EXPLAINING DATASETS THROUGH HIGH-ACCURACY REGIONS

Ina Fiterau, Carnegie Mellon University
Artur Dubrawski, Carnegie Mellon University

Work under review at the SIAM Data Mining Conference
OUTLINE

- Motivation of need for interpretability
- Explanation-Oriented Partitioning (EOP)
- Evaluation of EOP
**Example Application: Nuclear Threat Detection**

- Border control: vehicles are scanned
- Human in the loop interpreting results
Boosted Decision Stumps

- Accurate, but hard to interpret

How is the prediction derived from the input?

Image obtained with the Adaboost applet.
**Decision Tree – More Interpretable**

- **Radiation > x%**
  - yes
  - no

- **Payload type = ceramics**
  - yes
  - no

Consider balance of Th232, Ra226 and Co60

- **Uranium level > max. admissible for ceramics**
  - yes
  - no

- **Threat**

- **Clear**


**Motivation**

Many users are willing to trade accuracy to better understand the system-yielded results.

*Need:* simple, interpretable model

*Need:* explanatory prediction process
EXPLANATION-ORIENTED PARTITIONING (EOP)
EXPLANATION-ORIENTED PARTITIONING (EOP) EXECUTION EXAMPLE – 3D DATA
EOP Execution Example – 3D data

Step 1: Select a projection - \((X_1,X_2)\)
EOP Execution Example – 3D data

Step 1: Select a projection - \((X_1,X_2)\)
Step 2: Choose a good classifier - call it $h_1$
Step 2: Choose a good classifier - call it $h_1$
EOP Execution Example – 3D data

Step 3: Estimate accuracy of $h_1$ at each point
Step 3: Estimate accuracy of $h_1$ for each point
Step 4: Identify high accuracy regions
EOP Execution Example – 3D data

Step 4: Identify high accuracy regions
Step 5: Training points - removed from consideration
EOP Execution Example – 3D data

Step 5: Training points - removed from consideration
EOP Execution Example – 3D data

Finished first iteration
Iterate until all data is accounted for or error cannot be decreased
**Learned Model – Processing Query \([x_1x_2x_3]\)**

\([x_1x_2] \text{ in } R_1\) ?

- **no**
  - yes \(h_1(x_1x_2)\)

\([x_2x_3] \text{ in } R_2\) ?

- **no**
  - yes \(h_2(x_2x_3)\)

\([x_1x_3] \text{ in } R_3\) ?

- **no**
  - yes \(h_3(x_1x_3)\)

Default Value
PARAMETRIC REGIONS OF HIGH CONFIDENCE (BOUNDING POLYHEDRA)

- Enclose points in simple convex shapes (multiple per iteration)
  - Grow contour while train error is $\leq \varepsilon$

- Incorrectly classified
- Correctly classified
PARAMETRIC REGIONS OF HIGH CONFIDENCE (BOUNDING POLYHEDRA)

- Enclose points in simple convex shapes (multiple per iteration)
  - Grow contour while train error is \( \leq \varepsilon \)

- Calibration on hold out set - remove shapes that:
  - do not contain calibration points
  - over which the classifier is not accurate
PARAMETRIC REGIONS OF HIGH CONFIDENCE (BOUNDING POLYHEDRA)

- Enclose points in simple convex shapes (multiple per iteration)
  - Grow contour while train error is $\leq \varepsilon$

- Calibration on hold out set - remove shapes that:
  - do not contain calibration points
  - over which the classifier is not accurate

- Intuitive, visually appealing - hyper-rectangles/spheres
OUTLINE

- Motivation of need for interpretability
- Explanation-Oriented Partitioning (EOP)
- Evaluation of EOP
- Summary
Benefits of EOP
- Avoiding Needless Complexity -

Typical XOR dataset
BENEFITS OF EOP - AVOIDING NEEDLESS COMPLEXITY -

CART
- is accurate
- takes many iterations
- does not uncover or leverage structure of data
**Benefits of EOP - Avoiding Needless Complexity**

- **Typical XOR dataset**
  - CART
    - is accurate
    - takes many iterations
    - does not uncover or leverage structure of data
  - EOP
    - equally accurate
    - uncovers structure

![Typical XOR dataset](image1)

![Iteration 1](image2)

![Iteration 2](image3)
COMPARISON TO BOOSTING

What is the price of understandability?

Why boosting?
- It is an [arguably] good black-box classifier
- Learns an ensemble using any type of classifier
- Iteratively targets data misclassified earlier

Criterion: *Complexity of the resulting model*

= number of vector operations to make a prediction
COMPARISON TO BOOSTING - SETUP

- Problem: Binary classification
- 10D Gaussians/uniform cubes for each class
- Statistical significance: repeat experiment with several datasets and compute paired t-test p-values
- Results obtained through 5-fold cross validation
- EOP is often less accurate, but not significantly
- the reduction of complexity is statistically significant

Accuracy p-value: 0.832
Complexity p-value: 0.003
EOP (stumps as base classifiers) vs CART
Data from the UCI repository

- CART is the most accurate
- Parametric EOP yields the simplest models

<table>
<thead>
<tr>
<th>Dataset</th>
<th># of Features</th>
<th># of Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast Tissue</td>
<td>10</td>
<td>1006</td>
</tr>
<tr>
<td>Vowel</td>
<td>9</td>
<td>990</td>
</tr>
<tr>
<td>MiniBOONE</td>
<td>10</td>
<td>5000</td>
</tr>
<tr>
<td>Breast Cancer</td>
<td>10</td>
<td>596</td>
</tr>
</tbody>
</table>
EXPLAINING REAL DATA - SPAMBASE

- 1st Iteration
  - classifier labels everything as spam
  - high confidence regions do enclose mostly spam and
    - Incidence of the word ‘your’ is low
    - Length of text in capital letters is high
**EXPLAINING REAL DATA - SPAMBASE**

- **2nd Iteration**
  - the threshold for the incidence of `your' is lowered
  - the required incidence of capitals is increased
  - the square region on the left also encloses examples that will be marked as `not spam'
EXPLAINING REAL DATA - SPAMBASE

- 3rd Iteration
  - Classifier marks everything as spam
  - Frequency of ‘your’ and ‘hi’ determine the regions
SUMMARY

- EOP maintains classification accuracy but uses less complex models when compared to Boosting

- EOP with decision stumps finds less complex models than CART at the price of a small decrease in accuracy

- EOP gives interpretable high accuracy regions

- We are currently testing EOP in a range of practical application scenarios
THANK YOU
EXTRA RESULTS
EXPLAINING REAL DATA - FUEL