Feature Selection in Learning Using Privileged Information

Rauf Izmailov, Blerta Lindqvist, Peter Lin
rizmailov@vencorelabs.com
Phone: 908-748-2891
Agenda

• Learning Using Privileged Information (LUPI)
• Feature Selection Problem
• Our Approach and Results
• Summary
Basic Binary Classification: Problem Formulation

Given training data (observations, facts)

\[(x_1, y_1), \ldots, (x_L, y_L)\]

where \(x \in R^n\) and \(y \in \{-1, +1\}\)

Generalize data to a rule (function \(y = f(x)\))

- normal system
- compromised system

System measurements

Rule separating normal from compromised system

verification

Normal system

Compromised system
Modern machine learning techniques for data analysis problems require construction of decision rules that operate in high dimensional spaces.

- To obtain good decision rules, one has to train learning algorithms using a huge number of data points.
- Meanwhile, humans can learn from a significantly smaller number of training examples.

Why the discrepancy?

Humans use a fundamentally different learning paradigm than machines.
What’s the difference?

**Machine Learning Paradigm**

- Here are some examples of cats
- Here are animals that are not cats

**Human Learning Paradigm**

- Here are some examples of cats
- Here are animals that are not cats

Learn a *decision rule*:

**Input** → Decision Rule → **Output**

- Not a cat

Some additional information about cats:
- Cute
- Tail
- Whiskers

**ADDITIONAL (PRIVILEGED) INFORMATION**
**Learning Using Privileged Information (LUPI)**

- Classical pattern recognition problem: training data and test data are from the same space, with have same attributes etc.

- Given training data (observations, facts) 
  \((x_1, y_1), \ldots, (x_L, y_L)\)

- Generalize data to a rule (function) 
  \[ y = f(x) \]

- where \( x \in X \) and \( y \in \{-1,+1\} \)

- Given training data (observations, facts) and additional privileged data 
  \((x_1^*, \ldots, x_L^*)\)

- Generalize data to a rule (function) 
  \[ y = f(x) \]

- where \( x \in X, x^* \in X^* \) and \( y \in \{-1,+1\} \)

- New paradigm of learning with privileged information: additional information is available **ONLY** with training data, but **NOT** with test data
Learning Using Privileged Information (LUPI)

**Training data:**
- Off-line processing
- High-quality data
- Additional features used as privileged information

**Test data:**
- On-line processing
- Reduced-quality data

**Traditional ML**
- Standard features
- Test data: class -1
- Test data: class +1

**LUPI**
- Standard features
- Privileged features
- Test data: class -1
- Test data: class +1

LUPI uses fundamental asymmetry between training & test data and leverages high-quality privileged information available during training for better performance.

LUPI converges to the solution much faster than alternatives (needs 33 examples instead of 1,000, or 100 instead of 10,000).

Privileged data have the same properties as labels.

**Standard:** digitized X-ray images of patients’ lungs & their classifications

**Privileged:** pathologist’s report of the same pictures

**Output:** presence/absence of cancer

**Decision rule:** cancer identification rule based ONLY on digitized X-ray images
General LUPI Mechanism

**Traditional Machine Learning**
- Rule is learned only on **standard** features.
- Rule works only on **standard** features.
- **Privileged** features, if they exist, are ignored.

**Learning Using Privileged Information**
- Rule learned on both **standard** and **privileged** features.
- **Privileged** features partially learned (approximated) from **standard** ones.
- Works on **standard** features and approximated **privileged** ones.

**Classification algorithm**
### Some Examples of LUPI Applications

<table>
<thead>
<tr>
<th>Privileged information</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Future events</td>
<td>LUPI has been applied to prediction of quasi-chaotic time series (future-in-the-past was privileged information)</td>
</tr>
<tr>
<td>Detailed description of events (semantic information) produced by human experts</td>
<td>LUPI has been applied to image classification (semantic description of images was privileged information)</td>
</tr>
<tr>
<td>Time-consuming probing of data</td>
<td>LUPI has been applied to protein classification (3D protein folding was privileged information)</td>
</tr>
<tr>
<td>Heterogeneous sources of information, some of which may unavailable during test</td>
<td>LUPI has been applied to human detection on a combination of electro-optical and infra-red sensors (one type of sensor was privileged)</td>
</tr>
<tr>
<td>Expensive sensors</td>
<td>LUPI has been applied to human detection on a combination of expensive (high quality) and cheap (low quality) video cameras</td>
</tr>
<tr>
<td>Detailed user behavior information</td>
<td>LUPI has been applied to cyber analytics, where user behavior was privileged information, while traffic measurements were treated as standard information</td>
</tr>
</tbody>
</table>
LUPI Performance Gain: Illustration

Current ML paradigm:
use both available sensors

LUPI paradigm:
use unavailable sensor

Current ML paradigm:
use one available sensor

LUPI Gain
Problem: given amino-acid sequences of proteins; classify them into families of proteins

Assumptions:
- The training data: the space of amino-acid sequences (relatively easy to obtain)
- The privileged information: the space of 3D structures of the proteins (difficult to obtain)

Source:
- SCOP database (structural classification of proteins): sequences and their hierarchical organizations
- 80 superfamilies with the largest number of sequences in each

3D-structures are close

sequences may not be close
Future Events as Privileged Information

- The goal is to predict the development of the system (time series) at the specified time in the future.
- Privileged data include the evolution / trajectory of the system from current moment to the targeted time in the future.
- Given archived data, privileged data can be viewed as “future-in-the-past”.

Data generated by the Mackey-Glass equation:

\[
\frac{dx(t)}{dt} = -ax(t) + \frac{bx(t - \tau)}{1 + x^{10}(t - \tau)},
\]

where \(a\), \(b\), and \(\tau\) (delay) are parameters.

The training triplets:

- \(x_t = (x(t), x(t-1), x(t-2), x(t-3))\)
- \(x_t^* = (x(t + \Delta - 1), x(t + \Delta - 2), x(t + \Delta + 1), x(t + \Delta + 2))\)

**current value and past values**

**values in future (around \(\Delta\))**
Example: EO/IR Monitoring

- EO/IR benchmark dataset from OSU
- 3 paired (EO and IR) surveillance videos in the form of sequential images (total number of frames about 8,000)
- The goal is to detect humans on video
Application of LUPI to Cyber Analytics

- **LUPI-Based Approach:**
  - At training time, collect wide range of observables, including user behavior and related host/network data \(\rightarrow\) **privileged** features used only for training
  - At test time, use a reduced feature set based solely on traffic generated by host that is guaranteed to be observable from outside the host

**Results:** *Order of magnitude higher detection accuracy using LUPI*
Application of LUPI to Target Recognition

- Minor Area Motion Imagery (MAMI) dataset (AFRL, 2014)
- Standard/Privileged features correspond to different resolution/quality of videos
- Consistent LUPI performance advantage for larger training sizes
- Partial LUPI performance advantage for small training size: 80
- Exception: high error rate of standard case and small training size

![Diagram showing LUPI performance improvement over standard approach for different datasets.]
LUPI: Current Status

- LUPI was first introduced about 10 years ago. Initial LUPI framework (called SVM+) was limited just to SVM architecture, and had limited scalability up to 200-300 points in the training dataset (the corresponding matrix was ill-conditioned)
- Current LUPI framework is as scalable as standard classifiers, not restricted to SVM (LUPI works for neural networks, etc.)
- Wide scope of LUPI applications (a few calibration sets already made public).
- Reasonable performance gains of LUPI (20%-50%)
- Open Source version of LUPI (in scikit-learn) has been developed and released within DARPA program PPAML.
- First Workshop on Privileged Information “Beyond Labeler” last year: http://smileclinic.alwaysdata.net/ijcai16workshop/
Agenda

• Learning Using Privileged Information (LUPI)
• Feature Selection Problem
• Our Approach and Results
• Summary
LUPI Feature Selection Problem

- Privileged information does not mean it’s relevant or well-approximated (SSN)
- LUPI feature selection: not one metric, but two:
  - Relevance of the feature
  - Quality of approximation
- Interplay between metrics:
  - Privileged feature can be relevant for classification, but its quality of approximation can be poor
  - Privileged feature can be well approximated but have low relevance for classification
- Both metrics can be expressed as adjusted mutual information and decision to include the feature can be made based on both
- But how to select a reasonable combination of these metrics given scarcity of calibration datasets in order?
Agenda

• Learning Using Privileged Information (LUPI)
• Feature Selection Problem
• Our Approach and Results
• Summary
Feature Engineering Approach

- Some of the observed variables (features) depend on the others according to the corresponding model.
- These dependencies are ignored in data-driven approach.

- If model is known/assumed, a regressed variable in it has two representations:
  - observed view (in data-driven approach),
  - regressed view (on model-driven approach, based on applying regression to observed input values).
Simple Regression Models

- Ideally, we should consider datasets for which **both** elements are known:
  - Observed data (for data-driven approach)
  - Established knowledge-based model (for model-driven approach)
- That ideal approach requires careful selection of candidate datasets and their models:
  - Only a few such datasets can be investigated in a given time
  - Hardening & fine-tuning of the approach will be less reliable

**Observed features**

<table>
<thead>
<tr>
<th>Observed features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Derived features</td>
</tr>
</tbody>
</table>

- Derived model-based features, in this “primitive” regression model, double the number of variables / features
- Each variable has two representations:
  - Observed (what we see in data)
  - Derived (what we should see in the model)

- Primitive regression approximate all possible regression outputs. Some are “noise”:
  - No meaningful regression for a particular variable
  - Regression is obscured by irrelevant variables
<table>
<thead>
<tr>
<th>Datasets</th>
<th>Train</th>
<th>Test</th>
<th>Features</th>
<th>SVM error</th>
<th>SVM-L error</th>
<th>SVM-LN error</th>
<th>Cols-L</th>
<th>Cols-LN</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collins [20]</td>
<td>375</td>
<td>125</td>
<td>22</td>
<td>25.56%</td>
<td>20.12%</td>
<td>18.60%</td>
<td>2.00</td>
<td>4.00</td>
<td>27.23%</td>
</tr>
<tr>
<td>Synthetic_control [23]</td>
<td>480</td>
<td>120</td>
<td>60</td>
<td>0.79%</td>
<td>0.71%</td>
<td>0.58%</td>
<td>31.70</td>
<td>64.00</td>
<td>26.32%</td>
</tr>
<tr>
<td>Parkinsons [24]</td>
<td>175</td>
<td>20</td>
<td>22</td>
<td>5.75%</td>
<td>4.50%</td>
<td>4.25%</td>
<td>9.20</td>
<td>18.40</td>
<td>26.09%</td>
</tr>
<tr>
<td>mu284 [20]</td>
<td>213</td>
<td>71</td>
<td>10</td>
<td>3.31%</td>
<td>3.31%</td>
<td>2.82%</td>
<td>0.45</td>
<td>0.90</td>
<td>14.89%</td>
</tr>
<tr>
<td>fl2000 [20]</td>
<td>50</td>
<td>17</td>
<td>15</td>
<td>20.29%</td>
<td>17.94%</td>
<td>19.71%</td>
<td>2.40</td>
<td>4.80</td>
<td>11.59%</td>
</tr>
<tr>
<td>Prm-synth [20]</td>
<td>187</td>
<td>63</td>
<td>2</td>
<td>14.13%</td>
<td>14.05%</td>
<td>12.62%</td>
<td>2.00</td>
<td>4.00</td>
<td>10.67%</td>
</tr>
<tr>
<td>Wilt [25]</td>
<td>500</td>
<td>4339</td>
<td>5</td>
<td>2.65%</td>
<td>2.50%</td>
<td>2.45%</td>
<td>1.40</td>
<td>2.80</td>
<td>7.55%</td>
</tr>
<tr>
<td>Meta-ensembles [20]</td>
<td>62</td>
<td>12</td>
<td>62</td>
<td>41.67%</td>
<td>41.67%</td>
<td>39.17%</td>
<td>2.65</td>
<td>5.85</td>
<td>6.00%</td>
</tr>
<tr>
<td>fri_c0_100_5 [20]</td>
<td>80</td>
<td>20</td>
<td>5</td>
<td>13.50%</td>
<td>12.75%</td>
<td>13.50%</td>
<td>5.00</td>
<td>10.00</td>
<td>5.56%</td>
</tr>
<tr>
<td>Cloud [20]</td>
<td>81</td>
<td>27</td>
<td>7</td>
<td>46.48%</td>
<td>45.74%</td>
<td>44.44%</td>
<td>1.00</td>
<td>2.00</td>
<td>4.38%</td>
</tr>
<tr>
<td>Spectrometer [20]</td>
<td>398</td>
<td>133</td>
<td>101</td>
<td>45.04%</td>
<td>43.91%</td>
<td>43.27%</td>
<td>92.90</td>
<td>185.80</td>
<td>3.92%</td>
</tr>
<tr>
<td>Sonar [26]</td>
<td>187</td>
<td>21</td>
<td>60</td>
<td>13.10%</td>
<td>12.38%</td>
<td>12.62%</td>
<td>1.55</td>
<td>3.10</td>
<td>3.66%</td>
</tr>
<tr>
<td>ar4 [20]</td>
<td>90</td>
<td>17</td>
<td>29</td>
<td>16.47%</td>
<td>15.88%</td>
<td>16.47%</td>
<td>0.80</td>
<td>1.60</td>
<td>3.57%</td>
</tr>
<tr>
<td>Glass [27]</td>
<td>160</td>
<td>54</td>
<td>10</td>
<td>8.98%</td>
<td>8.80%</td>
<td>8.70%</td>
<td>0.35</td>
<td>0.85</td>
<td>3.12%</td>
</tr>
<tr>
<td>Analcatdata-wildcat [20]</td>
<td>130</td>
<td>33</td>
<td>5</td>
<td>22.88%</td>
<td>22.27%</td>
<td>22.73%</td>
<td>1.70</td>
<td>3.20</td>
<td>2.65%</td>
</tr>
<tr>
<td>pc1 [28]</td>
<td>831</td>
<td>278</td>
<td>21</td>
<td>6.83%</td>
<td>6.69%</td>
<td>6.69%</td>
<td>2.95</td>
<td>2.90</td>
<td>2.11%</td>
</tr>
<tr>
<td>Pasture [20]</td>
<td>30</td>
<td>6</td>
<td>22</td>
<td>40.00%</td>
<td>39.17%</td>
<td>39.17%</td>
<td>7.80</td>
<td>16.15</td>
<td>2.08%</td>
</tr>
<tr>
<td>kc2 [28]</td>
<td>417</td>
<td>105</td>
<td>21</td>
<td>16.95%</td>
<td>17.90%</td>
<td>16.62%</td>
<td>4.15</td>
<td>8.30</td>
<td>1.97%</td>
</tr>
<tr>
<td>kc1 [28]</td>
<td>500</td>
<td>1609</td>
<td>21</td>
<td>15.28%</td>
<td>15.24%</td>
<td>15.18%</td>
<td>3.85</td>
<td>6.10</td>
<td>0.67%</td>
</tr>
<tr>
<td>chsclass_census [20]</td>
<td>300</td>
<td>100</td>
<td>6</td>
<td>41.85%</td>
<td>41.60%</td>
<td>41.60%</td>
<td>3.30</td>
<td>6.60</td>
<td>0.60%</td>
</tr>
<tr>
<td>Magic telescope [20]</td>
<td>700</td>
<td>18320</td>
<td>10</td>
<td>15.95%</td>
<td>15.90%</td>
<td>15.97%</td>
<td>2.35</td>
<td>4.50</td>
<td>0.30%</td>
</tr>
<tr>
<td>Pollen [20]</td>
<td>500</td>
<td>3348</td>
<td>5</td>
<td>50.58%</td>
<td>50.54%</td>
<td>50.45%</td>
<td>2.50</td>
<td>5.00</td>
<td>0.26%</td>
</tr>
<tr>
<td>Tamilnadu-electricity [20]</td>
<td>700</td>
<td>45081</td>
<td>3</td>
<td>32.46%</td>
<td>32.44%</td>
<td>32.39%</td>
<td>1.75</td>
<td>3.50</td>
<td>0.22%</td>
</tr>
<tr>
<td>Disclosure [29]</td>
<td>496</td>
<td>166</td>
<td>3</td>
<td>44.43%</td>
<td>44.34%</td>
<td>44.43%</td>
<td>3.00</td>
<td>6.00</td>
<td>0.20%</td>
</tr>
<tr>
<td>Steel plates fault [20]</td>
<td>500</td>
<td>14</td>
<td>33</td>
<td>0.14%</td>
<td>0.16%</td>
<td>0.12%</td>
<td>2.15</td>
<td>4.30</td>
<td>0.13%</td>
</tr>
<tr>
<td>Blood transfusion [30]</td>
<td>598</td>
<td>150</td>
<td>4</td>
<td>24.93%</td>
<td>24.90%</td>
<td>24.90%</td>
<td>0.20</td>
<td>0.20</td>
<td>0.13%</td>
</tr>
<tr>
<td>Mushroom [19]</td>
<td>500</td>
<td>7624</td>
<td>21</td>
<td>34.58%</td>
<td>34.58%</td>
<td>34.58%</td>
<td>1.00</td>
<td>2.00</td>
<td>1.64%</td>
</tr>
<tr>
<td>Connect-4 [20]</td>
<td>500</td>
<td>67057</td>
<td>42</td>
<td>28.79%</td>
<td>28.79%</td>
<td>28.79%</td>
<td>0.70</td>
<td>1.40</td>
<td>1.40%</td>
</tr>
<tr>
<td>Ada-agnostic [20]</td>
<td>500</td>
<td>4062</td>
<td>48</td>
<td>21.94%</td>
<td>21.94%</td>
<td>21.94%</td>
<td>3.60</td>
<td>7.20</td>
<td>3.60%</td>
</tr>
<tr>
<td>Cardiotocography [31]</td>
<td>500</td>
<td>1626</td>
<td>35</td>
<td>0.17%</td>
<td>0.17%</td>
<td>0.17%</td>
<td>0.10</td>
<td>0.20</td>
<td>0.10%</td>
</tr>
<tr>
<td>Ecoli [20]</td>
<td>252</td>
<td>84</td>
<td>7</td>
<td>16.90%</td>
<td>16.90%</td>
<td>16.90%</td>
<td>0.25</td>
<td>0.50</td>
<td>0.25%</td>
</tr>
<tr>
<td>Kr-vs-kp [20]</td>
<td>500</td>
<td>2696</td>
<td>36</td>
<td>4.07%</td>
<td>4.07%</td>
<td>4.07%</td>
<td>1.20</td>
<td>2.40</td>
<td>1.20%</td>
</tr>
<tr>
<td>Triazines [20]</td>
<td>139</td>
<td>47</td>
<td>60</td>
<td>28.72%</td>
<td>28.72%</td>
<td>28.72%</td>
<td>2.30</td>
<td>4.30</td>
<td>2.30%</td>
</tr>
<tr>
<td>Datatrieeve [28]</td>
<td>110</td>
<td>20</td>
<td>8</td>
<td>9.75%</td>
<td>9.75%</td>
<td>9.75%</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05%</td>
</tr>
<tr>
<td>Meta-batchincremental [20]</td>
<td>62</td>
<td>12</td>
<td>62</td>
<td>35.00%</td>
<td>35.00%</td>
<td>35.00%</td>
<td>2.70</td>
<td>5.95</td>
<td>2.70%</td>
</tr>
</tbody>
</table>
UCI Machine Learning Repository: Parkinsons, WDBC

**SVM** Standard error rate
**SVM** Privileged error rate
**SMV** LUPI error

\[
\frac{S - L}{S - P}
\]

performance gap recovered by LUPI
Agenda

- Learning Using Privileged Information (LUPI)
- Feature Selection Problem
- Our Approach and Results
- Summary
Summary

- LUPI feature selection problem was identified
- A two-metric approach to LUPI feature selection problem was proposed and tested
- A LUPI-like feature engineering and selection process for non-LUPI learning problems was formulated and tested
- More refined and robust methods of both constructing and selecting features in LUPI framework will be studied next
TRANSFORMATIVE RESEARCH