothkopf, an intelligence and emerging markets advisor who served as Senior Trade Official in the Clinton Administration, the leading cause for political tension in the world today is the increasing relative inequality between social classes. Globalization has created a global super-class of the wealthiest and most elite with the most power. They run governments, the largest corporations, the powerhouses of international finance, the media, world religions, and, from the shadows, the world's most dangerous criminal and terrorist organizations. These super-class members have more in common with one another than with their own countrymen. They have globalized more rapidly than any other group and control globalization. Nationalist critics have accused them of feeding the growing economic and social inequity that divides the world [7].

Rothkopf argues that the history of mankind is one of elites rising up, overreaching, and being brought back down. They have been brought back down in revolutions and through (government) innovation. He claims that today there are many signs of backlash against these relative inequalities in

the world, as evidenced by the rise to power of rulers like Hugo Chavez in Venezuela, Mahmoud Ahmadinejadseen in Iraq, and Vladimir Putin in Russia [7].

Lee argues that geopolitics plays an important role in terrorism. For example, following the demise of the Soviet Union in 1991, many of the Marxist-Leninist terrorist attacks in Europe from the 1970s and early 1980s subsided. Today, the Marxist-Leninist terrorists of the 1970's have been replaced with more religiously oriented jihad-style groups [8].

This study focuses on terrorism activity in Iraq. Figure 1 below shows a logarithmic chart with the historical annual number of terrorist attacks in Iraq and the world since 1975 [9].

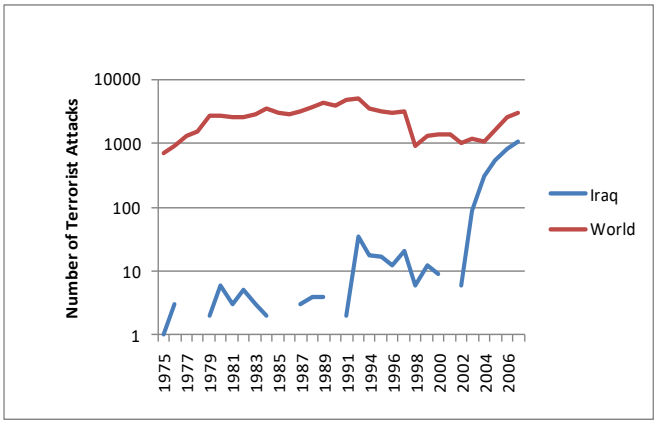


Figure 1. Historical annual number of terrorist attacks in Iraq.

As evidenced by the chart above, the wars of 1991 and 2003 have caused exponential growth in terrorist activity in Iraq. The geopolitics of the region, the large inequities in the country and the current US military occupation make Iraq a "hotbed" for terrorism.

III. Analyzing Terrorism Data

Prior to 1990, terrorism analysis was qualitative and used simple summary statistics to show trends in a particular variable over time. In 1999, Enders and Sandler used time-series number of domestic terrorist incidents to show that the end of the cold war resulted in a decrease in transnational terrorism [10]. They later used a linear model and a threshold autoregressive model to search for correlations between terrorism incidents and significant policy changes [11].

In 1996, O'Brien described the relationship between foreign policy crisis outcomes and terrorist incidents using Box - Jenkins time-series methods [12]. In 2002, Major used game theory to develop a probability distribution for insurance losses related to terrorism [13]. In 2003, Sandler and Arce used game theory to describe the interaction between terrorist groups and targeted states as a strategic game between rational actors. They concluded that tourists, civilians, and businesses are most likely targets because they have the least deterrence capability [14].

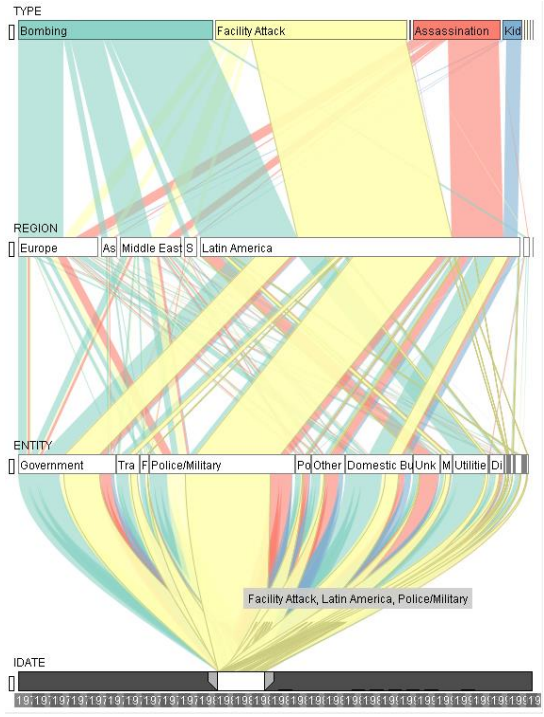
According to Guo, these modeling approaches are limited in their ability to formulate and test a valid hypothesis. They build on rigorous statistical or mathematical models, but provide limited understanding of the causes, development, and diffusion of terrorism activities. Moreover, terrorism data is often incomplete or inaccurate and only represents the outcome, not the process [15].

To counter these limitations, new approaches for visual and computational analysis have been developed for the exploration of terrorism data. These approaches can reveal unknown trends or regularities, prompt new thoughts, and help the analyst gain insights to formulate better hypotheses and models.

For example, Villanova developed a visual analytical system that focuses on depicting one of the most fundamental concepts in investigative analysis, the five W's (who, what, where, when, and why) [16]. With this approach, an investigator can interactively explore terrorist activities efficiently and discover reasons of attacks (why) by identifying patterns temporally (when), geo-spatially (where), between multiple terrorist groups (who), and across different methods or modes of attacks (what).

Similarly, Ziemkiewicz developed an interconnected visual analysis tool that provides three interconnected views of the Global Terrorism Database (GTD) to support the investigative process [17]. Figure 2 below shows one of the views from this tool used by investigators to explore the data in an abstract way by examining correlations across multiple dimensions.

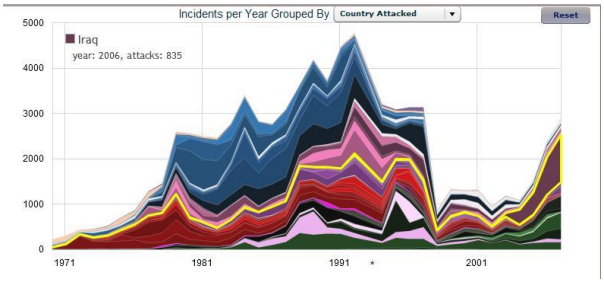
Figure 2. Visual analysis of correlations across data dimensions. [17]



The ribbons between categorical dimensions in Figure 2 show the proportion of cases from one category that fall under a category from another dimension. The visual analysis above highlights the high rate of incidents in Latin America during the early 1980s, and how those incidents break down in terms of type and target.

A. *The Global Terrorism Database*

The GTD is an open-source database including information on terrorist events around the world from 1970 through 2007. It includes systematic data on more than 80,000 domestic as well as transnational and international terrorist incidents that have occurred during this time period. For each GTD incident, information is available on the date and location of the incident, the weapons used and nature of the target, the number of casualties, and—when identifiable—the group or individual responsible. The GTD is the most comprehensive unclassified data base on terrorist events in the world. It includes information on more than 27,000 bombings, 12,000 assassinations, and 2,900 kidnappings since 1970, with information on at least 45 variables for each case. More recent incidents include information on more than 120 variables. The database is supervised by an advisory panel of 12 terrorism research experts. These experts have collected incident data from over 3,500,000 news articles and 25,000 news sources from 1998 to 2007 alone.



The Human-Computer Interaction Lab and START, both at the University of Maryland, have developed an exploratory interactive visual analysis tool for the GTD called Data Rivers. This tool allows users to investigate temporal trends in terrorism in the Global Terrorism Database (GTD) by aggregating important variables from the database and visualizing them as a comprehensible stack chart. Five different stack charts can be selected: Countries Attacked, Regions Attacked, Target Nationalities, Types of Targets and Types of Weapons. A stack chart analyzes every incident in the GTD, both domestic and international, from 1970 to 2007 (over 85,000 discrete events) and aggregates them according to the selection of the user. A unique layer is created for each data stream, with each stream reflecting a value of the variable being displayed. The thickness of each layer represents its frequency in the database.

Figure 3 below shows the GTD Data Rivers default chart (Countries Attacked) with Iraq highlighted in yellow.

Figure 3. Visual analysis of terrorism temporal trends using Data Rivers.

The chart creates a layer for every country in the GTD that has experienced a terrorist attack at some point since 1970. Thick layers represent countries that have had many attacks, whereas thin layers signify countries that had fewer attacks. One of the benefits of stack charts is that aggregate trends emerge as layers stack up like a histogram [18].

B. *Missing Data in the Global Terrorism Database*

One of the main problems of the GTD is that it is missing a large amount of data. Figure 4 below shows a distribution of missing values in the GTD by variable. Missing values are shown as white space.

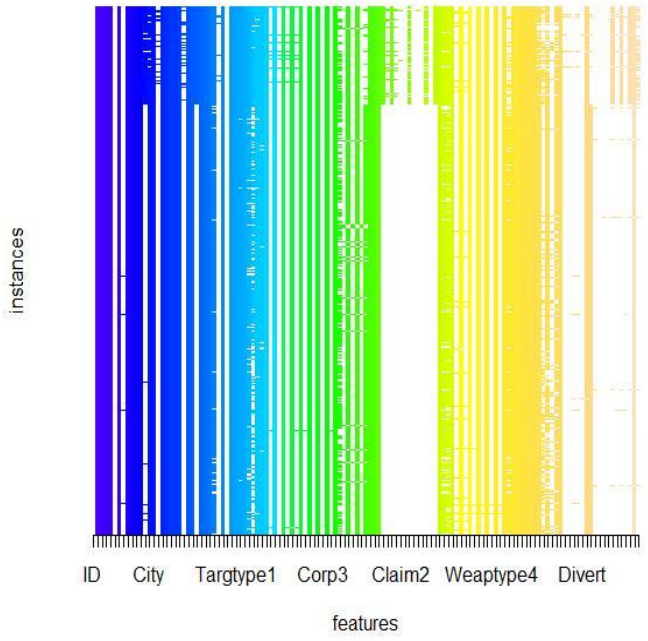


Figure 4. Global Terrorism Database distribution of missing values.

As may be seen from the chart above, 42.7% of the values in the GTD are missing, while 74% of the features and 100% of the instances have missing values. In addition, all data for the year 1993 is missing due to an office move. Since the database includes incidents covered by the media, the perpetrator remains unidentified in 39% of the terrorist attacks. Without information concerning the perpetrator of the event, it is difficult to accurately classify the incident as terrorism [8].

IV. Data Preprocessing

In this study, we pre-process raw data from the GTD by eliminating outliers, estimating missing values and discretizing the data.

A. *Eliminating Outliers*

Data objects that are grossly different from or inconsistent with the remaining set of data are called outliers. Many data mining algorithms attempt to minimize the influence of outliers or totally eliminate them. However, such action can lead to the loss of important hidden information [19].

There are several techniques available for identifying outliers and smoothing noisy data, including clustering, binning and regression. Clustering groups attribute values in clusters and then detects and removes outliers; binning sorts the attribute values and partitions them into bins; and

regression smoothes data by using regression functions. Another technique is combined human computer interaction; possible outliers are detected by the computer and checked by a human [20].

B. Treating Missing Data

Missing data is one of the biggest challenges in statistical analysis. Improper handling of missing values will distort analysis. Until proven otherwise, the researcher must assume that missing cases differ in analytically important ways from cases were values are present. The problem with missing data is the possibility that the remaining data set is biased.

Several methods have been developed for dealing with missing data, but some have drawbacks when applied to classification tasks. In 1972, Chan and Dunn considered the treatment of missing values in supervised classification using the LDA classifier for two class problems [21]. Dixon introduced the KNN imputation technique in 1979 for dealing with missing values in supervised classification [22]. In 1995, Tresp also considered the missing value problem in a supervised learning context for neural networks [23].

In 2004, Acuña analyzed the treatment of missing values in supervised classification using the LDA and KNN classifiers. According to Acuña, missing data rates of less than 1% are generally considered trivial, rates of 1-5% are manageable, and rates of 5-15% require handling with sophisticated methods. Missing data rates of more than 15% may severely impact any kind of interpretation [24].

In this study we use four methods to treat missing values: Case Deletion, Mean Imputation, Median Imputation and K-Nearest Neighbor Imputation.

Case Deletion (CD) consists of discarding all instances (cases) with missing values for at least one feature. CD should be applied only in cases in which data are missing completely at random. It is usually used when the class label is missing (when doing classification). CD is not effective when the percent of missing values per attribute varies considerably [20].

Mean Imputation (MI) consists of replacing the missing data for a given feature (attribute) by the mean of all known values of that attribute in the class where the instance with missing attribute belongs. One of the drawbacks of mean imputation is that replacing all missing records with a single value will deflate the variance and artificially inflate the significance statistical tests.

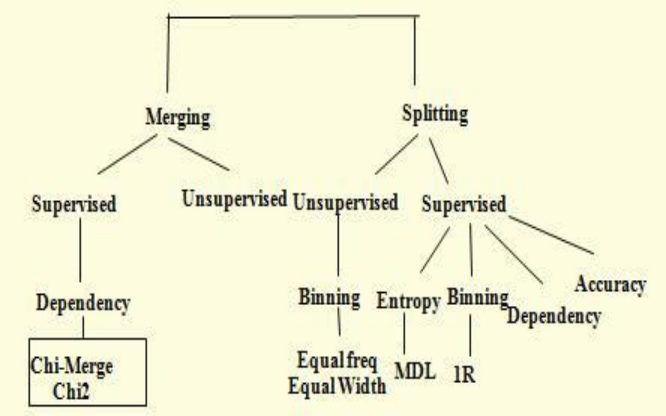
Median Imputation (MDI) consists of replacing the missing data for a given feature with the median of all known values of that attribute in the class where the instance with the missing feature belongs. This method is the recommended when the distribution of the values of a given feature is skewed because it is not affected by the presence of outliers.

KNN Imputation (KNNI) consists of imputing the missing values of an instance considering a given number of instances that are most similar to the instance of interest. The similarity of two instances is determined using a distance function.

C. Discretization Techniques

Discretization is a process that transforms quantitative data into qualitative data to improve the learning process of machine learning algorithms. Figure 5 below provides a classification of various discretization methods [25].

Figure 5. Classification of discretization methods. [25]



Splitting methods start with an empty list of cut-points and keep on adding new ones to the list by 'splitting' intervals as the discretization progresses. Merging methods start with the complete list of all the continuous values of the feature as cut-points and remove some of them by 'merging' intervals as the discretization progresses. Supervised methods use the class information when selecting discretization cut points, while unsupervised methods do not. In this study, we use three discretization methods: 1R discretization, Entropy discretization and Equal Width discretization.

1R is a supervised discretization method using binning. After sorting the data, the range of continuous values is divided into a number of disjoint intervals and the boundaries of those intervals is adjusted based on the class labels associated with the values of the feature. The adjustment of the boundary continues until the next values belong to a class different to the majority class in the adjacent interval.

Entropy methods find the best split so that the bins are as pure as possible, i.e. the majority of the values in a bin have the same class label. It finds the split with the maximum information gain.

Equal Width discretization divides the range of each feature into k intervals of equal size. This method is very straight forward, but it allows outliers to dominate the presentation and does not handle skewed data well [17].

V. EXPERIMENTAL METHODOLOGY

In this paper we carry out experiments to evaluate the effect of using different missing value and discretization methods on the misclassification error rate when classifying terrorist attacks by perpetrator in Iraq. The methods used for dealing with missing values are case deletion (CD), mean imputation (MI), median imputation (MDI) and KNN imputation (KNNI). The discretiztion methods used were 1R, Entropy and Equal Width. The classifiers considered were Linear Discriminant Analysis (LDA), K-Nearest Neighbor (KNN), and Recursive Partitioning (RPART).

Our study was carried out using the GTD database provided by the START program from the University of Maryland. The data set used for the study is a subset of the GTD that includes terrorist attacks in Iraq from 1998 to 2007. This region was selected because it has become the new "hotbed" of terrorist attacks in the world. Also, post-1997 GTD records were selected because they use a slightly modified definition of terrorism and include additional information that identifies them as terrorist incidents.

The dataset was passed through a cleansing process to minimize the number of imputations necessary. Irrelevant attributes that do not add significance in the analysis processes were eliminated. The attributes maintained for the dataset were the date and city location of the incident, the type of weapons used to commit the terrorist act, the number of casualties, the amount of wounded victims, the type of attack and the identified terrorist group responsible. The dataset instances with missing terrorist group name information were eliminated from the database. To eliminate outliers, reduce noise and simplify the classification process, we only included the top five terrorist groups shown on Table I below.

TABLE I. TOP FIVE TERRORIST GROUPS IN IRAQ 1998-2007

Group Name	*No. of Incidents*
Al-Qa`ida in Iraq	101
Al-Qa`ida	18
Ansar al-Islam	17
Islamic State of Iraq (ISI)	17
Tawhid and Jihad	16
Al-Qa`ida in Iraq	101

These five groups account for 169 instances or 60% of all incidents with a known perpetrator in Iraq. The resulting dataset has 1.5% of missing values, with 28.6% of the features and 9.9% of the instances missing at least one value.

Once the data was cleansed, the four methods for treating missing values and the three discretization techniques were used to compute the ten-fold cross-validation estimates of the misclassification error for the LDA, KNN and RPART classifiers. All computations were run using the R functions available at www.math.uprm.edu. The results are shown in Table II below.

TABLE II. CROSS VALIDATION ERROR RATES FOR IRAQ DATASET

Missing Data	*Discretization*	*LDA*	*KNN*	*RPART*
CD	None	45.03%	38.41%	36.49%
CD	1R	41.72%	38.41%	29.14%
CD	Entropy	Error [a]	Error [a]	Error [a]
CD	EW	44.37%	38.28%	35.70%
MI	None	43.96%	40.24%	36.69%
MI	1R	42.90%	37.63%	28.70%
MI	Entropy	Error [a]	Error [a]	Error [a]
MI	EW	40.95%	38.88%	37.75%
MDI	None	43.55%	40.24%	36.15%
MDI	1R	44.08%	37.69%	28.11%
MDI	Entropy	Error [a]	Error [a]	Error [a]
MDI	EW	41.12%	37.75%	35.15%
KNNI	None	44.02%	40.24%	36.57%
KNNI	1R	44.73%	38.76%	28.99%
KNNI	Entropy	Error [a]	Error [a]	Error [a]
KNNI	EW	41.60%	38.46%	36.75%

a. Entropy is useless as it partitions variables into one bin.

As an additional test, we added the Global Positioning System (GPS) coordinates (latitude and longitude) for the location of the incidents to the dataset and ran the tests a second time. The results are shown in Table III below.

TABLE III. CROSS VALIDATION ERROR RATES WITH GPS DATA

Missing Data	*Discretization*	*LDA*	*KNN*	*RPART*
CD	None	39.30%	34.88%	32.40%
CD	1R	37.98%	34.88%	35.66%
CD	Entropy	Error [a]	Error [a]	Error [a]
CD	EW	40.54%	34.88%	31.86%
MI	None	39.82%	39.88%	31.79%
MI	1R	42.32%	34.35%	23.51%
MI	Entropy	Error [a]	Error [a]	Error [a]
MI	EW	36.07%	32.68%	31.49%
MDI	None	38.57%	39.88%	33.75%
MDI	1R	39.76%	37.08%	23.81%
MDI	Entropy	Error [a]	Error [a]	Error [a]
MDI	EW	33.39%	34.40%	35.12%
KNNI	None	42.20%	39.88%	36.79%
KNNI	1R	43.81%	38.15%	29.76%
KNNI	Entropy	Error [a]	Error [a]	Error [a]
KNNI	EW	39.05%	38.57%	34.64%

a. Entropy is useless as it partitions variables into one bin.

VI. CONCLUSIONS AND DISCUSSION

From Table II, we can conclude that the lowest classification error rate (28.11%) is obtained using MDI to treat missing values, 1R discretization and the RPART classifier. RPART provides the lowest classification error rate in all tests because it provides various advantages for working with this type of data. Among other things, it has the ability to work with any type of predictive variable, works well with missing data, makes automatic selection of variables, is unaffected by transformations of the predictive variables, and

is robust in the presence of outliers. RPART is a non-parametric classifier (i.e., it does not require assumptions) and takes advantage of the relationships that may exist among the predictive variables. We can therefore conclude that RPART is a better classifier than LDA and KNN for this problem, given the characteristics of the dataset.

We can also conclude that 1R is a better discretization method than Entropy and EW for this problem. 1R produces the lowest classification error rate in all tests, which may be explained by the fact that it is a supervised discretization method (i.e., it uses class information when selecting discretization cut points). EW discretization, on the other hand, cannot be used reliably on this classification problem. Although it reduced classification error rates for the LDA and KNN classifiers, it actually increased the error rate for the RPART classifier 50% of the time. This result may be explained by the fact that EW is an unsupervised discretization method which allows outliers to dominate the presentation and does not handle skewed data well. Entropy discretization is not useful for this problem because it partitions variables in this dataset into one bin, making them useless for classification purposes.

None of the methods used to treat missing values consistently reduced classification error rates by themselves. This erratic behavior may be explained by the fact that the amount of missing values was only 1.5% of the dataset and was concentrated in two variables. Nevertheless, the MDI methodology provides a 1% improvement over CD in the classification error rate when used with 1R discretization and the RPART classifier.

From Table III we can conclude that adding GPS coordinates for the location of the incidents reduces the classification error rate by another 4.6%. This result may be partly explained by the fact that some cities are not always entered into the database with the same name. The GPS coordinates allow classifiers to find stronger relationships between the predictive variables. These relationships are best illustrated by the RPART classification tree shown in Figure 6 below.

As may be seen from the tree, the year, number of wounded, month and latitude of each event are the key variables used by RPART to classify a terrorist group. Different groups are active in spurts at particular times and regions, with varying levels of violence characteristic of their group.

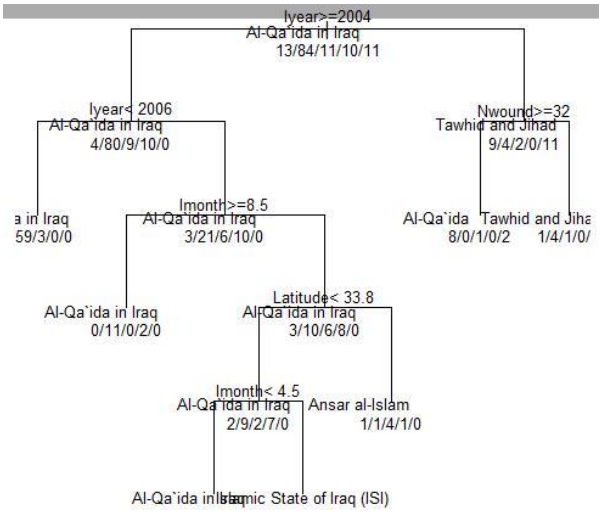


Figure 6. RPART Classification Tree for Iraq GPS Dataset

None of the top five groups identified in Iraq shows activity from 1998 to 2002. They all became active after the U.S. invasion in 2003. Based on these findings, it appears that terrorism in Iraq is the result of a broad and growing insurgency whose extremist faction increasingly resorts to violence.

We strongly recommend that the Global Terrorism Database include GPS coordinates in the future to facilitate the classification of terrorist groups.

REFERENCES

[1] P. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining, Pearson /Addison Wesley, 2006, p.19.

[2] C.E. Stout, The Psychology of Terrorism: Theoretical understandings and perspectives, Greenwood Publishing Group, 2002.

[3] M. Ranstorp, Mapping Terrorism Research: State of the Art, Gaps and Future Direction, 2006, chap. 1, National Defense College, Sweden.

[4] M. Crenshaw, "The causes of terrorism", Comparative Politics, 1981, vol. 13, No. 4, pp. 379-399.

[5] R.M. Williams, Jr., "The sociology of ethnic conflicts: comparative international perspectives", Annual Review of Sociology, vol. 20, 1994, pp. 49-70.

[6] M. Keet, "Sources of terrorism", sourced on May 6, 2010 from http://www.meteck.org/causesTerrorism.html.

[7] D. Rothkopf, Superclass: The Global Power Elite and the World They Are Making, 2009, Farrar, Straus and Giroux Publishing.

[8] J. Lee, "Exploring global terrorism data: a web-based visualization of temporal data", 2008, University of Maryland, College Park.

[9] Global Terrorism Database, START, sourced on April 29, 2010 from http://www.start.umd.edu/gtd/contact.

[10] W. Enders and T. Sandler, "Transnational terrorism in the post-cold war era", International Studies Quarterly, 1999, vol. 43, pp. 145-167.

[11] W. Enders and T. Sandler, "Patterns of transnational terrorism, 1970-99: alternative time series estimates", International Studies Quarterly, 2002, vol. 46, pp. 145-167.

[12] S. O'Brien, "Foreign policy crises and the resort to terrorism: a time-series analysis of conflict linkages", The Journal of Conflict Resolution, 1996, vol. 40, pp. 320-335.

[13] J.A. Major, “Advanced techniques for modeling terrorism risk”, Journal of Risk Finance, 2002, vol. 4, pp.15-24.

[14] T. Sandler and D. Arce, “Terrorism and game theory”' Simulation and Gaming, 2003, vol. 34, pp. 319-337.

[15] D. Guo, K. Liao, and M. Morgan, “Visualizing patterns in a global terrorism incident database”, 2006, Department of Geography, University of South Carolina, Columbia, SC

[16] A. Vilanova, A. Telea, G. Scheuermann, and T. Möller, “Investigative visual analysis of global terrorism”, 2008, Guest Editors, Volume 27, Number 3.

[17] C. Ziemkiewicz, et al., “Global terrorism visualization”, Southeast Visualization and Analytics Center (SRVAC), University of North Carolina at Charlotte, 2008.

[18] National Consortium for the Study of Terrorism and Responses to Terrorism (START), “Terrorist organization profiles,” sourced on April 29, 2010 from http://www.start.umd.edu/start/data/tops.

[19] A. Duhamel, M.C. Nuttens, P. Devos, M. Picavet, and R.A. Beuscart, “A preprocessing method for improving data mining techniques. application to a large medical diabetes databases”, Studies Health Technology and Informatics, 2003, vol. 95, pp. 269-74.

[20] I. Bichindaritz, “Data preprocessing”, University of Washington, 2007, sourced on May 6, 2010 from courses.washington.edu/tcss555/tcss555a_3.ppt

[21] L. Chan and O.J. Dunn, “The treatment of missing values in discriminant analysis”, Journal of the American Statistical Association, 1972, vol. 6, pp. 473-477.

[22] J. K. Dixon, “Pattern recognition with partly missing data”, IEEE Transactions on Systems, Man, and Cybernetics, SMC-9, 1979, vol. 10, pp. 617-621.

[23] V. Tresp, R. Neuneier, and S. Ahmad, “Efficient methods for dealing with missing data in supervised learning”, Advances in NIPS, vol. 7.

[24] E. Acuña and C. Rodriguez, “The treatment of missing values and its effect in the Classifier Accuracy,” University of Puerto Rico at Mayaguez, Mayaguez, PR, 2004.

[25] E. Acuña, “Data preprocessing: data reduction-discretization”, University of Puerto Rico, sourced on May 6, 2010 from math.uprm.edu/~edgar/dm8.ppt.