Predicting military conflicts by data-driven techniques

Adalbert F.X. Wilhelm,

Herbsttagung AG DANK, WIAS Berlin
November 19th, 2016
Outline

1. Military conflicts: data structures and data projects
2. Logistic regression
3. Requirements for “accepted” data-driven classifiers
4. Some empirical evaluations
5. Class imbalance
6. Conclusion
Modeling armed conflict

- one of the major topics in international relations
- events are of great importance
- modeling onset, duration, or termination

Definition:

- Armed Conflict: An armed conflict is a contested incompatibility that concerns government and/or territory where the use of armed force between two parties, of which at least one is the government of a state, results in at least 25 battle-related deaths in one calendar year.
- “Armed conflict” is also referred to as “state-based conflict”, as opposed to “non-state conflict”, where none of the warring parties are a government.
- War = Armed conflict with at least 1000 battle related deaths

Data typically in dyads (country-year) or triads (country-country-year)

modeling

- Onset
- duration
- termination of armed conflicts
Some data projects for conflict studies

- Correlates of War project
  - data from 1816 - 2010
  - military conflicts
    - between or among non-state entities (non-state war),
    - between states (inter-state war),
    - and within states (intra-state war).

- Militarized Interstate Disputes
  - all instances of when one state threatened, displayed, or used force against another.

- Uppsala Conflict Data Program (UCDP/Prio)
  - a conflict-year dataset with information on armed conflict where at least one party is the government of a state in the time period 1946-2013.
  - comprises 2134 conflicts
  - involving 116 states
  - involving 547 opponents
  - covering 68 years
Some data projects for conflict studies

- **KOSIMO**
  - A conflict-year dataset with information on violent and non-violent conflicts where at least one actor is nation-state in the time period 1945-1999.
  - Comprises 301 conflicts and 693 conflict episodes
  - Involving 171 states
  - Every conflict described by 28 variables

- **ICB International Conflict Behavior**
  - Four data sets covering the period from 1917 to 2001
  - Different units of analysis: nation-state, international system, nation-dyads, one-sided conflicts
What do we do with all this data?

Technology has brought us so much...

- Internet
- Moon Landing
- Human Genome Sequencing

But... it taps into our deepest FEARS.

- “smoking” can refer to cigarettes, marijuana, ribs, “hot women.”

Data Types:
- Images
- Text
- Video
- Audio

All are created by PEOPLE and they require CONTEXT.

Example: If you want to run an anti-smoking campaign, you have to know how people talk about it.

Orwell feared that the TRUTH would be concealed from us.

(A captive culture)

Huxley feared that we would be drowned in a sea of IRRELEVANCE.

(A trivial culture)

Facts are vulnerable to misuse.

For example, a mother's PROXIMITY to freeways has been correlated to AUTISM. (!)

Facts are...
- Stubb* vs stubborn things.
- Stupid things.
- Stuck in the middle.

Ho, misspoke, meaning to quote John Adams at the Boston Massacre trial that FACTS are STUBBORN.

Ronald Reagan once said... Does a set of data make you feel more comfortable? More successful? Then your interpretation of it is probably WRONG.

Assessments & Analytics can value one form of METRIC over another.

Data doesn't create meaning, WE do!

We have a RESPONSIBILITY to SHARPEN our critical thinking skills.

The HUMANITIES give us CONTEXT for BIG DATA.

The humanities can make BAD decisions more quickly & efficiently than ever before.

S. Etlinger

Embedded APS for Course 990111 (Emp. Res.)

IRC – Library

Fall 2011
Standard approaches for modeling occurrence of events in the Social Sciences

- Logit model
  - Dichotomous response
  - A set of predictors (continuous and categorical)
  - Model formulation on the linear predictor level using the link function

\[
\log\left( \frac{p_i}{1 - p_i} \right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k
\]

- Model formulation on the response level using the inverse link function

\[
P(Y = 1) = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)}
\]
Standard approaches for modeling occurrence of events in the Social Sciences

- Which predictors are significant?
- Focus on specific predictors: are they complementary or is one of them redundant?
- Sequential model comparison

- Quality of models?
  - Prediction?
- Logistic regression misses out in predicting conflict cases!
The Role of Prediction in the Social Sciences

- Prediction is a contentious issue in the Social Sciences
  - focus on estimation of causal parameters
    - Priority is given to identifying causal effects (Beck et al. 2000; Ward et al. 2010)
    - Refinement of established models to evaluate additional/alternative causal mechanisms
  - model fit often neglected
  - P-value overuse (ASA Statement on statistical significance and p-values, 2016)
- Growing literature on predicting occurrence of events
  - civil war (Hegre et al. 2013; Shellman et al. 2013; Brandt et al. 2014; Clayton and Gleditsch 2014)
  - interstate disputes (Gleditsch and Ward 2012),
  - political instability (Goldstone et al. 2010)
Requirements for “accepted” data-driven classifiers in the social sciences

• improved prediction accuracy
• explanatory capability
• adaptability to class-imbalanced data

• ideally, allowing discussion of “causal effects”
The Single Model Philosophy

Motivation: Occam’s Razor

• “one should not increase, beyond what is necessary, the number of entities required to explain anything”
• Infinitely many models can explain any given dataset
• Might as well pick the smallest one…

Ensemble Philosophy

Build many models and combine them
Only through averaging do we get at the truth!
It’s too hard (impossible?) to build a single model that works best
Two types of approaches:

• Models that don’t use randomness
• Models that incorporate randomness
Support Vector Machines


- Predicting stability level of a state (three levels)
- KOSIMO data base: consists of eleven macro-structural indicators (factors, attributes, features) for 171 countries over the period 1975-1999.

- Comparison against multinomial logit and unrestricted fuzzy analysis of statistical evidence (UnFASE)

<table>
<thead>
<tr>
<th></th>
<th>Multinomial Logit</th>
<th>UnFASE</th>
<th>Proposed Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Overall</td>
<td>79%</td>
<td>79%</td>
<td>94%</td>
</tr>
<tr>
<td>Average Recall</td>
<td>69%</td>
<td>75%</td>
<td>91%</td>
</tr>
<tr>
<td>Average Precision</td>
<td>62%</td>
<td>66%</td>
<td>90%</td>
</tr>
</tbody>
</table>
Single model philosophy

Data: Occurrence of military conflicts in sub-Saharan Africa (Craft & Smaldone, 2002)

Different splits into training and test data

<table>
<thead>
<tr>
<th></th>
<th>CART</th>
<th>Logistic Regression</th>
<th>Naïve Bayes Classifier</th>
<th>Linear Discriminant Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.8898</td>
<td>0.8976</td>
<td>0.8819</td>
<td>0.8898</td>
</tr>
<tr>
<td>2</td>
<td>0.8701</td>
<td>0.9094</td>
<td>0.9016</td>
<td>0.9213</td>
</tr>
<tr>
<td>3</td>
<td>0.9016</td>
<td>0.874</td>
<td>0.8661</td>
<td>0.8937</td>
</tr>
<tr>
<td>4</td>
<td>0.874</td>
<td>0.878</td>
<td>0.8543</td>
<td>0.8819</td>
</tr>
<tr>
<td>5</td>
<td>0.9016</td>
<td>0.8976</td>
<td>0.874</td>
<td>0.8976</td>
</tr>
<tr>
<td>6</td>
<td>0.9134</td>
<td>0.9055</td>
<td>0.8898</td>
<td>0.9134</td>
</tr>
<tr>
<td>7</td>
<td>0.8937</td>
<td>0.8898</td>
<td>0.8701</td>
<td>0.8819</td>
</tr>
<tr>
<td>8</td>
<td>0.8819</td>
<td>0.8661</td>
<td>0.8346</td>
<td>0.8661</td>
</tr>
<tr>
<td>9</td>
<td>0.9409</td>
<td>0.9173</td>
<td>0.9055</td>
<td>0.9173</td>
</tr>
<tr>
<td>10</td>
<td>0.9016</td>
<td>0.8937</td>
<td>0.874</td>
<td>0.9094</td>
</tr>
</tbody>
</table>

AVG 0.8963 0.8928 0.8753 0.8959


Variables
- Warinvol: war involvement binary variable, from Gleditsch et al.
- Year: year, 1967 through 1997
- Colonial: colonial indicator, from Blanton et al.
- Country: country name
- Transiti: transition binary, from Polity IV
- Ethnopol: ethno-political groups indicator, from Minorities at Risk
- Repressi: repression indicator, from Polity IV
- Semidem: semi-democracy indicator, from Polity IV
- Arms: arms imports, from WMEAT (log values)
- Milspend: per capita military spending, from WMEAT (log values)
- Develop: per capita GNP, from WMEAT/World Bank (log values)
- Cumwar: 5-year moving average of war magnitude, from Gleditsch et al.
Ensemble Approaches

Bagging
  • **Bootstrap aggregating**

Boosting

Random Forests
  • Bagging reborn
  • Well-established
Bagging

Main Assumption:

- Combining many unstable predictors to produce a ensemble (stable) predictor.
- Unstable Predictor: small changes in training data produce large changes in the model.
  - e.g. Neural Nets, trees
  - Stable: SVM (sometimes), Nearest Neighbor.

Hypothesis Space

- Variable size (nonparametric):
  - Can model any function if you use an appropriate predictor (e.g. trees)
Boosting

• Originally developed by computational learning theorists to guarantee performance improvements on fitting training data for a weak learner that only needs to generate a hypothesis with a training accuracy greater than 0.5 (Schapire, 1990).
• Revised to be a practical algorithm, AdaBoost, for building ensembles that empirically improves generalization performance (Freund & Shapire, 1996).
• Key Insights:
  • Instead of sampling (as in bagging) re-weigh examples!
  • Examples are given weights. At each iteration, a new hypothesis is learned (weak learner) and the examples are reweighted to focus the system on examples that the most recently learned classifier got wrong.
  • Final classification based on weighted vote of weak classifiers
Each classifier $G_m(x)$ is trained from a weighted sample of the training data.

Each predictor is created by using a biased sample of the training data:

- Instances (training examples) with high error are weighted higher than those with lower error.

Difficult instances get more attention:

- This is the motivation behind boosting.
Random Forest

• Leo Breiman, Random Forests, Machine Learning, 45, 5-32, 2001
• Motivation: reduce error correlation between classifiers
• Main idea: build a larger number of un-pruned decision trees
• Key: using a random selection of features to split on at each node
How Random Forest Work

• Each tree is grown on a bootstrap sample of the training set of \( N \) cases.
• A number \( m \) is specified much smaller than the total number of variables \( M \) (e.g. \( m = \sqrt{M} \)).
• At each node, \( m \) variables are selected at random out of the \( M \).
• The split used is the best split on these \( m \) variables.
• Final classification is done by majority vote across trees.
Random Forest (part of it)
Advantages of random forest

• Error rates compare favorably to Adaboost
• More robust with respect to noise.
• More efficient on large data
• Provides an estimation of the importance of features in determining classification

• http://stat-www.berkeley.edu/users/breiman/RandomForests/cc_home.htm
Data sets

<table>
<thead>
<tr>
<th>Data set</th>
<th># cases</th>
<th># countries</th>
<th># time periods</th>
<th>incidence rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub-Saharan Africa I</td>
<td>1017</td>
<td>46</td>
<td>26</td>
<td>0.22</td>
</tr>
<tr>
<td>Sub-Saharan Africa II</td>
<td>743</td>
<td>41</td>
<td>19</td>
<td>0.27</td>
</tr>
<tr>
<td>Petrostates</td>
<td>7768</td>
<td>188</td>
<td>59</td>
<td>0.17</td>
</tr>
</tbody>
</table>

Table 1: Summary information about the data sets used in the evaluative comparison.
Sub-Saharan Africa I
Ensemble model philosophy

Data: Occurrence of military conflicts in Sub-Saharan Africa (Craft & Smaldone, 2002)

Random forests
10-fold Cross-validation

Variables Description of the variables.

Warinvol war involvement binary variable, from Gleditsch et.al.
Year year, 1967 through 1997
Colonial colonial indicator, from Blanton et. al.
Country country name
Transiti transition binary, from Polity IV
Ethnopol ethno-political groups indicator, from Minorities at Risk
Repressi repression indicator, from Polity IV
Semidem semi-democracy indicator, from Polity IV
Armsran arms imports, from WMEAT (log values)
Milspend per capita military spending, from WMEAT (log values)
Develop per capita GNP, from WMEAT/World Bank (log values)
Cumwar 5-year moving average of war magnitude, from Gleditsch et.al.
## Ensemble model philosophy

### Table 2: Predictive accuracy for some classification techniques as measured by AUC (area under the ROC curve) on presented data sets.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Logistic</th>
<th>AdaBag</th>
<th>AdaBoost</th>
<th>Random Forests</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub-Saharan Africa I</td>
<td>0.8307</td>
<td>0.743</td>
<td>0.7551</td>
<td>0.8078</td>
</tr>
<tr>
<td>Sub-Saharan Africa II</td>
<td>0.6017</td>
<td>0.5</td>
<td>0.7124</td>
<td>0.6048</td>
</tr>
<tr>
<td>Petrostates</td>
<td>0.7086</td>
<td>0.5</td>
<td>0.7124</td>
<td>0.6048</td>
</tr>
</tbody>
</table>

---

W. (2016)
Variable Importance Plots
(Sub-Saharan Africa Data)

Gini importance
mean Gini gain produced by $X_j$ over all trees
for variables of different types: biased in favor
of continuous variables and variables with
many categories (Strobl et al., 2007)

Permutation importance
mean decrease in classification accuracy after
permuting $X_j$ over all trees
for variables of different types: unbiased only
when subsampling is used (Strobl et al., 2007)

Separation Plots (Sub-Saharan Africa Data)

CRAFT AND SMALDONE (2002) -- logit

Random Forests

R Package: separationplot
Greenhill et al. (2015)

W. (2016)
Partial Dependence Plots (Civil War Data)

Muchlinski et al. (2016)
Class-imbalance

- Broad range of incidence rates
  - Restriction to politically relevant cases
  - Restriction to specific regions
  - Restriction to specific time frames
  - All these selections implicitly correct for class-imbalance!

<table>
<thead>
<tr>
<th>Data set</th>
<th># cases</th>
<th># countries</th>
<th># time periods</th>
<th>incidence rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub-Saharan Africa I</td>
<td>1017</td>
<td>46</td>
<td>26</td>
<td>0.22</td>
</tr>
<tr>
<td>Sub-Saharan Africa II</td>
<td>743</td>
<td>41</td>
<td>19</td>
<td>0.27</td>
</tr>
<tr>
<td>Petrostates</td>
<td>7768</td>
<td>188</td>
<td>59</td>
<td>0.17</td>
</tr>
<tr>
<td>Civil War Data</td>
<td>7141</td>
<td>172</td>
<td>45</td>
<td>0.0165</td>
</tr>
<tr>
<td>ICOW</td>
<td>36156</td>
<td>34</td>
<td>109</td>
<td>0.0077</td>
</tr>
<tr>
<td>UCDP</td>
<td>4314736</td>
<td>116</td>
<td>68</td>
<td>0.0005</td>
</tr>
</tbody>
</table>
Rare events correction – Class imbalance problem

- For logistic regression, options to correct predicted probabilities for imbalanced data or to use penalized logistic regression (Firth’s method)
- yields unbiased estimates for class-imbalanced data
Rare events correction – Class imbalance problem

- For data-driven classifiers use sampling
  - Down-sampling
    - Loses information on majority class
  - Up-sampling
    - Repeats information of minority class

- Specify sampling counts per strata
  - Balanced design
  - Over-sampling minority class
Class imbalance problem - Solution approaches

- Sampling
  - Up-Sampling (Oversampling)
    - Repeats information of minority class
  - Down Sampling (Undersampling)
    - Loses information of majority class
  - SMOTE (Synthetic Minority Over Sampling Technique)
  - Cluster-based or strata based sampling
Class imbalance problem - Solution approaches

SMOTE-Algorithm (k-NN approach)

Algorithm SMOTE(T, N, k)

Input: Number of minority class samples T;
Amount of SMOTE N%;
Number of nearest neighbors k

Output: \((N/100)\times T\) synthetic minority class samples

1. (* If \(N\) is less than 100%, randomize the minority class samples as only a random percent of them will be SMOTE'd.*)
2. if \(N < 100\)
3. then Randomize the \(T\) minority class samples
4. \[T = (N/100) \times T\]
5. \[N = 100\]
6. endif
7. \[N = \text{int}(N/100)\]
(*Amount of SMOTE is in integral multiples of 100.*)
8. \(k\) = Number of nearest neighbors
9. \(\text{numattrs}\) = Number of attributes
10. \(\text{Sample}[][]\) : array for original minority class samples
11. \(\text{newindex}\) : keeps a count of number of synthetic samples generated, initialized to 0
12. \(\text{Synthetic}[][]\) : array for synthetic samples
(* Compute \(k\) nearest neighbors for each minority class sample. *)

13. for \(i = 1\) to \(T\)
14. Compute \(k\) nearest neighbors for \(i\),
and save the indices in the \(\text{nnarray}\)
15. Populate\((N, i, \text{nnarray})\)
16. endif

Populate\((N, i, \text{nnarray})\)

(*Function to generate the synthetic samples.*)

17. while \(N != 0\)
18. Choose a random number between 1 and \(k\), call it \(\text{nn}\).
(*This step chooses one of the \(k\) nearest neighbors of \(i\).*
19. for \(\text{attr} \leftarrow 1\) to \(\text{numattrs}\)
20. Compute: \(\text{dif} = \text{Sample[nnarray[nn]][attr]} - \text{Sample}[i][\text{attr}]\)
21. Compute: \(\text{gap} = \text{random number between 0 and 1}\)
22. \(\text{Synthetic[newindex][attr]} = \text{Sample}[i][\text{attr}] + \text{gap}\times\text{dif}\)
23. endfor
24. \(\text{newindex}++\)
25. \(N = N - 1\)
26. endwhile
27. return(* End of Populate. * )
Class imbalance problem - Solution approaches

- Cost-sensitive learning
  - Weighted learning

<table>
<thead>
<tr>
<th>True</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Class i</td>
</tr>
<tr>
<td>Class i</td>
<td>0</td>
</tr>
<tr>
<td>Class j</td>
<td>$\lambda_{ji}$</td>
</tr>
</tbody>
</table>

Table 1: Cost matrix

- Recognition based learning
- Ensemble methods
- Combinations of the above
## Class imbalance problem - Solution approaches

<table>
<thead>
<tr>
<th>Method</th>
<th>Advantages</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under-sampling</td>
<td>• Independent on underlying classifier.</td>
<td>• May remove significant patterns and cause loss of useful information</td>
</tr>
<tr>
<td></td>
<td>• Can be easily implemented</td>
<td></td>
</tr>
<tr>
<td>Over-sampling</td>
<td></td>
<td>• Time consuming: Introduce additional computational cost</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• May lead to over-fitting</td>
</tr>
<tr>
<td>Cost sensitive</td>
<td>• Minimize the cost of misclassification (by biasing the classifier toward the minority class)</td>
<td>• The misclassification costs (the actual cost of errors) often are unknown</td>
</tr>
<tr>
<td>Recognition based</td>
<td>• Have better performance especially on high dimensional data</td>
<td>• Many classifiers such as decision trees and Naive Bayes cannot be built by one class learning.</td>
</tr>
<tr>
<td>Ensemble</td>
<td>• Better classification performance than individual classifiers</td>
<td>• Time consuming</td>
</tr>
<tr>
<td></td>
<td>• More resilience to noise</td>
<td>• Over fitting</td>
</tr>
</tbody>
</table>

## Case study 2

Data: Sub-Saharan Africa I
Random Forest

<table>
<thead>
<tr>
<th></th>
<th>mtry</th>
<th>ROC</th>
<th>Sens</th>
<th>Spec</th>
<th>ROCSD</th>
<th>SensSD</th>
<th>SpecSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>0.9334589</td>
<td>0.9852186</td>
<td>0.7716912</td>
<td>0.04887503</td>
<td>0.01210057</td>
<td>0.09989737</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>0.9298289</td>
<td>0.9818579</td>
<td>0.7900735</td>
<td>0.05555949</td>
<td>0.01448013</td>
<td>0.07759298</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>0.9263345</td>
<td>0.9736885</td>
<td>0.8025735</td>
<td>0.05418566</td>
<td>0.02213473</td>
<td>0.08566255</td>
</tr>
</tbody>
</table>

### Downsampling

<table>
<thead>
<tr>
<th></th>
<th>mtry</th>
<th>ROC</th>
<th>Sens</th>
<th>Spec</th>
<th>ROCSD</th>
<th>SensSD</th>
<th>SpecSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>0.9325959</td>
<td>0.9621858</td>
<td>0.8209559</td>
<td>0.05274956</td>
<td>0.01733350</td>
<td>0.07974260</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>0.9177396</td>
<td>0.9621311</td>
<td>0.8213235</td>
<td>0.06242765</td>
<td>0.01754704</td>
<td>0.09307596</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>0.9157309</td>
<td>0.9621585</td>
<td>0.8147059</td>
<td>0.06426289</td>
<td>0.01355468</td>
<td>0.09577972</td>
</tr>
</tbody>
</table>
## Case study 2

**Data:** Sub-Saharan Africa II  
**Random Forest**

<table>
<thead>
<tr>
<th>mtry</th>
<th>ROC</th>
<th>Sens</th>
<th>Spec</th>
<th>ROCSD</th>
<th>SensSD</th>
<th>SpecSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.9489306</td>
<td>0.9733449</td>
<td>0.7766667</td>
<td>0.04110217</td>
<td>0.02414699</td>
<td>0.1275474</td>
</tr>
<tr>
<td>2</td>
<td>0.9502497</td>
<td>0.9709059</td>
<td>0.7971429</td>
<td>0.04190666</td>
<td>0.02231801</td>
<td>0.1087551</td>
</tr>
<tr>
<td>3</td>
<td>0.9488522</td>
<td>0.9635889</td>
<td>0.7904762</td>
<td>0.04606556</td>
<td>0.03093548</td>
<td>0.1105883</td>
</tr>
</tbody>
</table>

### Downsampling

<table>
<thead>
<tr>
<th>mtry</th>
<th>ROC</th>
<th>Sens</th>
<th>Spec</th>
<th>ROCSD</th>
<th>SensSD</th>
<th>SpecSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.9419993</td>
<td>0.9297329</td>
<td>0.8380952</td>
<td>0.03681699</td>
<td>0.04016076</td>
<td>0.1176104</td>
</tr>
<tr>
<td>2</td>
<td>0.9458408</td>
<td>0.9346109</td>
<td>0.8523810</td>
<td>0.03399345</td>
<td>0.03592463</td>
<td>0.1192062</td>
</tr>
<tr>
<td>3</td>
<td>0.9447466</td>
<td>0.9297329</td>
<td>0.8523810</td>
<td>0.03302378</td>
<td>0.04016076</td>
<td>0.1192062</td>
</tr>
</tbody>
</table>
Case study 2

Data: Petrostates
Random Forest

<table>
<thead>
<tr>
<th></th>
<th>mtry</th>
<th>ROC</th>
<th>Sens</th>
<th>Spec</th>
<th>ROCSD</th>
<th>SensSD</th>
<th>SpecSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>0.6419646</td>
<td>1.0000000</td>
<td>0</td>
<td>0.08496380</td>
<td>0.00000000</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
<td>0.6577368</td>
<td>0.9968446</td>
<td>0</td>
<td>0.05122113</td>
<td>0.003071956</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>17</td>
<td>0.6619219</td>
<td>0.9957919</td>
<td>0</td>
<td>0.05567687</td>
<td>0.003224936</td>
<td>0</td>
</tr>
</tbody>
</table>

Downsampling

<table>
<thead>
<tr>
<th></th>
<th>mtry</th>
<th>ROC</th>
<th>Sens</th>
<th>Spec</th>
<th>ROCSD</th>
<th>SensSD</th>
<th>SpecSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>0.6789208</td>
<td>0.6529996</td>
<td>0.6516667</td>
<td>0.06018751</td>
<td>0.02529584</td>
<td>0.08022442</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
<td>0.6767277</td>
<td>0.6729960</td>
<td>0.6112500</td>
<td>0.06561405</td>
<td>0.02320574</td>
<td>0.12981543</td>
</tr>
<tr>
<td>3</td>
<td>17</td>
<td>0.6756234</td>
<td>0.6600098</td>
<td>0.6320833</td>
<td>0.06742040</td>
<td>0.02359248</td>
<td>0.13999628</td>
</tr>
</tbody>
</table>
Class imbalance for conflict data

- dependency on magnitude of class imbalance
- correction needed for strong imbalances
- for pre-adjusted data sets correction may actually harm
- balanced design produces stable results
What can we conclude?

- Machine learning classifiers (in particular, random forests) improve prediction accuracy for onset of conflicts.
- Variable importance results are fairly stable and a reasonable alternative to predictor significance in regression models.
- Partial dependence plots enhance interpretability of "causal effects".
- Existing non-linearities in relationships can be easily handled.
- Theoretically existing rare event situations are avoided by sample pre-selection.
- Rare event situations can be tackled by down-/up-sampling.
- Data-driven classifiers are a valuable addition to the tool-kit of the quantitative-oriented social scientist.
- First step towards a paradigmatic shift between explanation, prediction and modeling.
- Wider acceptance of data-driven classifiers in the social sciences needs additional linkage to theory-driven approaches and their results.
Future work?

- Causal Random Forests (Duncan, 2014)
- Mixed-effects random forests for clustered data (Haijem et al., 2014)
  - to address
    - Serial correlation
    - Spatial correlation
    - Clustering
    - Hierarchical data
    - Panel structure
- Further evaluation of class imbalance effects
References:

References:

R packages used:

- library(randomForest) # for random forests
- library(caret) # for CV folds and data splitting
- library(ROCR) # for diagnostics and ROC plots/stats
- library(pROC) # same as ROCR
- library(stepPlr) # Firth’s logit implemented thru caret library
- library(doMC) # for using multiple processor cores
- library(separationplot)
Thank you very much for your attention!

Questions?

Comments?