Homework 8

Matt Dickenson
Department of Political Science
Duke University
Durham, NC 27708
mcd31@duke.edu

Abstract

Can classification methods help to automate the production of political indicators in near real time? The Militarized Interstate Disputes (MID) dataset, produced by the Correlates of War project, has been widely used in political research over the past three decades and is increasingly used in policy applications. Despite its value for understanding conflict, MID data coding is performed in iterative batches by human coders that lag behind the present by several years. However, reliance solely on human coders is neither necessary nor desirable. Using automated classification methods (Support Vector Machines) to classify real-time event data (GDELT), this project hopes to obtain a close approximation to the MID dataset at a fraction of the cost in both time and money.

1 Methods

1.1 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 [1]. To measure the dependent variable of conflict, the Militarized Interstate Disputes (MID) dataset will be split into subsets for training and validation [2]. 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.

1.2 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 [3]. 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.

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

1.3 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.

1.4 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.

Acknowledgments

Thanks to Michael D. Ward and National Science Foundation Grant #3331808 for support during this project. Any conclusions or errors are the sole responsibility of the author.

References

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

[2] Faten Ghosn, Glenn Palmer, and Stuart A Bremer. The mid3 data set, 1993–2001: Procedures, coding rules, and description. *Conflict Management and Peace Science*, 21(2):133–154, 2004.

[3] Deborah J. Gerner, Philip A. Schrodt, Yilmaz Ömür, and Rajaa Abu-Jabr. Conflict and Mediation Event Observations (CAMEO): A New Event Data Framework for a Post Cold War World. Boston, MA, August, 29-September 1 2002. Annual Meetings of the American Political Science Association.