Conflict Prediction with Spike and Slab prior

Anh Le

December 8, 2014

Introduction

Being able to predict political violence is very useful for governments and international organizations to deploy their limited resources effectively. Traditional models of violence in political science have relied on frequentist panel data methods, which have not had good predictive performance because:

- The null hypothesis of $\beta = 0$ being tested is not as important as the substantive effect of the variable
- Many social science variables are highly correlated. For example, *competitive elections*, *civil liberties*, and *property rights* are three conceptually distinct but empirically correlated variables. This results in (logistic) regression with large coefficient variance and meaningless p-value
- The problem of collinearity is compounded by the recent explosion of data available, with hundreds of predictors to be considered

Therefore, this project will use two methods that work well given the large number of predictors:

1. sparse regression with spike-and-slab prior
2. (boosted) classification tree

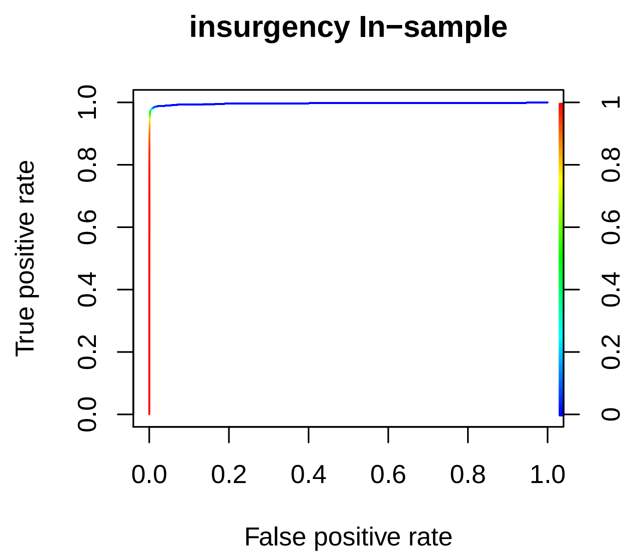
These models are found to perform better than existing models used in our lab.[1]

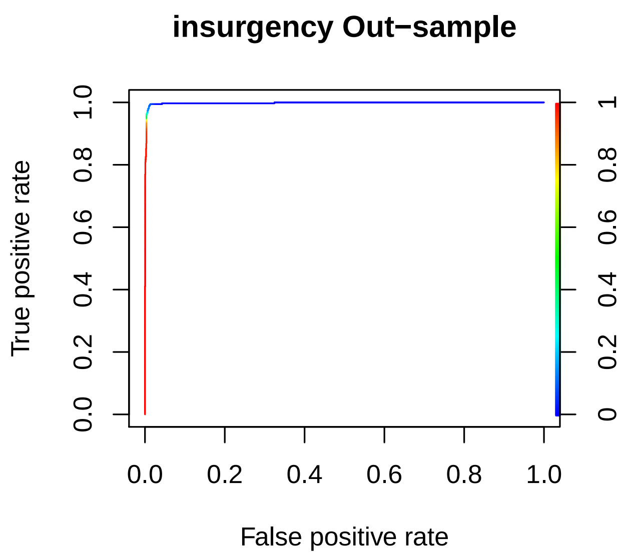
1 Description of Data

The dataset contains monthly data of 167 countries from 2001 to present (dimension = $27,000 \times 550$). The training cutoff date is `2013-06-01`.[2]

[1] We currently use hierarchical linear mixed-effect models, combined together with Bayesian ensemble. The predictors are hand-picked.

[2] This cutoff date is to comply with existing models used in our lab.





The labels are binary indicators of whether an event happens. We are interested in five types of events: insurgency, rebellion, dpc (domestic political crisis), erv (ethnic and religious violence), and mp (massive protest).

The predictors include a country's politics, economics, and financial status. I also include spatial lags (constructed from Gower similarity, 4 nearest neighbors, and centroid distance) and temporal lags (by up to 2 months).

2 Spike-and-Slab

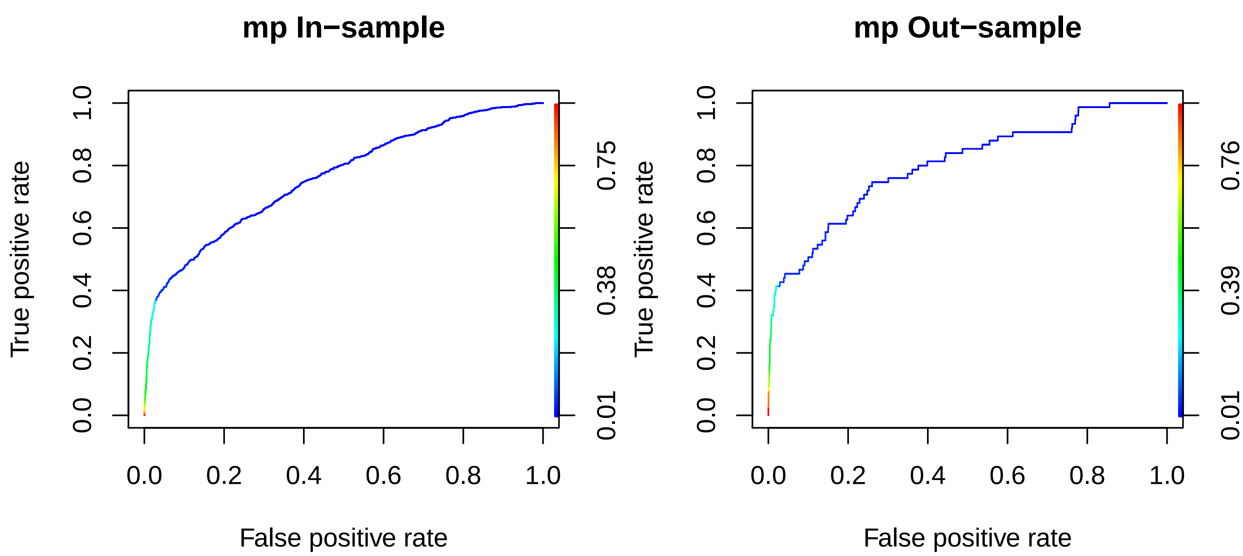
2.1 Model building

I pre-process the data by removing extraneous variables (country names, time ID, etc.) and, for each country in the dataset, add a binary variable that indicates whether an observation belongs to a country. This allows the model to have different intercepts for each country, essentially adding country fixed effect.

I fit the logistic model with spike-and-slab prior using the package `BoomSpikeSlab`. To simultaneously fit one model for each of my five labels, I use packages `doMC` and `foreach`.

Result

The results below come from a MCMC chain with iterations = 1000 and burn-in = 100. Even though the MCMC chain is not that long, due to the size of the dataset the computational time is already over 1.5 hour.



	insurgency	rebellion	dpc	erv	mp
brier	0.005	0.006	0.042	0.008	0.036
auc.C	0.996	0.999	0.927	0.989	0.764
precision	0.981	0.957	0.789	0.961	0.548
recall	0.767	0.769	0.410	0.681	0.068

Table 1: In-sample predictive performance

	insurgency	rebellion	dpc	erv	mp
brier	0.008	0.020	0.097	0.033	0.024
auc.C	0.998	0.930	0.865	0.975	0.801
precision	0.976	0.907	0.544	0.907	0.647
recall	0.946	0.789	0.548	0.490	0.147

Table 2: Out-sample predictive performance