STA561: Probabilistic machine learning

Automated Production of Political Indicators (9/25/13)

Authors: Matt Dickenson, Department of Political Science

1 Motivation

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. For example, the most recent version was released in 2004 and contains data through 2001. An update through 2010 was expected last summer but is delayed indefinitely. However, reliance solely on human coders is neither necessary nor desirable. Using automated classification methods to classify real-time event data, this project hopes to obtain a close approximation to the MID dataset at a fraction of the cost in both time and money.

2 Problem definition

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 event data for interstate disputes since 2001. The event data input used for classification will include GDELT, which begins with 1979 and is updated daily, and ICEWS, which begins in 2000 and is updated on an approximately monthly basis.

3 Models and methods

This project will use classification methods to categorize country-country-months (e.g. `USA-China-2012-May`) as either in conflict or not. The method used will most likely be HMM, with conflict treated as a latent state. Observed data from GDELT and ICEWS include information how much actors from each country interacted, whether acts were material or verbal, and how conflictual or cooperative each interaction was. Another alternative method to explore is Bayesian Hierarchical Association Rule Mining (HARM).

4 Results and validation

The results of the classification procedure will be a matrix of country-country-months that are categorized as interstate disputes or not. Results through 2001 can be validated against MID data to see whether they accurately classify disputes. Results for dyad-months after 2001 can be easily checked for face validity by checking whether observations classified as interstate disputes match with events that actually occurred. When the next update to the MID dataset is released it will provide a true out-of-sample validation opportunity.