Conflict Prediction via Machine Learning: Addressing the Rare Events Problem with Bagging

Nils B. Weidmann

PhD candidate, International Conflict Research,
ETH Zurich, Switzerland
Research fellow, Department of Political Science,
University of Washington, Seattle, U.S.A.
`weidmann@icr.gess.ethz.ch`

Abstract

Machine learning deals with the development of algorithms for classification and prediction. However, these algorithms have only in rare cases been used in political science. This poster demonstrates the application of state-of-the-art machine learning techniques to the prediction of conflict. In order to address the rare events problem, I use an ensemble of classifiers built on subsets of the training data. These subsets include all positive cases, and a random selection of negative ones. Although I focus primarily on decision tree learning, the proposed method can be used in conjunction with different classification algorithms in order to improve the prediction of conflict onset.

1. Why Machine Learning?

Prediction

One way to make quantitative conflict research more policy-relevant is by providing risk assessments. However, standard regression models are often poor at predicting conflict (Beck, King and Zeng, 2000; Ward and Bakke, 2005). Machine learning methods, on the other hand, are tailored to prediction tasks.

Complex Processes

Conflict is the result of complex interdependencies of a multitude of factors. The processes that bring about civil war are likely to be nonlinear, interactive and context-dependent (Beck, King and Zeng, 2000). Machine Learning methods can deal with relationships of high complexity.

2. Machine Learning

What is Machine Learning?

Machine Learning is a subfield of Computer Science and deals with the development of algorithms that improve their performance with experience (Mitchell, 1997). A "supervised" machine learning task is equivalent to classification: Given certain attributes of a learning instance, the algorithm seeks to predict the correct class.

Evaluation of a Classifier

- Out-of-sample prediction: Separate *training* and *test* sets
- Maximize use of data by K-fold cross-validation
- Measures of predictive accuracy (for a two-class problem):
  1. *Precision* (Proportion of correct positive predictions): TP/(TP+FP)
  2. *Recall* (Proportion of positive cases correctly classified as positive): TP/(TP+FN)

3. Decision Trees

Example

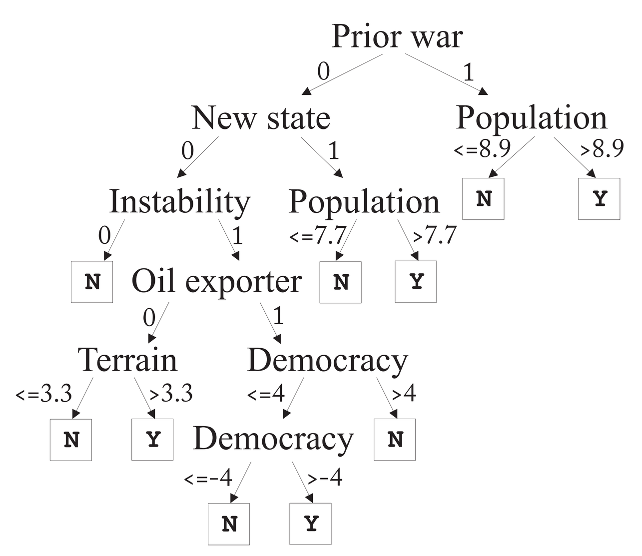

Figure 1: *Decision tree predicting onset (`Yes`/`No`) of civil war, built on the Fearon and Laitin (2003) dataset.*

Decision tree learning (Quinlan, 1986)

At each node, select the attribute A that maximizes the information gain g on the sample S:

$$g(S,A) = \text{entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{S_v}{S} \text{entropy}(S_v)$$

The entropy is the *impurity* of a set of cases. Entropy function for a two-class problem (Mitchell, 1997):

$$\text{entropy}(S) = p_Y log_2 p_Y - p_N log_2 p_N$$

4. The Rare Events Problem

Predicting Conflict

When predicting conflict with machine learning algorithms, we do not obtain useful classifiers because of the high proportion of negative cases. For example, decision tree learners produce a degenerated tree with one leaf (`No`).

A Modified Bagging Procedure

Trees constructed on a balanced subsample (all positive training cases, random selection of negative cases) display a more meaningful structure. However, since the tree structure is dependent on the random selection of negative cases, the tree does not perform well on the full sample. We can repeat the sampling process many times ("Bagging"):

- Create a balanced sample with random selection from the negative cases
- Build a decision tree on it

The predictions of the individual trees are aggregated by majority vote (Breiman, 1996).

Illustration

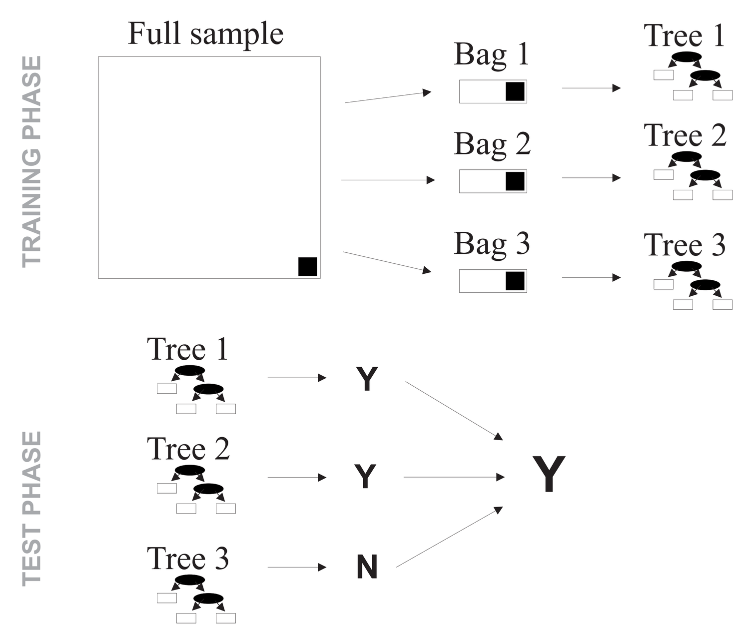

Figure 2: *Training and test phase of the ensemble classification method. Black squares indicate the positive cases.*

5. Results

Bagging and Decision Trees

Prediction task: Onset of civil war (Fearon and Laitin, 2003), Model 1. N=6610, 111 conflict onsets (1.68%). All results are out-of-sample predictions (10-fold cross-validation). Software: `Weka` machine learning package (Witten and Frank, 2005) with `RWeka` interface.

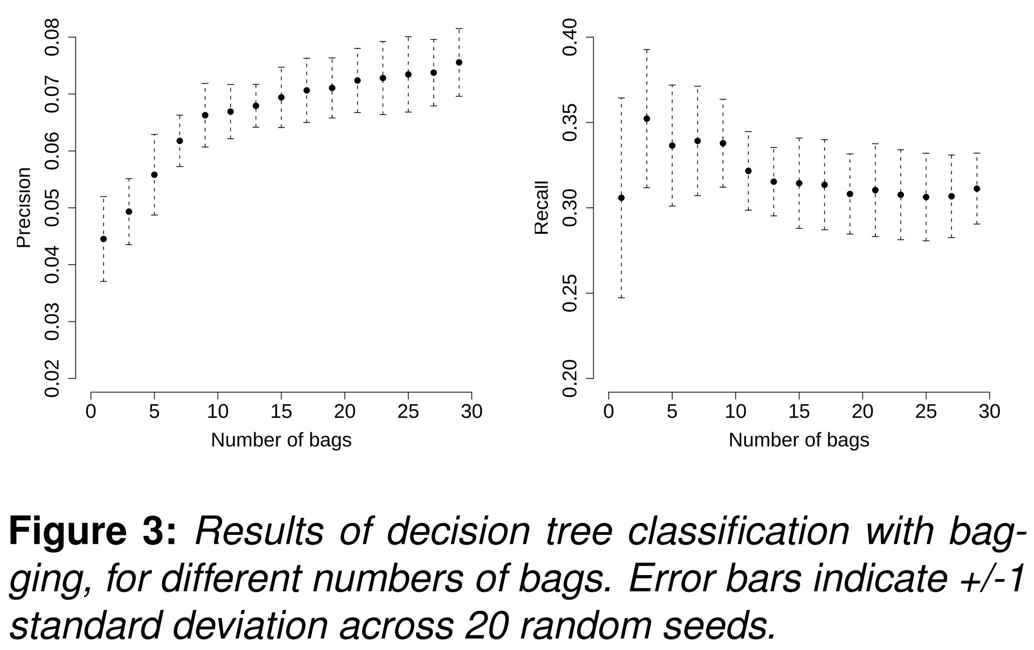

Figure 3: *Results of decision tree classification with bagging, for different numbers of bags. Error bars indicate +/-1 standard deviation across 20 random seeds.*

Assessing the Impact of Individual Variables

Approach: Compare predictive performance of restricted models (with single variables left out) to full model.

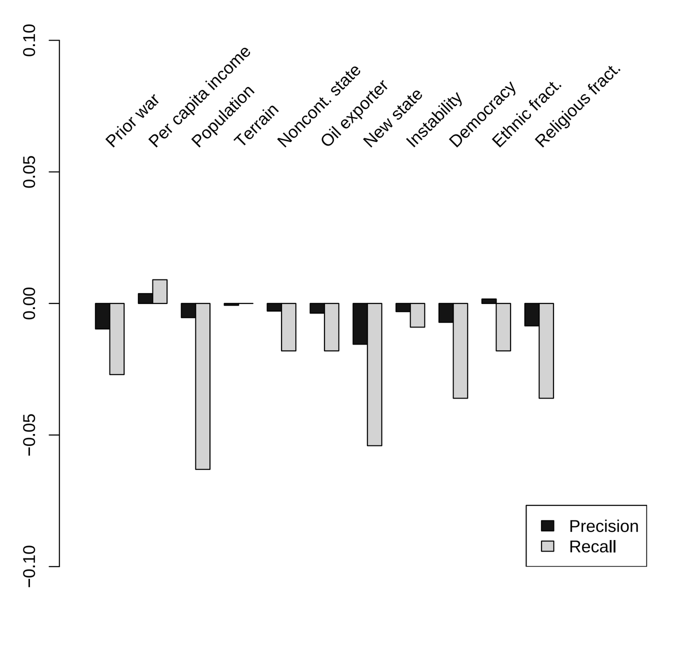

Figure 4: *Changes in precision and recall compared to the full model, when leaving out the respective variable.*

Different Proportions of Positive and Negative Cases in a Bag

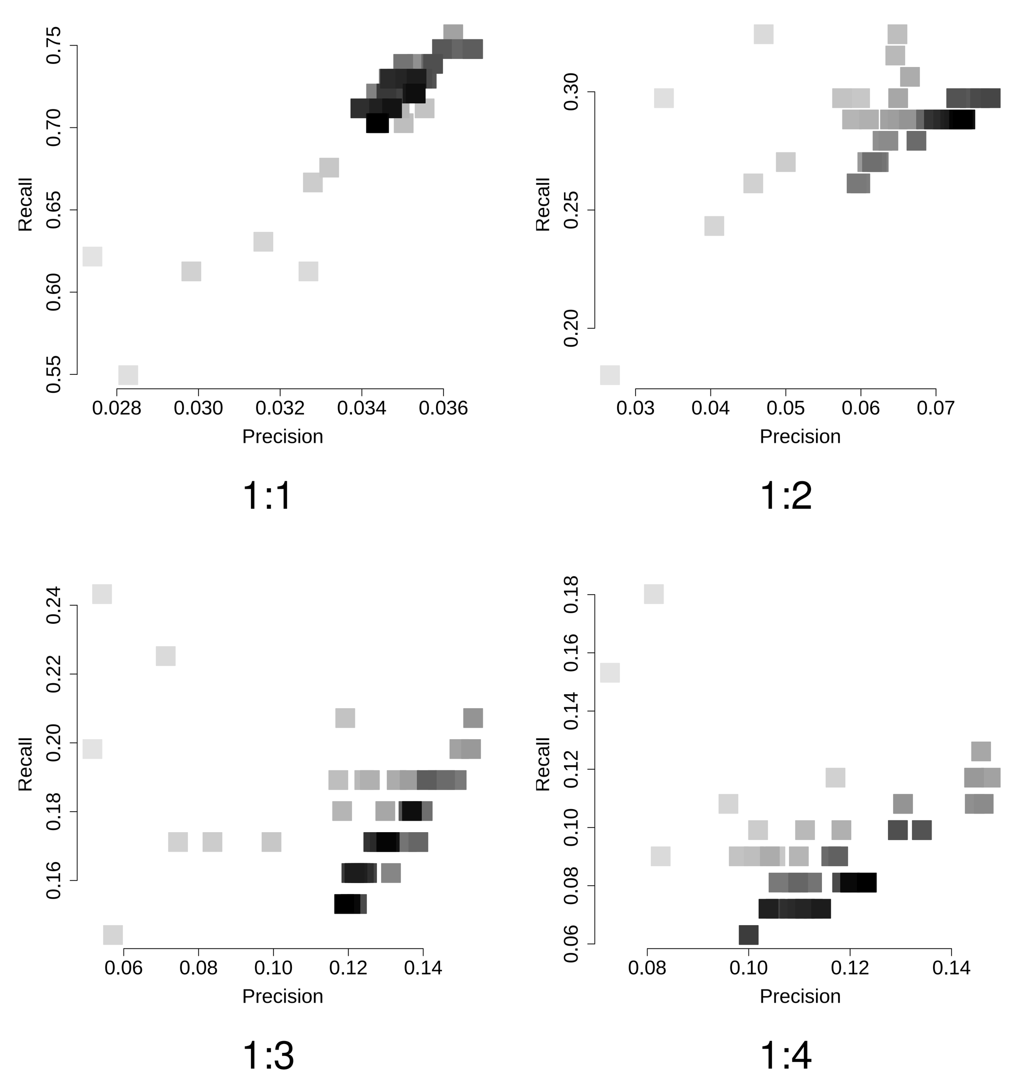

Figure 5: *Results of decision tree classification with bagging, for different ratios of positive vs negative cases in a bag. Darker shadings correspond to higher numbers of bags.*

Alternative Base Classifiers

The proposed methods can also be used in conjunction with other classification algorithms.

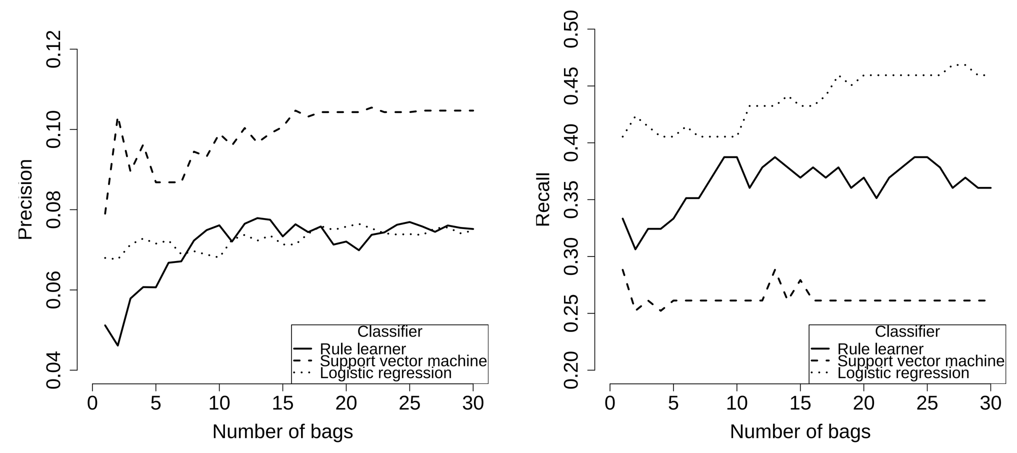

Figure 6: *Precision and Recall for different classification algorithms with bagging, for different numbers of bags.*

6. Discussion

The proposed method makes machine learning algorithms applicable to conflict prediction and can be used in conjunction with different base classifiers. However, the use of an ensemble of classifiers makes the effect of individual variables more difficult to assess.

Acknowledgements

To Michael D. Ward, for his advice on this project.

References

Beck, Nathaniel, Gary King and Langche Zeng. 2000. "Improving Quantitative Studies of International Conflict: A Conjecture." *American Political Science Review* 94(1):21–35.

Breiman, Leo. 1996. "Bagging Predictors." *Machine Learning* 24:123–140.

Fearon, James D. and David D. Laitin. 2003. "Ethnicity, Insurgency and Civil War." *American Political Science Review* 97(1):75–90.

Mitchell, Tom. 1997. *Machine Learning.* McGraw-Hill.

Quinlan, Ross. 1986. "Induction of Decision Trees." *Machine Learning* 1(1):81–106.

Ward, Michael D. and Kristin Bakke. 2005. "Predicting Civil Conflicts: On the Utility of Empirical Research." Paper prepared for the Conference on Disaggregating the Study of Civil War and Transnational Violence, San Diego, CA.

Witten, Ian and Eibe Frank. 2005. *Data MIning: Practical Machine Learning Tools and Techniques.* 2nd ed. Morgan Kaufmann.