Defining the AIM: An Abstraction for Improving Machine Learning Prediction

Victoria Stodden
University of Illinois at Urbana-Champaign

Symposium on Data Science and Statistics
Reston, VA
May 18 2018
Imagine: Querying the Scholarly Record

1. Show a table of effect sizes and p-values in all phase-3 clinical trials for Melanoma published after 1994;

2. Name all of the image denoising algorithms ever used to remove white noise from the famous “Barbara” image, with citations;

3. List all of the classifiers applied to the famous acute lymphoblastic leukemia dataset, along with their type-1 and type-2 error rates;

4. Create a unified dataset containing all published whole-genome sequences identified with mutation in the gene BRCA1;

5. Randomly reassign treatment and control labels to cases in published clinical trial X and calculate effect size. Repeat many times and create a histogram of the effect sizes. Perform this for every clinical trial published in the 2003 and list the trial name and histogram side by side.

Donoho & Gavish, “Three Dream Applications of Verifiable Computational Results,” CiSE, 2012
The Acute Lymphoblastic Leukemia Dataset

Introduced in Golub et al. “Molecular classification of cancer: class discovery and class prediction by gene expression monitoring” (1999): “cancer classification based on gene expression monitoring by DNA microarrays is described and applied to human acute leukemias [to] discover the distinction between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL)"

In joint work with Xiaomian Wu and April Tang, we tried query 3.
The ALL Dataset Query

We wanted:

• A list of all classifiers applied to the Golub dataset (with citations);

• A comparison of their misclassification rates.

A literature search produced 30 articles, but they did not give comparable misclassification rates.

Our next step was to create the table of comparable misclassification rates. We identified 5 articles for which this seemed possible.
Our (Naive) Expectation

We found that the articles implemented (at least) three steps, each varying from one article to the next:

1. data preprocessing,

2. feature selection,

3. application of machine learning algorithm.
Analysis Steps

Paper 1

Paper 6

50 re-randomization comparison (including a test train split in each iteration) based on the whole process also conducted.

Paper 29

Obs: 38 Pnd: 7129
Train Data
Pre processing
Feature Selection: (Same genes as 1)
Dimension Reduction (PCA/PLS, K=3)
Model Building (LOOCV)
LOOCV Accuracy

Obs: 38 Pnd: 50
Test Data
Classification Result
Test Accuracy

Obs: 38 Pnd: 7129
Train Data
Pre processing
Feature Selection
Model Building (in LOOCV)
LOOCV Accuracy

Obs: 48 Pnd: 3571
Test Data
Classification Result
Test Accuracy

Obs: 38 Pnd: 50
Model Building (in LOOCV)

Obs: 34 Pnd: 50
Classification Result
Test Accuracy

Obs: 34 Pnd: 50
Test Data
Classification Result
Test Accuracy

Obs: 72 Pnd: 7129
Merge Data
Preprocessing
Train/Test split
Test Data
Classification Result
Test Accuracy

Repeat 200 times and using accuracy upper quartile to compare the classifiers.

Paper 9
Learning Algorithms Applied (typically 47ALL; 25AML)

<table>
<thead>
<tr>
<th>Paper</th>
<th>Data Size</th>
<th>Algorithm(s) Applied</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>72 x 6817</td>
<td>Golub Classifier: informative genes+weighted vote</td>
</tr>
<tr>
<td>2</td>
<td>72 x 6817</td>
<td>Golub Classifier: informative genes+weighted vote</td>
</tr>
<tr>
<td>3</td>
<td>72 x 7129</td>
<td>Nearest Neighbor; SVM(linear kernel, quadratic kernel);</td>
</tr>
<tr>
<td>4</td>
<td>72 x 7129</td>
<td>SVM(top 25, 250, 500, 1000 features)</td>
</tr>
<tr>
<td>5</td>
<td>72 x 7070</td>
<td>MVR(median vote relevance); NBGR(naive bayes global relevance);</td>
</tr>
<tr>
<td>6</td>
<td>72 x 6817</td>
<td>Logistic and Quadratic discriminant analysis</td>
</tr>
<tr>
<td>7</td>
<td>72 x 7129</td>
<td>SVM</td>
</tr>
<tr>
<td>9</td>
<td>72 x 6817</td>
<td>Linear and Quadratic discriminant analysis; Classification trees;</td>
</tr>
<tr>
<td>10</td>
<td>72 x 7129</td>
<td>Decision Trees; AdaBoost</td>
</tr>
<tr>
<td>11</td>
<td>72 x 7129</td>
<td>MAVE-LD, DLDA, DQDA, MAVE-NPLD</td>
</tr>
<tr>
<td>12</td>
<td>72 x 7129</td>
<td>SIMCA classification</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>
## Comparable Classification Rates

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Paper1 classifier</td>
<td>0.912</td>
<td>0.941</td>
<td>0.971</td>
<td>0.971</td>
<td>0.882</td>
<td>0.735</td>
</tr>
<tr>
<td>Paper 3 Adaboost</td>
<td>0.912</td>
<td>0.912</td>
<td>0.971</td>
<td>0.971</td>
<td>0.912</td>
<td>0.912</td>
</tr>
<tr>
<td>Paper 3 NN</td>
<td>0.971</td>
<td>0.941</td>
<td>0.912</td>
<td>0.941</td>
<td>0.971</td>
<td>0.971</td>
</tr>
<tr>
<td>Paper 3 SVM Linear</td>
<td>0.971</td>
<td>0.971</td>
<td>0.941</td>
<td>0.971</td>
<td>0.971</td>
<td>0.765</td>
</tr>
<tr>
<td>Paper 3 SVM Quadratic</td>
<td>0.971</td>
<td>0.882</td>
<td>0.971</td>
<td>0.971</td>
<td>0.971</td>
<td>0.912</td>
</tr>
<tr>
<td>Paper 6 logit</td>
<td>0.971</td>
<td>0.971</td>
<td>0.971</td>
<td>0.971</td>
<td>0.971</td>
<td>0.882</td>
</tr>
<tr>
<td>Paper 6 qda</td>
<td>0.941</td>
<td>0.912</td>
<td>0.941</td>
<td>0.971</td>
<td>0.971</td>
<td>0.853</td>
</tr>
<tr>
<td>Paper 9 bagging</td>
<td>0.941</td>
<td>0.912</td>
<td>0.971</td>
<td>0.971</td>
<td>0.912</td>
<td>0.765</td>
</tr>
<tr>
<td>Paper 9 bagging with CPD</td>
<td>0.735</td>
<td>0.853</td>
<td>0.8235</td>
<td>0.912</td>
<td>0.765</td>
<td>0.677</td>
</tr>
<tr>
<td>Paper 9 decision tree</td>
<td>0.912</td>
<td>0.912</td>
<td>0.971</td>
<td>0.971</td>
<td>0.912</td>
<td>0.735</td>
</tr>
<tr>
<td>Paper 9 DLDA</td>
<td>0.971</td>
<td>0.941</td>
<td>0.971</td>
<td>0.971</td>
<td>0.971</td>
<td>0.882</td>
</tr>
<tr>
<td>Paper 9 DQDA</td>
<td>0.971</td>
<td>0.941</td>
<td>0.971</td>
<td>0.971</td>
<td>0.971</td>
<td>0.882</td>
</tr>
<tr>
<td>Paper 9 FLDA</td>
<td>0.882</td>
<td>0.882</td>
<td>0.971</td>
<td>0.971</td>
<td>0.882</td>
<td>0.882</td>
</tr>
<tr>
<td>Paper 9 nn</td>
<td>0.971</td>
<td>0.912</td>
<td>0.853</td>
<td>0.971</td>
<td>0.941</td>
<td>0.941</td>
</tr>
<tr>
<td>Paper 29 Bayesian Network</td>
<td>0.735</td>
<td>0.882</td>
<td>0.971</td>
<td>0.971</td>
<td>0.824</td>
<td>0.618</td>
</tr>
</tbody>
</table>
Learnings..

• Classification rates are hard to synthesize (200+ student hours)

• Many points of variability: starting dataset; preprocessing steps; feature selection methods; algorithm choice; tuning of algorithm and parameters...

• Details not well-captured in the traditional article, making comparisons difficult or impossible.

• Would be easier if:
  ➡ there was prior agreement on the dataset,
  ➡ prior agreement on hold-out data for testing,
  ➡ full disclosure of feature selection steps,
  ➡ full disclosure of algorithm application and parameter tuning.
Conclusions

• *Serious* problems in reporting standards

• Framework for Analysis:
  ➡ Agreement on datasets prior to analysis, conferences around those datasets,
  ➡ Hold-out data held by a neutral third party (e.g. NIST), not seen by researchers,
  ➡ Researchers distinguish and specify feature selection and preprocessing vs learning algorithm application,
  ➡ Send code to the third party who returns your misclassification rate on the test data.

Side effect: training data and code/algorithm shared.

• Results and analysis at https://github.com/AIM-Project/AIM-Manuscript/