European Policy Analysis - Volume 2, Number 1 - Spring 2016

Decision Trees and Random Forests: Machine Learning Techniques to Classify Rare Events

Simon Hegelich

Technical University of Munich, Bavarian School of Public Policy, Munich, Germany

Abstract

The article introduces machine learning algorithms for political scientists. These approaches should not be seen as a new method for old problems. Rather, it is important to understand the different logic of the machine learning approach. Here, data is analyzed without theoretical assumptions about possible causalities. Models are optimized according to their accuracy and robustness. While the computer can do this work more or less alone, it is the researcher's duty to make sense of these models afterward. Visualization of machine learning results, therefore, becomes very important and is in the focus of this paper. e methods that are presented and compared are decision trees, bagging, and random forests. e latter are more advanced versions of the former, relying on bootstrapping procedures. To demonstrate these methods, extreme shi s in the US budget and their connection to the attention of political actors are analyzed. e paper presents a comparison of the accuracy of different models based on ROC curves and shows how to interpret random forest models with the help of visualizations. e aim of the paper is to provide an example, how these methods can be used in political science and to highlight possible pitfalls as well as advantages of machine learning.

doi: 10.18278/epa.2.1.7

***Keywords**: Machine learning, methods, punctuated equilibrium, statistics for the 21st century*

References

Abedin, Jaynal. 2014. *Data Manipulation with R*. Packt Publishing Ltd, Birmingham.

Berk, Richard A. 2006. "An Introduction to Ensemble Methods for Data Analysis." *Sociological Methods and Research* 34 (3): 263–295.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32.

Cantú, Francisco, and Sebastián M. Saiegh. 2011. "Fraudulent Democracy? An Analysis of Argentina's Infamous Decade Using Supervised Machine Learning." *Political Analysis* 19 (4): 409–433.

Chen, Chao, Andy Liaw, and Leo Breiman. 2004. *Using Random Forest to Learn Imbalanced Data*. Berkeley: University of California.

Conway, Drew, and John White. 2012. *Machine Learning for Hackers*. O'Reilly Media, Inc., Sebastopol

Ergül, Özgür. 2013. *Guide to Programming and Algorithms Using R*. Springer, New York.

Frohwein, Hendrik I., and James H. Lambert. 2000. "Risk of Extreme Events in Multiobjective Decision Trees Part 1. Severe Events." *Risk Analysis* 20 (1): 113–124.

Giles, Jim. 2012. "Making the Links." *Nature* 488 (7412): 448–450.
Grabau, Martina, and Simon Hegelich. 2016. "The Gas Game: Simulating Decision- Making in the European Union's External Natural Gas Policy." *Swiss Political Science Review*. doi: 10.1111/spsr.12202.

Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Sata: e Promise and Pitfalls of Automatic Content Analysis Methods for Political Rexts." *Political Analysis* 21 (3): 267–297.

doi: 10.18278/epa.2.1.7

Hainmueller, Jens, and Chad Hazlett. 2014. "Kernel Regularized Least Squares: Reducing Misspecification Bias with a Flexible and Interpretable Machine Learning Approach." *Political Analysis* 22 (2):143–168.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. 2 vol. New York: Springer.

Hegelich, Simon, Cornelia Fraune, and David Knollmann. 2015. "Point Predictions and the Punctuated Equilibrium eory: A Data Mining Approach—US Nuclear Policy as Proof of Concept." *Policy Studies Journal* 43 (2): 228–256.

Hill, Daniel W., and Zachary M. Jones. 2014. "An Empirical Evaluation of Explanations for State Repression." *American Political Science Review* 108 (03): 661–687.

Hopkins, Daniel J., and Gary King. 2010. "A Method of Automated Nonparametric Content Analysis for Social Science." *American Journal of Political Science* 54 (1): 229–247.

Jacoby, William G., and David A. Arm- strong. 2014. "Bootstrap Con dence Regions for Multidimensional Scaling Solutions." *American Journal of Political Science* 58 (1): 264–278.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. Springer, New York.

Lazer, David, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, and Myron Gutmann. 2009. "Life in the Network: The Coming age of Computational Social Science." *Science* (New York, NY) 323 (5915): 721.

Mayer-Schönberger, Viktor, and Kenneth Cukier. 2013. *Big Data: A Revolution that will Transform how We Live*, Work, and ink. Houghton Mi in Harcourt, London.

Mitchell, Tom M. 1997. *Machine Learning*. 1997. Burr Ridge, IL: McGraw-Hill, 45.

Montgomery, Jacob M., Florian M. Hollenbach, and Michael D. Ward. 2012. "Improving Predictions using Ensemble Bayesian Model Averaging." *Political Analysis* 20 (3): 271–291.

doi: 10.18278/epa.2.1.7

Mooney, Christopher Z. 1996. “Bootstrap Statistical Inference: Examples and Evaluations for Political Science.” *American Journal of Political Science* 40(2)570–602.

Shikano, Susumu. 2006. “Bootstrap und Jackknife.” in: Behnke, Joachim/Gwschend, Thomas/Schindler, Delia/ Schnapp, Kai-Uwe (Eds.) *Methoden der Politikwissenschaft. Neuere qualitative und quantitative Analyseverfahren. Baden- Baden* Nomos 69 –79.

Silver, Nate. 2012. e Signal and the Noise: *Why so Many Predictions Fail-But Some Don’t*. New York: Penguin.

Siroky, David S. 2009. “Navigating Random Forests and Related Advances in Algorithmic Modeling.” *Statistics Surveys* 3: 147–163.

Suzuki, Takafumi. 2009. “Extracting Speaker-Specific Functional Expressions from Political Speeches Using Random Forests in Order to Investigate Speakers’ Political Styles.” *Journal of the American Society for Information Science and Technology* 60 (8): 1596–1606.

True, James L. 2009. "Historical budget records converted to the present functional categorization with actual results for FY 1947-2008." *Policy Agendas Project*.

doi: 10.18278/epa.2.1.7