QUALITY OF CLASSIFICATION APPROACHES FOR THE QUANTITATIVE ANALYSIS OF INTERNATIONAL CONFLICT

Adalbert F.X. Wilhelm [1]

[1] Department of Psychology and Methods, Jacobs University Bremen, (e-mail: a.wilhelm@jacobs-university.de)

ABSTRACT: We provide an evaluative comparison of some modern classification algorithms, such as CART, AdaBoost, random forests, and support vector machines, to predict the incidences of military conflicts. While modern classification procedures are able to improve the prediction accuracy as compared to the traditionally used logistic regression, the logistic regression still holds a large advantage in terms of interpretability of the variables' relevancy.

KEYWORDS: logistic regression, classification trees, boosting, rare events.

1 Introduction

The increase in available data about military conflicts, e.g. collected by the correlates of war project (Palmer *et al.* , 2015) or the Uppsala Conflict Data Program (Themnér & Wallensteen, 2014), provides the political science researcher with ample opportunities to study the onset of wars and other armed conflicts. This has led to a significant increase in application papers using quantitative analysis techniques to model armed conflicts as one major aspects of international relations. The logistic regression has become kind of a panacea for these analyses. Robust estimation correcting for heteroscedasticity and the cluster structure of the panel data have become standard specifications over the past years.

Over the last decades a number of modern classification algorithms, such as CART, AdaBoost, neural nets, and support vector machines, have been proposed. Here, we compare the traditional logistic regression approach to some modern classification methods as implemented in the `caret` package (Kuhn, 2008) in `R` (R Core Team, 2014). The data sets are chosen in such a way that they cover the most prominent data collection projects in the field of international relations.

2 Data sets

Sub-Saharan Africa I The first data set we have chosen covers arms trade, military expenditure and their impact on the occurrence of armed conflicts in sub-Saharan Africa. The data set has been prepared and originally analysed by Craft & Smaldone (2002) and is available in the replication data archive of the *Journal of Peace Research*.

A primary goal of the original research with this data was to determine whether arms trade (`ARMSTRAN`) is a predictor of political violence in sub-Saharan Africa (cf. Craft & Smaldone, 2002, p. 696) over and above the effect of military expenditure (`MILSPEND`). Common theories on military conflicts agree on the tendency of instable regimes to be more likely involved in military conflicts. Instability of a regime is here operationalized by variables on regime transition (`TRANSITI`), presence of ethno-political groups (`ETHNOPOL`), on political violence exerted by the government (`REPRESSI`) and semi-democracy (`SEMIDEM`). Economic development (`DEVELOP`) is widely seen as an influential factor for keeping peace. The variable `CUMWAR` has been included to control for auto-correlation of war involvement. Some authors have argued that the colonial legacy (`COLONIAL`) also plays a role in this context due to the different styles of ruling in the past and the result different possibilities of political organisation of minorities and political opponents.

Sub-Saharan Africa II The second data set deals with the same region but focuses on the relationship between economic development and civil wars in this region. A special emphasis is put on the endogeneity question here. The data set has been originally analysed by Miguel *et al.* (2004) and is available in the replication data archive of the *Journal of Political Economy*.

Petrostates The third data set deals with non-democracies and the effect of oil on the stability of regimes. Colgan (2015) used this data to shed light on the fact that civil war is more frequent in petrostates and at the same time petrostates experience fewer regime changes. The data is available on the replication data archive of the *Journal of Peace Research*.

ICOW The fourth data set comprises militarized interstate disputes in the time period from 1806 to 2010. It is rarely analysed in total but subsets of it are widely used. In its total, it differs from the other three data sets by its low incidence rate.

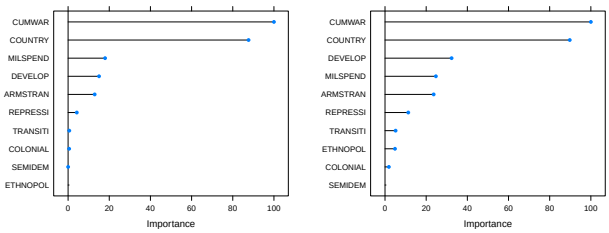

Figure 1. *Variable importance plot to compare the explanatory aspects of the various classifiers. The plots above relate to the data set* `Sub-Saharan Africa I`*; the plot on the left shows the results for the AdaBoost models, the one on the right for the random forest models.*

3 Evaluation

As mentioned before, the typical analysis in this application field aims at explaining the relationship between various potential predictors and the occurrence of conflict. The potential predictors are mainly associated with specific aspects of political theories and the quantitative analysis is performed to provide empirical corroboration for political theories and certain policies. Typically, prediction is not in the major focus of the analysis. The rare cases in which also the models' predictive aspects have been mentioned show that the logistic regression models perform rather poorly in this aspect. King & Zeng (2001) discuss methodological consequences in this field for rare events. The data sets we are looking at, however, show incidence rates that appear large enough to avoid any problematic issues here.

Our evaluative comparison is hence based on two main aspects: the importance of variables within the classifier as well as the prediction accuracy. Other performance measures will be used to differentiate between classifiers in more details. As primary measure for prediction accuracy we took the overall percentage of correctly classified instances in our test sample which consisted of 25% of cases from the corresponding data set.

Evaluation of variable importance is done by comparing the results of the generic function `varImp` from the `R` package `caret`. This function uses different measures for each model. For details, please see Kuhn (2008). The results can be summarized graphically, see Fig 1.

Table 1. *Predictive accuracy for some classification techniques on presented data sets.*

Data set	No-Information rate	Logistic	AdaBoost	Random Forests
Sub-Saharan Africa I	0.7795	0.8937	0.9055	0.8228

4 Conclusion

The algorithms produce rather stable results on the various data sets. The preliminary conclusion so far states that the logistic regression model shows slightly lower predictive accuracy than most of the other algorithms. Since in particular the prediction of the conflict cases is of major interest the performance improvement is larger than it is visible in the overall predictive accuracy. This loss in predictive accuracy is however balanced by the ease in interpretative explanation of the logistic regression model.

References

COLGAN, JEFF D. 2015. Oil, domestic conflict, and opportunities for democratization. *Journal of Peace Research*, **52**(1), 3–16.

CRAFT, CASSADY, & SMALDONE, JOSEPH P. 2002. The Arms Trade and the Incidence of Political Violence in sub-Saharan Africa, 1967-97. *Journal of Peace Research*, **39**(6), pp. 693–710.

KING, GARY, & ZENG, LANGCHE. 2001. Explaining Rare Events in International Relations. *International Organization*, **55**(6), 693–715.

KUHN, MAX. 2008. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, **28**(5), 1–26.

MIGUEL, EDWARD, SATYANATH, SHANKER, & SERGENTI, ERNEST. 2004. Economic Shocks and Civil Conflict: An Instrumental Variables Approach. *Journal of Political Economy*, **112**(4), pp. 725–753.

PALMER, GLENN, D'ORAZIO, VITO, KENWICK, MICHAEL, & LANE, MATTHEW. 2015. The MID4 Data Set: Procedures, Coding Rules, and Description. *Conflict Management and Peace Science*. Forthcoming.

R CORE TEAM. 2014. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

THEMNÉR, LOTTA, & WALLENSTEEN, PETER. 2014. Armed conflicts, 1946-2013. *Journal of Peace Research*, **51**(4), 541–554.