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 This project will analyze an algorithm and implementation of classification trees and discuss their application to automated processing of political event data. 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 two related methods (generalized linear models and random forests). Section 6 concludes the paper with implications for future research.

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 accu-

racy 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

4 Statistical Analyis

4.1 Problem Definition and Data Sources

The problem that this project attempts to solve is the classification of country dyad months (e.g. `USA-China-2012-May`) as either in conflict or not. To achieve this, we will 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.

Table 1: CAMEO event categories and descriptions

	Cooperative	**Conflictual**
Verbal	public statement, appeal, express intent to cooperate, consult, engage in diplomatic cooperation	demand, disapprove, reject, threaten, protest
Material	cooperate materially, provide aid, yield, investigate	exhibit fore posture, reduce relations, coerce, assault, fight, use conventional mass violence

4.2 Features of the Data

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. Examples of these categories are presented in Table 1.

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.

4.3 Model

The mathematical model for this project is that a binary indicator of conflict, y, between country i and country j at time t is a function of observed interactions between them in month t. Formally,

$$\hat{y}_{i,j,t}|x \quad = \quad f(\Delta x_{i,j,t} + z_{i,j,t})$$

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:

$$z_{i,j,t} = \frac{\sum_{k=1}^{n} x_{i,j,t,k} \mathbb{I}(\text{conflictual})}{\sum_{k=1}^{n} x_{i,j,t-1,k}}.$$

Both the x and z 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.

4.4 Machine Learning Method

The inference problem is to compute a function $f(\cdot)$ that maps country interactions to estimate an indicator of whether conflict occurred. To accomplish this, this project will use a support vector machine (SVM). This method is appropriate for binary classification with real-valued predictors, which makes it well suited for this project.

5 Results

6 Conclusion

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.

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

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.

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*.