CPS 430/590.06 **Prof. John Reif**
Design and Analysis of Algorithms **Fall 2013**
Final Project **Matt Dickenson**

Machine Learning Algorithms with Application to Political Event Data

Abstract How do computational and statistical efficiency interact, and what are the tradeoffs between them? In the era of "big data" this question has become increasingly relevant. This project explores that question in the context of a particular machine learning algorithm (classification trees) and a particular problem (automated processing of political indicators). Applying machine learning to the classification of political event data can greatly reduce the cost in human effort, time, and money. The motivation for this project is to update the Militarized Interstate Disputes (MID) dataset, which has been widely used in academic research and policy discussions. The MID project relies on humans reading journalistic accounts and manually entering the classification of the event according to a defined schema. This dependence on humans is both less accurate, less efficient, and more expensive than automated methods. The results of this analysis suggest that automated methods can provide a first-pass classification of political indicators at a huge savings of time and money, without sacrificing accuracy or interpretability.

1 Introduction

Can computational methods detect political and social upheaval through automated text processing and machine learning? If so, can this process be done with both statistical and computational efficiency? This project seeks to answer these questions in one particular application area: international conflicts. In this section, we provide context for the relevance of this application and the methods under discussion. Section 2 discusses existing methods for predicting disputes, surveys the extent of machine learning within political science, and explains why classification trees are an appropriate method for this problem. Section 3 analyzes the computational complexity of classification trees in general, and Section 4 describes the implementation of classification trees used in this project. Section 5 presents the statistical results and compares their efficiency to generalized linear models. Section 6 concludes the paper with implications for future research.

2 Background and Motivation

2.1 Production of MID Data

The current production of the Militarized Interstate Disputes (MID) dataset is overseen by the Correlates of War Project, headquartered at Penn State University. The most recent version of the data set, MID 3.0, was released in 2003 with coverage through 2001. An update was expected during the summer of 2012, but has been delayed indefinitely. The time-intensive nature of the current MIDS production pipeline inevitably delays important scholarly research on the nature of interstate conflict.

A "militarized incident is defined as a single military action involving an explicit threat, display, or use of force by one system member state towards another system member state." These incidents form the basis of a dispute in the data set. The current "algorithm" for producing MID data updates is as follows (Ghosn et al., 2004):

1. Search the Lexis-Nexis database for news articles mentioning possible militarized incidents. An event classifies as a militarized interstate incident if it:

    - occurs among two or more interstate system members;
    - is an overt action taken by the official military forces or government representatives of a state; and
    - involves the threat, display, or use of forces.

2. Human coders assigned to a specific geographical region read news articles and list all possible militarized incidents.

3. Principal invesigators classify incidents as "OK" or "under revision." More precise information (e.g. exact start date of a conflict) is sought for incidents in the latter category.

4. Events from different regions are aggregated into the final dataset.

For the most recent update, this process took about three years. By contrast, a single run of a classification tree like the one used in the paper takes less than ten minutes.

Decision trees can help to imitate this human thought process, but with obvious advantages in scale and speed (King and Lowe, 2003; Mikhaylov et al., 2012; Ruggeri et al., 2011). Furthermore, automated methods follow their decision rules perfectly, without falling prey to accidental errors of human classifiers. These trees also maintain interpretability relative to other "black-box" machine learning methods, meaning that a well-defined tree could be incorporated into a production pipeline that also includes human classifiers. For all of these reasons, classification trees are worth exploring in this project.

2.2 Previous Research

As one of the most widely used dependent variables in international conflict studies, much effort has been devoted to estimating models of MID onset and duration. However, this work

suffers from several common weaknesses that this project attempts to ameliorate: virtually all projects, especially before the present decade, used a fixed functional form (typically from the family of generalized linear models); out-of-sample testing and cross-validation is used only rarely, making claims of 'prediction' somewhat dubious in many cases; and often the independent variables are measured at the annual level with high levels of serial correlation, meaning that there is little temporal variation in the predictors, while the dependent variable tends to exhibit more sudden onsets (Ward et al., 2010). A recent shift toward event data has helped to address the latter two of these issues: with frequent updates (often measured at the daily level), there is substantial variation in the independent variables, validation requires only a brief waiting period for new sets of test data (Gerner et al., 1994, 2002; King and Lowe, 2003; Ruggeri et al., 2011; Schrodt and Leetaru, 2013).

With this transition toward event data as predictors, the political forecasting community has become attune to new challenges and has responded with several established practices. Coding the sentiment of interactions can now be done in near real-time (NRT) using the Tabari system, which aggregates and deduplicates news reports (Obrien, 2010; Schrodt, 2009). Sentiment coding can be done according to two widely used systems. The Goldstein scale assigns a score of -10 (highly conflictual) to +10 (highly cooperative) to events, but it is difficult to employ this scale for aggregations or permutations of the data (Goldstein, 1992). CAMEO classifies events into a pre-defined schema of material/verbal and cooperative/conflictual actions, that makes aggregation simpler because we can count events within each category (Gerner et al., 2002). These event classifications provide a principled, automated method for exhaustively categorizing the types of events that may consitute an interstate dispute (Ghosn et al., 2004).

The community has also dealt with challenges when aggregating event data up to various temporal levels. Although there is no single best practice, monthly aggregation has become a common strategy (Arva et al., 2013; Yonamine, 2013) and is used in this project. Modfiying the features by transforming the raw counts into month-to-month changes (i.e. first-differencing) and measuring the balance between conflictual and cooperative interactions

as a percentage of the total also helped to simplify the feature set (Box and Jenkins, 1976).

Interpretability is an important concern in this project due to the policy-relevant nature of the problem and the (potential) need to compare the resulting model to the process used by human coders involved in creating the MID dataset (Ghosn et al., 2004). For this reason, "black box" methods such as Support Vector Machines were judged to be inappropriate. Classification trees (and their continuous counterpart, regression trees, collectively known as CART) offer a nice alternative that is more flexible than GLMs and more interpretable than Random Forests (these two methods should provide lower and upper bounds, respectively, on CART) (Klebanov et al., 2008). CART has been used for event data within conflict studies, and in public health where researchers encounter similar issues of unbalanced and missing data (Schrodt, 1990; Speybroeck, 2012; Trappl et al., 1996).

In later stages of this project, several additional tools may help to improve the predictive accuracy of the model. International conflict is a relatively rare event, meaning that in k-fold cross validation it is possible that some subsets will have no instances of conflict; to prevent this, synthetic minority over-sampling (SMOTE) could be used (Chawla et al., 2002). To incorporate interdependencies not captured at the dyadic level, future iterations could also include lags that measure conflict in social or spatial neighbors (Gleditsch and Ward, 2000, 2001; Hoff and Ward, 2004; Ward and Gleditsch, 1998; Ward et al., 2007, 2011). A Bayesian ensemble model of several classification trees could also improve performance while still maintaining more interpretability than is available in random forests (Arva et al., 2013; Montgomery et al., 2012; Raftery, 1995; Raftery et al., 2005). If these methods are successful, the general processing of automating political indicators through the use of event data could also be applied to other widely used indices such as the Polity and Freedom House regime scales (measuring democracy and autocracy).

3 Computational Complexity of Classification Trees

A classification tree provides a set of decision nodes used to assign observations to one or more categories based on features of the observation. We wish to build a tree that minimizes classification error. The problem of finding the optimal tree is NP-complete (Mohri et al., 2012). However, we can use a greedy algorithm to grow the tree iteratively, starting at the root.

A schematic tree consists of a node that *splits* the data based on a *test*. Values satisfying the test are classified into the left subtree, while values failing the test are assigned to the right subtree. Once an observation reaches a leaf node, it is classified probabilistically based on the preponderance of observations assigned to that leaf.

[Figure 1 about here.]

CART greedily grows a tree T from a root node by adding splits based on features. At each stage, there is an impurity function $I(T)$ that measures the (in)accuracy of the existing classifications. When adding another split to the tree, the CART algorithm seeks to maximize the impurity reduction ΔI. The measure of impurity reduction is supplied to the algorithm, as is a complexity parameter, κ. We can think of κ as the cost of adding another variable to the tree. When $\kappa = \infty$ the tree will consist of only the root node, and as we reduce κ toward zero we allow the tree to grow. However, setting κ too low can result in overfitting, which can be assessed via cross-validation. For more on classification trees in general and the implementation used for this project, see (Murphy, 2012; Olshen and Stone, 1984; Therneau and Atkinson, Therneau and Atkinson). For n observations, classification is at least as complex as sorting, giving us a lower bound of $\Theta(n \log n)$. Careful implementation with greedy algorithms can reach $\Omega(n \log n)$ (Latkowski, 2003).

Given a set of features H, a cost function $I(T)$, and a minimum complexity κ, the greedy top-down tree-building algorithm at is as follows:

1. Initialize the empty tree, T_0

2. Until $\Delta I \leq kappa$:

   (a) Given T_{i-1}, choose the leaf l and feature test $h \in H$ that maximizes ΔI

   (b) Replace leaf l in T_i with the new subtree created by test h

Pruning is a bottom up procedure that replaces nodes with leaves, minimizes the error in another data set (not the one used to train the tree). This second data set can be an entirely new test set or a bootstrapped version of the original training set. Pruning is generally exponential, but this can be reduced using dyanmic programming. Starting with the errors at each leave, we prune backwards until we reach a threshold (provided as input) for the maximum number of errors, e. This gives us the simplest tree possible that makes no more than e errors. For data input of size n and tree size $|T|$, greedy pruning can be done in $O(n^2|T|)$. Having seen how classification trees perform in theory and their computational efficiency, we now turn to an applied case to examine their statistical efficiency and accuracy.

4 Applied Analyis

4.1 Problem Definition and Data Sources

The classification tree attempts to categorize country dyad months (e.g. `USA-China-2012-May`) as either in conflict or not. To achieve this, we use real time (daily) event data from the Global Database of Events, Language, and Tone (GDELT), aggregated up to the dyad month level for 1992-present (Schrodt and Leetaru, 2013). To measure the dependent variable of conflict, the Militarized Interstate Disputes (MID) dataset will be split into subsets for training and validation (Ghosn et al., 2004). The goal of this project is to replicate and extend MID data coding as accurately as possible using automated procedures. If a reliable method can be developed to replicate the MID data up to 2001, it can then be extended to generate data for interstate disputes since 2001.

In work on this project thus far, several important features of the GDELT data have been identified. All events in GDELT are classified according to the CAMEO coding scheme

(Gerner et al., 2002). Within this scheme, there are two major distinctions along two dimensions: acts can be material or verbal, and interactions can be cooperative or conflictual. These four categories provide a rough characterization of how two countries interact within a given period of time. More fine-grain classification, into twenty subcategories, is also provided. Verbal cooperative events include public statements, appeals, expressions of intent to cooperate, consultations, and engaging in diplomatic cooperation. Materiale cooperation includes providing aid, yielding, and investigations. Verbal conflict includes demands, disapproval, rejections of offers, threats, and protests. Material conflict includes exhibiting force posture, reducing relations, coercion, assaults, fights, and the use of unconventional mass violence. These categories are mutually exclusive and exhaustive.

During the process of aggregating GDELT records into dyad months, the absolute number of events within each of the four major and twenty minor categories was counted. From these raw counts, the monthly change in counts and percentages, as well as the relative frequency of each interaction type was computed. These features–proportion of interactions that were conflictual versus cooperative, and how sharply events changed from the previous month–will be used as predictors for the classification procedure.

MID hostilities are measured on a five-point scale, which was collapsed into a binary with the cutoff set at four. Events above this threshold involve the use of force, typically associated with "war," while events below this threshold require only the threat or display of force (Ghosn et al., 2004). Further classification of exact hostility levels will be attempted in a later stage of this project.

4.2 Model

The conflict indicators $y_{i,j,t}$ are binary $(0, 1)$. The observations $x_{i,j,t}$ consist of the month-to-month change in interactions between i and j within each of the event categories described above ($\Delta x_{i,j,t} = x_{i,j,t} - x_{i,j,t-1}$). Thus, x is a count variable that can take on positive or negative values ($x \in \mathbb{Z}$). The observations $z_{i,j,t}$ measure the relative frequency of conflictual

interactions as a percentage of the total n observations for the dyad-month:

The predictor values are observed in the GDELT data. The indicator of conflict, y, is observed in the MID data, and predicted indicators of conflict $\hat{y}$ will be estimated. The predicted values $\hat{y}$ for the test set can be compared to the actual MID data to assess how well the model works out-of-sample. This will give us a sense of how accurate the classifications for post-2001 data will be. Even though these values will not be perfectly accurate, they should give us a good approximation of which countries experienced conflict since 2001 and can help speed up the production of the next generation of MID data. A classification tree was fit to this data, with results presented in the next section.

5 Results

To fit the classification tree, this project used the `rpart` library in $\mathcal{R}$ (Therneau and Atkinson, Therneau and Atkinson). The minimum complexity parameter, κ, was initialized to be 10^{-4}. By cross-validation, the optimal κ was found to be 0.00155 (see Figure 2). This value was used to prune the tree from its maximum of 176 splits down to nine. The resulting tree is shown in Figure 3, with leaves shaded by whether MID hostilities ("war") or peace is more likely. Each leaf also indicates the relative frequency of war in the training set for observations assigned to that leaf.

[Figure 2 about here.]

[Figure 3 about here.]

5.1 Interpretation

To interpret the tree, we begin at the top (root) node and work downward. The inequality at listed at each node splits the data. If the observation satisfies the inequality, we visit the left subtree; otherwise, we visit the right subtree. We continue recursively until we reach a leaf, which indicates whether that dyad month should be classified as peaceful (green) or at

war (red). For example, an observation with less than 14 investigations in that dyad-month would cause us to trace the left subtree of the root note, and classify that observation as peaceful (war occurs in less than one percent of such observations). By contrast, a month with between 34 and 176 investigations, more than 28 statements, and at least 32 acts of force has about a 77 percent chance of being at war in the training data.

All of the twenty event types described above were supplied as candidate variables to the model, as were the percentage of observations within each quadrant (i.e. material cooperation, verbal cooperation, material conflict, and verbal conflict). Of these, seven variables are included in the final tree. Investigations feature prominently in the first three splits. At first this seemed curious, given that investigations are classified as material cooperation by CAMEO (compare node eight). Upon closer examination, this category includes investigations of crime, corruption, human rights abuses, military action, and war crimes. Given that humanitarian interventions feature prominently in the MID dataset during the period under consideration, the influence of this variable becomes less surprising. For example, NATO efforts in the Balkans involve a large number of dyads involved in the investigation of war crimes and other human rights violations.

The other splits are less surprising. Demands (node four) are likely to be more associated with domestic conflicts, such as protest movements, rather than interstate disputes. Large amounts of aid make conflict less likely (node seven), as would be expected. Unconventional violence (e.g. mass killings and ethnic cleansing) and the use of force (e.g. fighting with small arms, artillery, or aircraft) are associated with MIDs for obvious reasons. Overall the model seems to match the types of conflict observed in the 1990s, but might do less well in later periods.

5.2 Model Diagnostics

How does the tree perform relative to other classification models? Table 1 compares the tree above to a null model (all observations are predicted to be peaceful) and a logistic regression

using the same features as the CART model. The data was split into two parts, a training set for fitting the models, and a test set for out-of-sample comparison. This prevents a model that overfits the data from appearing superior to one with more generality.

Mean-squared error is the average squared deviation of predictions from the observed values. It is calculated as $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. The ideal MSE is zero, and in this case the worst possible value is 1. In the case of binary classification, this is equal to the misclassification rate. Precision and recall are particularly useful metrics for rare events, as they penalize the model for making incorrect prediction. The ideal for both is a value of 1. Precision is calculated as the proportion of true positives relative to the number of true and false positives (i.e. $\frac{\#\text{accurate predictions of war}}{\#\text{ total predictions of war}}$). Recall is similar, but for the proportion of true positives relative to the number of true positives and false negatives (i.e. $\frac{\#\text{accurate predictions of war}}{\#\text{ total observed cases of war}}$).

The classification tree outperforms the other two models in the training data (1992-1998, $n = 372,271$). For the test data (1999-2001, $n = 213,218$), the tree has the same mean-squared error (MSE) as the null model, but better precision and recall. Compared to logistic regression, the tree has worse recall but much better precision for the test data. This suggests that the classification tree generated more false negatives (i.e. missed the occurrence of some conflicts) in the test data.

[Table 1 about here.]

6 Conclusion

Classification trees strike a nice balance between computational and statistical efficiency. The computational complexity of greedy tree-building is $\Omega(n \log n)$, which is the best that can be hoped for. Statistically, CART outperforms logisitic regression using a generalized linear model (GLM). In the application here, this is likely because the classification tree handled dependencies between features better than the linear model.

The interpretation of the tree suggests that it fits the 1990s very well, but may perform

less well in the future. Perhaps an ensemble of trees based on a few months of data at a time (with the addition of more trees as the data extend in to the future) would perform better and account for changes in the data generating process over time. A Bayesian ensemble of several classification trees could improve performance while still maintaining more interpretability than is available in random forests (Arva et al., 2013; Montgomery et al., 2012; Raftery, 1995; Raftery et al., 2005).

Other adjustments can also help to refine the model at later stages. International conflict is a relatively rare event, meaning that in k-fold cross validation it is possible that some subsets will have no instances of conflict; to prevent this, synthetic minority over-sampling (SMOTE) could be used (Chawla et al., 2002). To incorporate interdependencies not captured at the dyadic level, future iterations could also include lags that measure conflict in social or spatial neighbors (Gleditsch and Ward, 2000, 2001; Hoff and Ward, 2004; Ward and Gleditsch, 1998; Ward et al., 2007, 2011).

The findings thus far indicate that automated production of political indicators is feasible to within a close approximation. The application of classification trees offers a quick and inexpensive tool that can save much human effort. A final pass by human coders will still be required to obtain fully accurate classifications, but the cost can be substantially reduced. With the additional steps proposed above, the process of automating political indicators using event data could also be applied to other widely used indices such as the Polity and Freedom House regime scales (measuring democracy and autocracy).

References

Arva, B., J. Beieler, B. Fisher, G. Lara, P. A. Schrodt, W. Song, M. Sowell, and S. Stehle (2013). Improving forecasts of international events of interest. In *EPSA 2013 Annual General Conference Paper*, Volume 78.

Box, G. E. P. and G. M. Jenkins (1976). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer (2002). Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research 16*(1), 321–357.

Gerner, D. J., P. A. Schrodt, R. A. Francisco, and J. L. Weddle (1994). The analysis of political events using machine coded data. *International Studies Quarterly 38*(1), 91–119.

Gerner, D. J., P. A. Schrodt, Y. Ömür, and R. Abu-Jabr (2002, August, 29-September 1). Conflict and Mediation Event Observations (CAMEO): A New Event Data Framework for a Post Cold War World. Boston, MA. Annual Meetings of the American Political Science Association.

Ghosn, F., G. Palmer, and S. A. Bremer (2004). The mid3 data set, 1993–2001: Procedures, coding rules, and description. *Conflict Management and Peace Science 21*(2), 133–154.

Gleditsch, K. S. and M. D. Ward (2000). War and peace in space and time: The role of democratization. *International Studies Quarterly 44*(1), 1–29.

Gleditsch, K. S. and M. D. Ward (2001). Measuring space: A minimum-distance database and applications to international studies. *Journal of Peace Research 38*(6), 739–758.

Goldstein, J. S. (1992). A conflict-cooperation scale for weis events data. *Journal of Conflict Resolution 36*(2), 369–385.

Hoff, P. D. and M. D. Ward (2004). Modeling dependencies in international relations networks. *Political Analysis 12*(2), 160–175.

King, G. and W. Lowe (2003). An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design. *International Organization 57*(03), 617–642.

Klebanov, B. B., D. Diermeier, and E. Beigman (2008). Lexical cohesion analysis of political speech. *Political Analysis 16*(4), 447–463.

Latkowski, R. (2003). High computational complexity of the decision tree induction with many missing attribute values. In *Proceedings of CS&P*, pp. 25–27.

Mikhaylov, S., M. Laver, and K. R. Benoit (2012). Coder reliability and misclassification in the human coding of party manifestos. *Political Analysis 20*(1), 78–91.

Mohri, M., A. Rostamizadeh, and A. Talwalkar (2012). *Foundations of machine learning*. The MIT Press.

Montgomery, J. M., F. M. Hollenbach, and M. D. Ward (2012). Improving predictions using ensemble bayesian model averaging. *Political Analysis 20*(3), 271–291.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. The MIT Press.

Olshen, L. B. J. F. R. and C. J. Stone (1984). Classification and regression trees. *Wadsworth International Group*.

Obrien, S. P. (2010). Crisis early warning and decision support: Contemporary approaches and thoughts on future research. *International Studies Review 12*(1), 87–104.

Raftery, A. E. (1995). Bayesian model selection in social research (with discussion). In P. V. Marsden (Ed.), *Sociological Methodology 1995*. Cambridge, MA: Blackwell.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski (2005). Using bayesian model

averaging to calibrate forecast ensembles. *Monthly Weather Review 133*(5), 1155–1174.

Ruggeri, A., T.-I. Gizelis, and H. Dorussen (2011). Events data as bismarck's sausages? intercoder reliability, coders' selection, and data quality. *International Interactions 37*(3), 340–361.

Schrodt, P. and K. Leetaru (2013). Gdelt: Global data on events, location and tone, 1979-2012. *International Studies Association*.

Schrodt, P. A. (1990). Predicting interstate conflict outcomes using a bootstrapped id3 algorithm. *Political Analysis 2*(1), 31–56.

Schrodt, P. A. (2009). Tabari: Textual analysis by augmented replacement instructions, version 0.7.

Speybroeck, N. (2012). Classification and regression trees. *International journal of public health 57*(1), 243–246.

Therneau, T. M. and E. J. Atkinson. An introduction to recursive partitioning using the rpart routines. Technical report.

Trappl, R., J. Fürnkranz, and J. Petrak (1996). Digging for peace: Using machine learning methods for assessing international conflict databases'. In *ECAI*, pp. 453–457. PITMAN.

Ward, M., B. Greenhill, and K. Bakke (2010). The perils of policy by p-value: Predicting civil conflicts. *Journal of Peace Research 47*(4), 363–375.

Ward, M., R. Siverson, and X. Cao (2007). Disputes, democracies, and dependencies: A reexamination of the kantian peace. *American Journal of Political Science 51*(3), 583–601.

Ward, M. D. and K. S. Gleditsch (1998). Democratizing for peace. *American Political Science Review*, 51–61.

Ward, M. D., K. Stovel, and A. Sacks (2011). Network analysis and political science. *Annual Review of Political Science 14*, 245–264.

Yonamine, J. E. (2013). Working with event data: A guide to aggregation choices. *Ph.D. Thesis*.

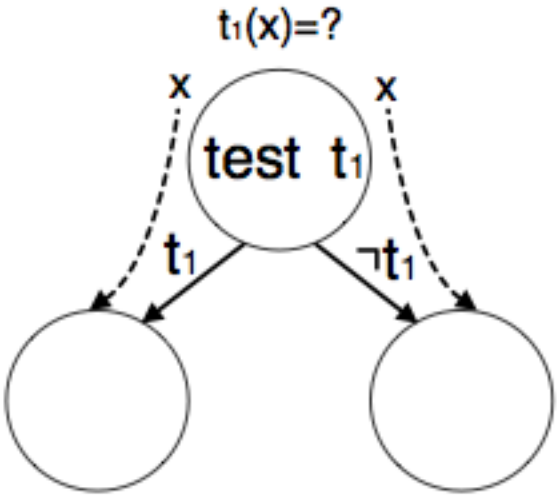

Figure 1: A Schematic of a Classification Tree (adapted from Latkowski, 2003)

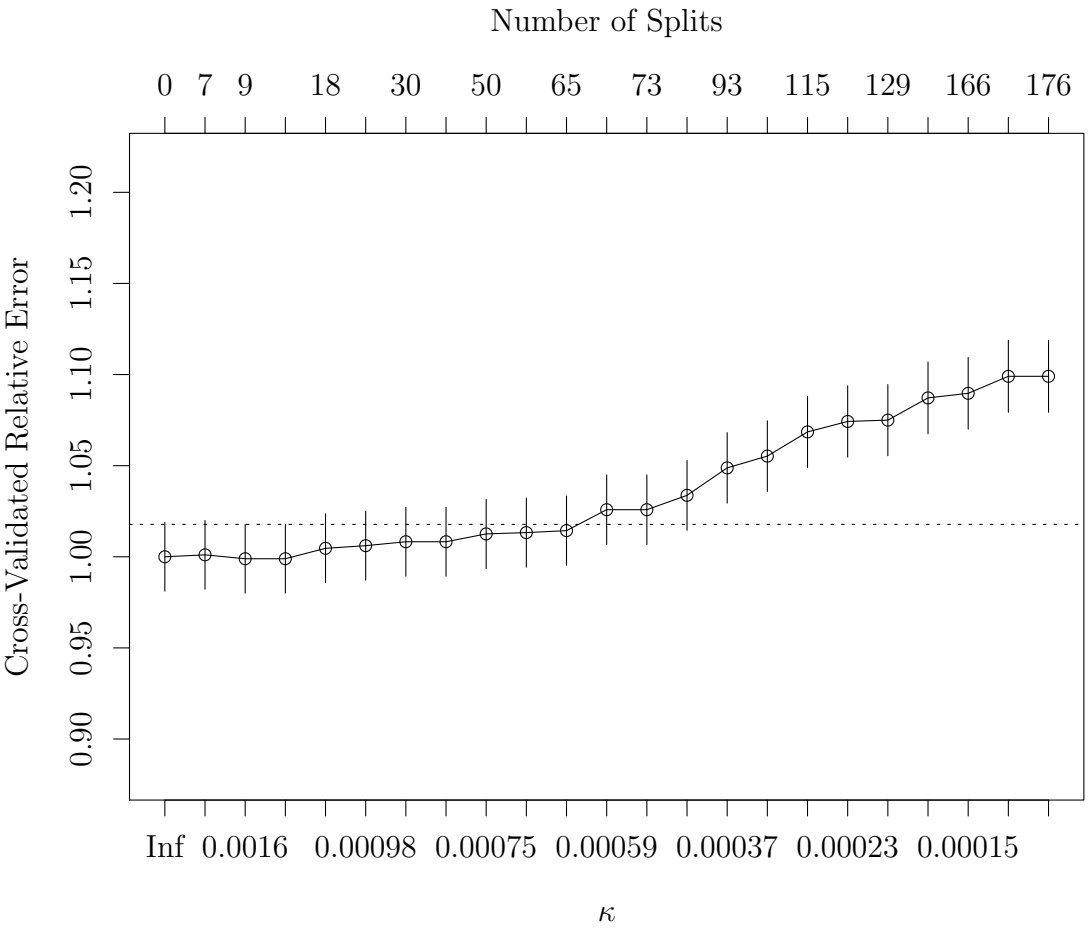

Figure 2: Cross-Validation Error and Number of Splits as a Function of κ

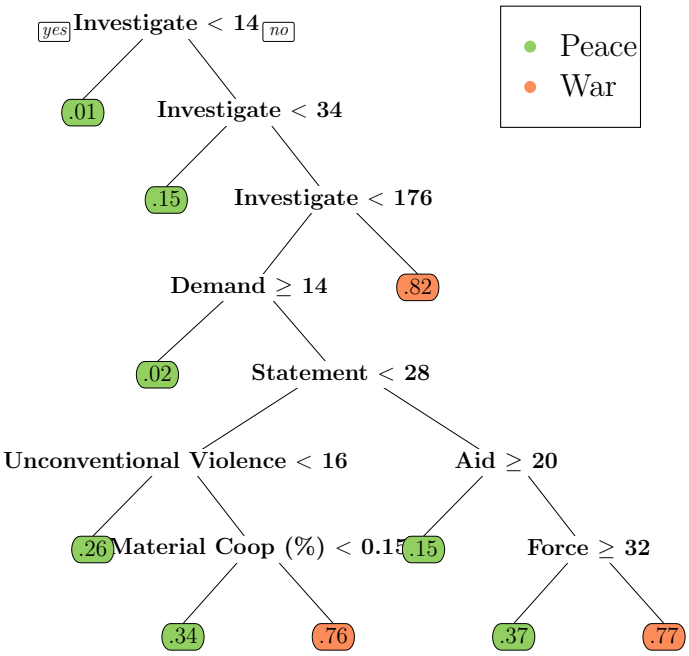

Figure 3: Fitted and Pruned Classification Tree

Table 1: Comparison of CART to Alternative Models

	Training Data			Test Data		
Model	MSE	Precision	Recall	MSE	Precision	Recall
Null	0.0075	0.000	0.000	0.0066	0.000	0.000
GLM	0.0082	0.158	0.022	0.0079	0.142	0.038
CART	0.0067	0.702	0.192	0.0066	0.422	0.027