Learning Using Privileged Information

Rauf Izmailov
rizmailov@vencorelabs.com
Phone: 908-748-2891
Basic Binary Classification: Notations

- **labels**
  - positive class
  - negative class

- **N dimensions (features, variables)**

- **L samples**
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

\[
\begin{align*}
\text{Normal system} & \quad \text{Compromised system} \\
+ & \quad -1
\end{align*}
\]
There are two and only two factors responsible for induction:
1) Percentage of errors on the training sample (trivial factor)
2) Capacity (VC dimension) of the set of rules (non-trivial factor).

\[ R(f) \leq R_{\text{emp}}(f) + O\left(\sqrt{\frac{h}{L}}\right) \]

Confidence term:
- Increases with capacity (VC dimension) \( h \)
- Decreases with sample size \( L \)
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

**Learn a decision rule:**

**Human Learning Paradigm**

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

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

---

**ADDITIONAL (PRIVILEGED) INFORMATION**

- Not a cat
Model-Driven and Data-Driven Approaches

**Problem of Model-driven approach:**

- **Statistical method:** model is specified → estimate parameters → construct decision
- **Advantages:**
  - Well-established mathematical tools to convert models into decision rules.
- **Disadvantages:**
  - Actual system structure & distributions may differ from simplified model
  - Parameters may be hard to estimate (non-convex problem, insufficient data)
- **Problem:** How to leverage data-driven information in model-driven setting?

**Problem of Data-driven approach:**

- **Optimization method:** model is not specified => find the best approximation function
- **Advantages:**
  - Direct method (one hop versus two hops in model-driven approach)
  - Typically better performance than model-driven approach
- **Disadvantages:**
  - Ignores domain knowledge information on real structures
- **Problem:** How to leverage model-driven information in data-driven approach?
Learning Using Privileged Information (LUPI)

- 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 \(y \in \{-1, +1\}\)

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

- New paradigm of learning with privileged information: additional information is available **ONLY** with training data, but **NOT** with test data

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

- 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 \(y \in \{-1, +1\}\)
Why Bother With LUPI?

Traditional ML:
$$R(f) \leq R_{\text{emp}}(f) + O\left(\frac{h}{\sqrt{L}}\right)$$

LUPI:
$$R(f) \leq R_{\text{emp}}(f) + O\left(\frac{h}{L}\right)$$

Bayes (infinite sample size):
$$\lim_{L \to \infty} R_{\text{Bayes}} = R_{\text{emp}}$$

LUPI approach:
Add privileged information

Traditional approach:
Significantly increase sample size

- SVM converges (as $1/\sqrt{L}$) to the default Bayes rule as sample size $L$ goes to infinity
- In reality, almost all samples are small; additional samples are expensive/impossible to get
- LUPI uses privileged info instead of additional samples to converge much faster (as $1/L$)
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
- Class -1
- Class +1

**LUPI**
- Standard features
- Privileged features

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:** hand-written digits and their classification

**Privileged:** Semantic description of digits shapes

**Output:** classification of pixel image into 0,1,2,...,9

**Decision rule:** classification based **ONLY** on pixel image.
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

Applying decision rule

Knowledge transfer
LUPI Mechanism for Neural Networks

ANN on privileged space

ANN on decision space

Knowledge Transfer
\[ \varphi_1(x^1, x^2, x^3) \leftarrow x^*1 \]
\[ \varphi_2(x^1, x^2, x^3) \leftarrow x^*2 \]
### 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 checp (low quality) video cameras</td>
</tr>
</tbody>
</table>
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

Advanced Model as Privileged Information
## Classification or Protein Families

<table>
<thead>
<tr>
<th>Protein superfamily pair</th>
<th>SVM</th>
<th>LUPI</th>
<th>SVM (full)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a.26.1-vs-c.68.1</td>
<td>7.3</td>
<td>7.3</td>
<td>0</td>
</tr>
<tr>
<td>a.26.1-vs-g.17.1</td>
<td>16.4</td>
<td>14.3</td>
<td>0</td>
</tr>
<tr>
<td>a.118.1-vs-b.82.1</td>
<td>19.2</td>
<td>6.4</td>
<td>0</td>
</tr>
<tr>
<td>a.118.1-vs-d.2.1</td>
<td>41.5</td>
<td>24.5</td>
<td>3.8</td>
</tr>
<tr>
<td>a.118.1-vs-d.14.1</td>
<td>13.1</td>
<td>13.1</td>
<td>2.2</td>
</tr>
<tr>
<td>a.118.1-vs-e.8.1</td>
<td>22.8</td>
<td>2.3</td>
<td>2.3</td>
</tr>
<tr>
<td>b.1.18-vs-b.55.1</td>
<td>14.6</td>
<td>13.5</td>
<td>0</td>
</tr>
<tr>
<td>b.18.1-vs-b.55.1</td>
<td>31.5</td>
<td>15.1</td>
<td>0</td>
</tr>
<tr>
<td>b.18.1-vs-c.55.1</td>
<td>36.2</td>
<td>36.2</td>
<td>0</td>
</tr>
<tr>
<td>b.18.1-vs-c.55.3</td>
<td>38.1</td>
<td>36.6</td>
<td>0</td>
</tr>
<tr>
<td>b.18.1-vs-d.92.1</td>
<td>25</td>
<td>11.8</td>
<td>0</td>
</tr>
<tr>
<td>b.29.1-vs-b.30.5</td>
<td>16.9</td>
<td>16.9</td>
<td>3.6</td>
</tr>
<tr>
<td>b.29.1-vs-b.55.1</td>
<td>10</td>
<td>5.5</td>
<td>0</td>
</tr>
<tr>
<td>b.29.1-vs-b.80.1</td>
<td>8.3</td>
<td>5.9</td>
<td>0</td>
</tr>
<tr>
<td>b.29.1-vs-b.121.4</td>
<td>35.9</td>
<td>16.8</td>
<td>5.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Protein superfamily pair</th>
<th>SVM</th>
<th>LUPI</th>
<th>SVM (full)</th>
</tr>
</thead>
<tbody>
<tr>
<td>b.29.1-vs-c.47.1</td>
<td>10.2</td>
<td>3.2</td>
<td>0</td>
</tr>
<tr>
<td>b.30.5-vs-b.80.1</td>
<td>43.3</td>
<td>6.7</td>
<td>0</td>
</tr>
<tr>
<td>b.30.5-vs-b.55.1</td>
<td>25.5</td>
<td>14.6</td>
<td>0</td>
</tr>
<tr>
<td>b.55.1-vs-b.82.1</td>
<td>11.8</td>
<td>10.3</td>
<td>0</td>
</tr>
<tr>
<td>b.55.1-vs-d.14.1</td>
<td>20.9</td>
<td>19.4</td>
<td>0</td>
</tr>
<tr>
<td>b.55.1-vs-d.15.1</td>
<td>17.7</td>
<td>12.7</td>
<td>0</td>
</tr>
<tr>
<td>b.80.1-vs-b.82.1</td>
<td>4.7</td>
<td>4.7</td>
<td>0</td>
</tr>
<tr>
<td>b.82.1-vs-b.121.4</td>
<td>7.9</td>
<td>3.4</td>
<td>0</td>
</tr>
<tr>
<td>b.121.4-vs-d.14.1</td>
<td>29.5</td>
<td>23.9</td>
<td>0</td>
</tr>
<tr>
<td>b.121.4-vs-d.92.1</td>
<td>15.3</td>
<td>9.2</td>
<td>0</td>
</tr>
<tr>
<td>c.36.1-vs-c.68.1</td>
<td>8.9</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>c.36.1-vs-e.8.1</td>
<td>12.8</td>
<td>2.2</td>
<td>0</td>
</tr>
<tr>
<td>c.47.1-vs-c.69.1</td>
<td>1.9</td>
<td>0.6</td>
<td>0</td>
</tr>
<tr>
<td>c.52.1-vs-b.80.1</td>
<td>11.8</td>
<td>5.9</td>
<td>0</td>
</tr>
<tr>
<td>c.55.1-vs-c.55.3</td>
<td>45.1</td>
<td>28.2</td>
<td>22.5</td>
</tr>
</tbody>
</table>

3D structure is essential for classification; SVM+ does not improve classification of SVM

SVM+ provides significant improve improvement over SVM (several times)
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:

\(x_t = (x(t), x(t - 1), x(t - 2), x(t - 3))\)

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

\[x_t\] current value and past values

\[\*x_t\] values in future (around \(\Delta\))
LUPI for Decision Making with Unreliable Sensors

If one of the sensors is down, the corresponding features are not available.
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
Each 36*36 image was converted to a string of HOG features using partitions of multiple granularities: 2*2, 4*4, and 6*6:

For each image, two 504-dimensional vectors of features were computed: one for IR, another for EO:
LUPI Application for Missing Data

Current ML paradigm: use both *available* sensors

Current ML paradigm: use one *available* sensor

LUPI paradigm: use *unavailable* sensor
Classification Decision Results

Current ML paradigm: use both available sensors

- Error rate: 14.62%

LUPI paradigm: use unavailable sensor

- Error rate: 16.10%

Current ML paradigm: use one available sensor

- Error rate: 16.85%

Error rate reduction by 33%
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 → *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 Image Classification

- Data: 20 classes of objects (cows, bicycles, dogs, cars, cats, etc.)
- Objects downloaded, extracted and mapped to HOG features
- Selected two (arguably, most difficult in terms of separation) classes (motorcycles and bicycles)
- Effect of privileged features emulated by obscuring half of the image

The PASCAL Visual Object Classes Challenge

Motorcycles versus Bicycles

LUPI is worse

LUPI is better
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*

![Graph showing LUPI performance comparison]

- Standard
- Privileged

LUPI is better
LUPI is worse
LUPI: Summary

• 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) is being developed within a DARPA program and to be released soon.
• First Workshop on Privileged Information “Beyond Labeler” last year: http://smileclinic.alwaysdata.net/ijcai16workshop/
• Active on-going research in privileged information:
  – Mechanisms for approximation of privileged information
  – Mechanisms for selection/filtering of privileged information
  – Expansion to non-supervised learning applications
LUPI References


This material is based upon work partially supported by AFRL under contract FA9550-15-1-0502. The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.
TRANSFORMATIVE RESEARCH