Capturing Meaning & Impact in Real Time

Mixed-Methods & Supervised Machine Learning in Big Data Policy Analysis

Presenters: Amy Castro Baker, PhD, University of Pennsylvania
& Maria Y. Rodriguez, MSW, PhC, University of Washington
Association for Pubic Policy & Management Conference
Miami, FL
November 11th, 2015

Do We Still Need Theory?

- "This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves (Anderson, 2009, p.1)."
- "automatically discover insights – regardless of complexity– without asking questions... (Clark, 2013)."

Epistemologizing Big Data

- Positionality and Reflexivity in Big Data Sets.
- Social Orthogonality
- Anchoring methods in policy process theory

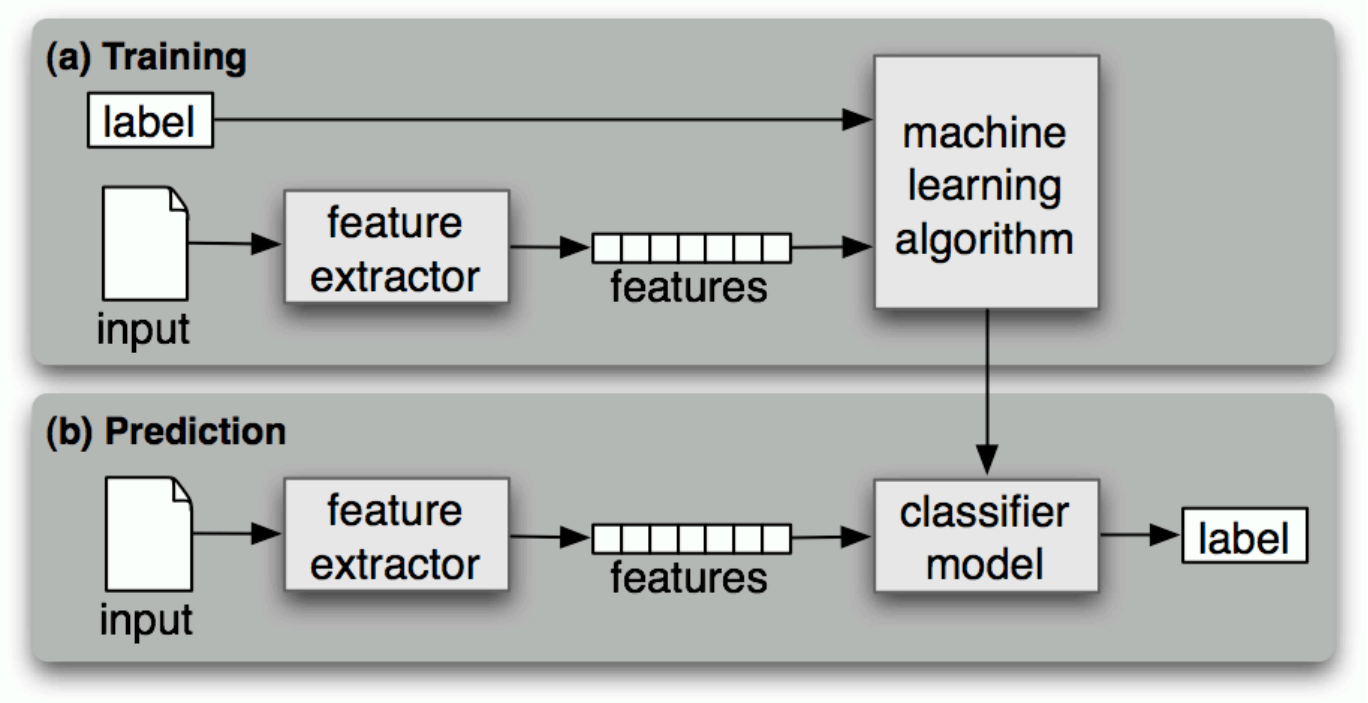
Big Data Mixed Methods

- Advocacy Coalition Framework
- Exploratory Research
- Simultaneous Generation and Confirmation of theory
- Sequential Mixed Methods Design

The ACA using SML Approach

- Addresses three main critiques of Big Data research:
  - Reproducibility
    - Open Source Software & Technical Support
  - Social Scientific Rigor
    - Theoretically guided
    - Researcher firmly in the analysis process
  - Ethical Considerations
    - Accessibility of the method to practitioners and academics alike
    - Allows for measured use of data generated by changing processes

SML Overview



The Supervised Machine Learning Process (Source: Bird et al. 2009)

Software & Algorithms

- RTextTools Package by Jurka et al (2011) in R
- 9 algorithms:
  - Support Vector Machine (SVM)
  - Maximum Entropy (MaxEnt)
  - GlmNet
  - SLDA
  - Booting
  - Bagging
  - Random Forest
  - Neural Networks
  - Classification Tree

Sequential Mixed-Method Design

- 5 Steps:
  - Manual Coding
    - Theory selection
    - Coding
    - Hypothesis generation*
  - Text Parsing
    - Test & Training set creation
  - Algorithm Training
  - Interpreting Results
  - Verification & Troubleshooting

Sample Results

Code	SVM			SLDA			Boosting			Bagging			Random Forests			GLMnet		
Coalitions	*Precision*	*Recall*	*F-Score*	*Precision*	*Recall*	*F-Score*	*Precision*	*Recall*	*F-Score*	*Precision*	*Recall*	*F-Score*	*Precision*	*Recall*	*F-Score*	*Precision*	*Recall*	*F-Score*
Coaliton A	0.830	0.930	0.880	0.780	0.780	0.780	0.770	0.910	0.830	0.780	0.920	0.840	0.780	0.960	0.860	0.780	0.870	0.820
Coalition B	0.770	0.560	0.650	0.500	0.500	0.500	0.670	0.410	0.510	0.700	0.440	0.540	0.820	0.410	0.550	0.610	0.450	0.520
Beliefs																		
Deep Core																		
Coalition A	0.780	0.960	0.860	0.810	0.820	0.810	0.780	0.940	0.850	0.780	0.870	0.820	0.780	0.980	0.870	0.800	0.930	0.860
Coaliton B	0.640	0.230	0.340	0.470	0.450	0.460	0.580	0.230	0.330	0.430	0.290	0.350	0.750	0.190	0.300	0.620	0.320	0.420
Policy Core																		
Coalition A	0.780	0.910	0.840	0.800	0.670	0.730	0.790	0.800	0.790	0.780	0.880	0.830	0.770	0.970	0.860	0.780	0.860	0.820
Coaliton B	0.570	0.330	0.420	0.390	0.560	0.460	0.450	0.440	0.440	0.500	0.330	0.400	0.750	0.230	0.350	0.470	0.330	0.390
Secondary																		
Coalition A	0.820	0.890	0.850	0.710	0.750	0.730	0.600	0.890	0.720	0.710	0.820	0.760	0.730	0.900	0.810	0.690	0.920	0.790
Coaliton B	0.830	0.740	0.780	0.630	0.580	0.600	0.560	0.190	0.280	0.690	0.550	0.610	0.810	0.550	0.660	0.790	0.430	0.560

Sample Results

Ensemble Classification								
n-Algorithms	**Coalitions**		**Deep Core Beliefs**		**Policy Core Beliefs**		**Secondary Beliefs**	
	Coverage	*Recall*	*Coverage*	*Recall*	*Coverage*	*Recall*	*Coverage*	*Recall*
1	1	0.8	1	0.77	1	0.76	1	0.8
2	1	0.8	1	0.77	1	0.76	1	0.8
3	1	0.8	1	0.77	1	0.76	1	0.8
4	1	0.8	1	0.77	1	0.76	1	0.8
5	0.94	0.82	0.96	0.77	0.91	0.78	0.94	0.81
6	0.84	0.85	0.87	0.8	0.74	0.83	0.83	0.85
7	0.71	0.87	0.74	0.85	0.58	0.88	0.55	0.86
8	0.44	0.91	0.58	0.89	0.33	0.89	0.27	0.79

Sample Verification

Deep Core Belief Confusion Matrices

Deep Core Beliefs : Consensus Coded			
		Predicted	
		Coalition A	Coalition B
Actual	Coaltion A	57	15
	Coalition B	14	39

Deep Core Beliefs :Probability Coded			
		Predicted	
		Coalition A	Coalition B
Actual	Coalition A	55	17
	Coalition B	23	30

Deep Core Beliefs : SVM Label			
		Predicted	
		Coalition A	Coalition B
Actual	Coalition A	64	8
	Coalition B	14	39

Deep Core Beliefs : MAXent Label			
		Predicted	
		Coalition A	Coalition B
Actual	Coalition A	70	2
	Coalition B	23	30

Deep Core Beliefs: Boosting Label			
		Predicted	
		Coalition A	Coalition B
Actual	Coalition A	85	5
	Coalition B	24	7

Deep Core Beliefs : GLMnet Label			
		Predicted	
		Coalition A	Coalition B
Actual	Coalition A	66	6
	Coalition B	30	23

Deep Core Beliefs : TREE Label			
		Predicted	
		Coalition A	Coalition B
Actual	Coalition A	51	21
	Coalition B	18	35

Deep Core Beliefs : SLDA Label			
		Predicted	
		Coalition A	Coalition B
Actual	Coalition A	54	18
	Coalition B	22	31

Deep Core Beliefs : Bagging Label			
		Predicted	
		Coalition A	Coalition B
Actual	Coalition A	59	13
	Coalition B	24	29

Deep Core Beliefs : Forests Label			
		Predicted	
		Coalition A	Coalition B
Actual	Coalition A	65	7
	Coalition B	70	29

Summary

- ACA using SML address some of the biggest concerns policy scholars may have concerning the use of Big Data in research
- Resource-light, the biggest cost is time to learn the method
- Built in mechanisms for reproducibility, adherence to the standards of social scientific rigor, and ethical concerns
- Accessible methodology, can be used by academics and practitioners alike

Acknowledgments

- This project was partially supported by the National Center For Advancing Translational Sciences of the National Institutes of Health under Award Number TL1TR000422. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.