3.8 Using Random Forest Models to Predict Organizational Violence

Using Random Forest Models to Predict Organizational Violence

Burton Levine, Georgiy Bobashev
Research Triangle Institute
Blevine@rti.org, Bobashev@rti.org

Abstract. We present a methodology to access the proclivity of an organization to commit violence against non-government personnel. We fitted a Random Forest model using the Minority at Risk Organizational Behavior (MAROB) dataset. The MAROB data is longitudinal; so, individual observations are not independent. We propose a modification to the standard Random Forest methodology to account for the violation of the independence assumption. We present the results of the model fit, an example of predicting violence for an organization; and finally, we present a summary of the forest in a “meta-tree.”

1.0 INTRODUCTION

We used Random Forest modeling in the context of social theories to predict the probability that an organization will commit violence against non-government security forces or civilians. By leveraging existing social theories to build predictive models our results have more face validity than an approach that is exclusively data driven. Effective implementation of these models could aid intelligence analysts in accessing terrorist threats. This paper is limited to the use of one Random Forest model. Our data source is longitudinal; consequently, implementation of the Random Forest model is complicated by the correlations within the data. We present methodology used to account for this correlation in the defining of nodes within a tree. Finally, we present the results of the model fit.

2.0 SOCIAL THEORIES

In social research, it is often unacceptable to apply modeling to data without a conceptual model. This is prudent for two reasons. First, the modeling procedures can aid in identifying associations between the outcome and causal variables but cannot identify causality. Consequently, we first identified the hypothesized causal relationships and then use the data to corroborate the conceptual model. Second, if we did not limit our analysis to the factors that we believe *a priori* are related to our outcome than we are likely to present spurious results. We limited the set of organizational characteristics we considered in our models to the characteristics that had been identified as relating to violence in the following papers [1-4].

3.0 DATA

We conducted this analysis on The Minorities at Risk Organizational Behavior (MAROB) dataset. MAROB contains information about the characteristics of ethnopolitical organizations that are most likely to employ violence and terrorism in the pursuit of their perceived grievances with local, national, or international authority structures. The data set contained information on 113 groups located in the Middle East and North Africa. The MAROB dataset covers the period between 1980 and 2004. Additional information about MAROB can be found at the Minorities at Risk (MAR) Web site: http://www.cidcm.umd.edu/mar.

4.0 RANDOM FOREST

4.1 Overview

We used *Random Forest* methodology to create large quantities of decision trees [5]. These trees are created by identifying organizational characteristics associated with violence. Then the tees are used to estimate the probability of violence given specific combinations of organizational characteristics. Random Forest methodology provides a mechanism for creating many decision trees using the same data source and the results are averaged over the trees. The two major advantages of the random forest methodology compared to using a single decision tree is a reduction in variability and a mechanism to validate the results.

4.2 Random Forest methodology

The Random Forest model is validated by allocating the sample into training and testing portions. We fitted a Random Forest model to the training portion. Then we quantified the model fit by comparing predictions from the Random Forest model between the training and the testing datasets. Organizations have a 75% chance of being assigned to the training dataset and 25% chance of being assigned to the testing dataset. Eighty-six of the 113 organizations were assigned to the training dataset.

Using the training dataset, 100 datasets were created by sampling 86 organizations with replacement. A decision tree was created for each of the 100 datasets. The name *Random Forest* is descriptive of the methodology since many decision trees are created by randomly sampling a dataset. Each decision tree partitions the dataset. A partition of a set was a division into non-overlapping and non-empty parts that cover the entire set. These subsets were both mutually exclusive and exhaustive of the set being partitioned. And, each partition had an associated predicted probability of violence. Consequently, for each tree, an observation on our data set (organization by year combination) fitted into one and only one leaf with an associated probability of violence based on the characteristics of the organization in that year.

4.3 Growing a tree

We limited our search for characteristics to use as nodes for *growing* each decision tree to variables that were identified in our literature review of social theories describing violence in the MAROB dataset. All of the independent variables were either binary or made into binary variables by creating indicator variables for organizational characteristics that were multinomial.

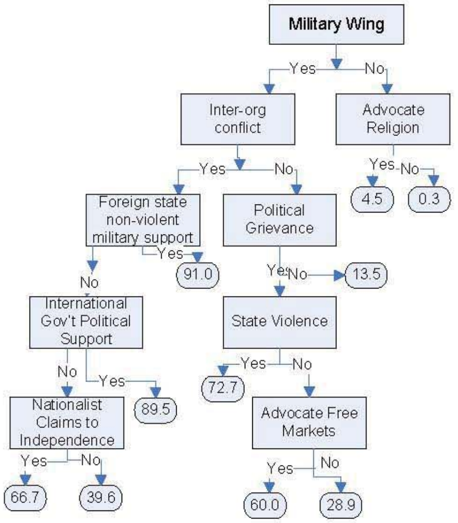
To create a tree from one data set we implemented the following procedure. First, for each variable considered, we fitted a generalized estimating equation (GEE) using the variable in question as the independent variable and next year violence as the dependent variable. Since the outcome is binary we used a logistic regression model: $\log\left(\frac{E\left[Y_{ij}\right]}{\left(1-E\left[Y_{ij}\right]\right)}\right)=x'_{ij}\cdot$

The following expression for the variance incorporates clustering: $V(\alpha)=\varphi A^{1/2}R(\alpha)A^{1/2}$ where $A(\mu)$ is a diagonal matrix of individual variances and φ is the scale parameter. We considered two type of correlations matrixes: exchangeable and autoregressive. We chose autoregressive because we believed that correlation of data within an organization is stronger for observations that are closer in time.

We determined which of the organizational characteristics was most strongly associated with violence by picking the characteristic that contained the lowest p-value from the GEE models. This characteristic is the first node. For the second node we divided the dataset into two groups based on the characteristics of the variable in the first node and calculate the characteristic that is most strongly related to violence on that subset of the data. Repeating this procedure for nodes 3 through 5 we identified the characteristic that was most strongly associated with violence on each subset of the data. We terminated this process when there were less than 20 observations in a node of the tree.

The following is an example of one of the decision trees in the Random Forest.

Figure 1: Example of a decision tree

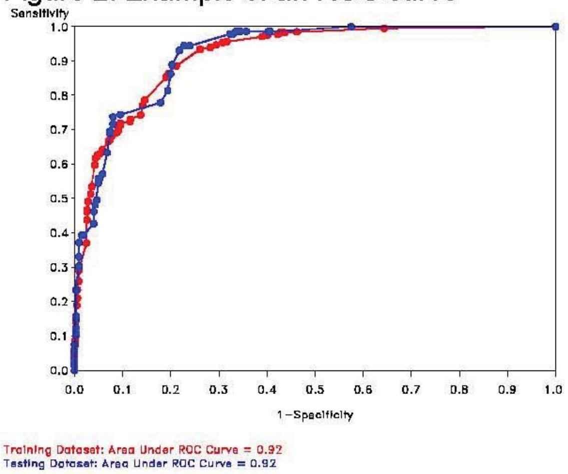


The tree above has 10 leafs. To assign a probability of violence to each leaf (the number in the ovals in the tree above) we created a variable that identified the leaf for every organization by year combination. Then we ran a GEE model with the "leaf" variable as the lone independent variable and with violence next year as the dependent variable.

The decision tree created a partition of the dataset such that each data point (each organization by year combination) was located in one and only one of the leaves of the decision tree. And, each leaf had an associated probability of violence. We used each tree to associate a probability of violence for each observation on the dataset based on the characteristics that made up the leaves of the tree. So, for each observation we generated 100 predictions, one for each tree, of the probability of violence in the next year. We then averaged over the 100 values to generate one prediction of violence. The trees were created using the training dataset exclusively. But, we generated predictions of violence for both the training and the testing dataset. Finally, we tested the quality of the predictions by generating ROC curves for the training and testing datasets as described in [6].

Below are graphs of the ROC curves. The area under the ROC curve for the training dataset is 0.92. This indicates that the model fits the data very well. The area under the ROC curve for the testing dataset is also 0.92. This indicates that the model is applicable to data that was not used to generate the predictions. In other words, we have obtained evidence that the model is generalizable.

Figure 2: Example of an ROC curve



4.4 Meta-tree

Using Random Forest methodology to predict the probability of violence has a beneficial characteristic of being robust. By sampling a large number of datasets the potentially deleterious effect of aberrant results is minimized. However, the prediction rule is very complicated since it requires a large amount of trees, 100 in our analysis. Consequently, in its present form Random Forests are not useful for identifying the combinations of characteristics that are related to violence. But, we can use the 100 trees to create a *meta-tree* that is useful in identifying the interactions that are most common to the 100 trees. The following table displays the distribution of the first node for the 100 trees.

Node	Frequency
Military Wing	86
Inter-organizational conflict	10
Other	4

In our meta-tree we have selected Military Wing as the first node.

The following tables display the distribution of the second node for the 86 trees that contained military wing in the first node.

First node: No Military Wing	
Second node	Frequency
Foreign state non-military support	34
Dominant economic grievance	34
Other	18

First node: Military Wing	
Second node	Frequency
Inter-organizational conflict	50
State violence	16
Other	20

In our meta-tree, for the second node under no military wing, we have the same number of trees using the nodes foreign state non-military support and dominant economic grievance. Both of these nodes lead to terminal leaves. So, we considered both possibilities for our meta-tree. We selected inter-organization conflict as our second node under military wing.

The follow table displays the distribution of the third node for the 50 trees that contained "has military wing" in the first node and inter-organizational conflict in the second node.

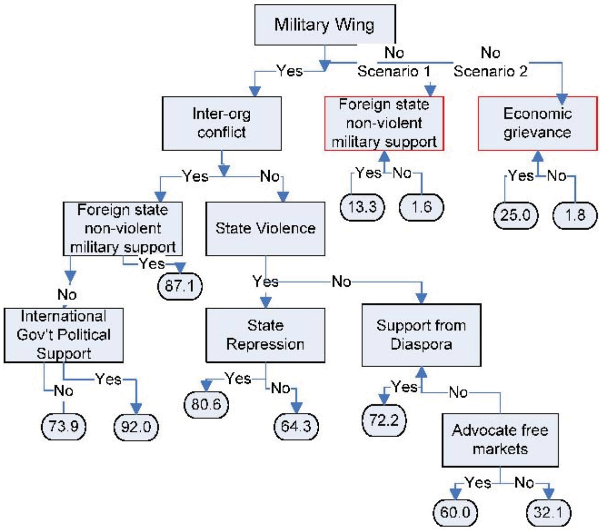
First node: Military Wing Second node: Inter-org conflict	
Node	Frequency
Foreign state non-military support	17
Received support from diaspora	13
Other	20

First node: Military Wing Second node: No inter-org conflict	
Node	Frequency
State violence	33
Other	17

We chose "foreign state non-military support" for the node under (Military wing=yes) and (Inter-organizational conflict=yes). And, chose "State violence" for the node under (Military wing=yes) and (Inter-organizational conflict=no).

Continuing the process we arrived at the following meta-tree. The boxes in red are two equally likely scenarios.

Figure 3: Example of a meta-tree



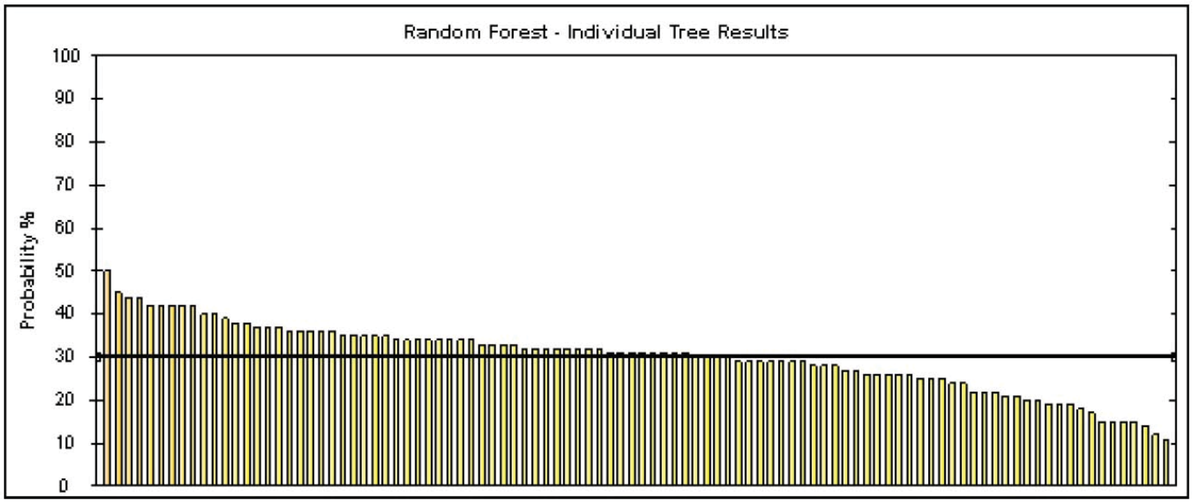
4.5 Making predictions

To use the Random Forest model to make predictions on an organization we need to have data on that organization. In the following table, the left column contains the nodes used in all 100 trees. The right column contains the value for the organization of interest. Rather than concoct a fictitious example, we used the MAROB data for the organization Fatah al-Intifada in 2004. We made predictions on whether this organization will commit violence on non-government personnel during 2005.

Criterion	Fatah al-Intifada 2004
Advocate free market policies	0
Dominant political grievance focused on creating separate state	1
Economic grievance focus on elimination of discrimination	0
Economic grievance of organization-remedial policies	0
Ethnic organization no claims of independence	1
Foreign state non-violent support	0
Foreign state support in current year	0
Has military wing	1
Inter organizational conflict	0
Inter organizational conflict severity	0
International government political support	0
International government organizational support	0
International non-government organizational support	0
Intra-organizational conflict	0
Organization is legal	1
Nationalist claims of independence	0
Organization advocates policies concerning religion	0
Organization open	1
Received support from diaspora	0
Repression by state	0
State violence against organization	0

To calculate this probability, for each of the 100 trees, we found the node associated with the organizational characteristics of Fatah al-Intifada in 2004 and the value of the probability of violence in that node. Then we averaged the probabilities for such nodes over all trees. It turned out that the probability of violence in the next year was 30%. A bar graph in Figure 4 presents the results of the prediction with the Random Forest model. Each vertical bar represents the prediction from one tree and the horizontal line is the average value over all 100 trees. An analyst can also explore individual tree models by examining the groups that form trees at either sides of the graph. For example, on the left hand side there are trees that produce high probability of violence (almost 50%). On the right end of the graph the trees contain groups that have lower probabilities (around 10%). Because each prediction is based on groups that share the same characteristics of the node with Fatah al-Intifada, one might want to take a closer examination of these groups' characteristics to think of possible scenarios of violent or not violent solution of the situation.

Figure 4. Bar graph of predicted probabilities for trees in the Random Forest predictive model



5.0 Discussion

Although it is not possible to predict the future, it is often possible to assess risks and probabilities of certain events to happen. All these forecasts and predictive probabilities carry a load of uncertainty error. Most of these errors we cannot control because model is always a simplification of the reality, however Random Forest methodology allows one to assess the within-model error, i.e. assuming that the model is correct. As most forecasting techniques this model is based on similarity rationale, i.e. similar (in some sense) groups would behave in a similar way. No two groups are exactly the same and a selected group could be similar to a variety of groups in different ways e.g., size, political goals, religious views, legal status, etc. The advantage of our Random Forest approach is that it examines a broad variety of similar groups, e.g. groups that are similar in many characteristics but are either very violent or not violent at all. Focusing on these opposite sides of the spectrum allows an analyst to assess what *additional* characteristics or government actions contribute to similar groups shift towards or away from violence.

In the intelligence analyst setting, assigning precise estimates are usually avoided. As a result, when these models are implemented the probability estimates will be mapped into the following categories: high, medium and low probability of violence.

The implementation of this methodology has several limitations. The results will only be valid for organizations that will fit the MAROB inclusion criteria. These inclusion criteria are described in the MAROB website. Furthermore, one must know a lot of information about the organizational characteristics of the group one is interested in to apply this model. If the user of the model is unsure of some of the characteristics they can either take their best guess or input all combinations of the missing data. Each combination will result in a different predicted probability of violence. Using these predicted probabilities one can create a range of predicted probabilities.

6.0 References

[1] Asal, V. (2007, August). *Organizational factors and the choice of terrorism: Minorities at Risk Organizational Behavior in the Middle East*. Paper presented at the annual meeting of the American Political Science Association, Chicago, IL.

[2] Post, J., Ruby, K., & Shaw, E. (2002a). The radical group in context: 1. An integrated framework for the analysis of group risk for terrorism. *Studies in Conflict and Terrorism*, *25*, 73–100.

[3] Post, J., Ruby, K., & Shaw, E. (2002b). The radical group in context: 2. Identification of critical elements in the analysis of risk for terrorism by radical group type. *Studies in Conflict and Terrorism*, *25*, 101–126.

[4] McCauley, C. (2009). The 21-item threat criterion list: Review and suggestions. In M. Schwerin, *Violent Intent Modeling and Simulation (VIMS) year-end report*. Prepared for the Human Factors/Behavioral Sciences Division, Science and Technology Directorate, U.S. Department of Homeland Security. Research Triangle Park, NC: RTI International. Appendix E.

[5] Breiman,L.(2001). Random Forests, Machine Learning, 45, 5-32. http://www.springerlink.com/content/u0p06167n6173512/fulltext.pdf.

[6] Zweig, M.H., Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine, Clinical Chemistry, 39: 561-577.

7.0 Acknowledgements

This work was supported by the VIMS contract with the Human Factors/Behavioral Sciences Division, Science and Technology Directorate, U.S. Department of Homeland Security.

The corresponding author is Burton Levine, (919) 541-1252; E-mail: blevine@rti.org; Contract Number HSHQDC-08-C-00100. The views expressed in this article are our own and do not reflect the official policy or position of the Department of Homeland Security or the U.S. Government. The study is approved for public release for unlimited distribution.