The Adaptive Data Cube & ISP Decision Trees

Carey E. Priebe
cerp@jhu.edu

Department of Applied Mathematics & Statistics
Department of Computer Science
Center for Imaging Science

Johns Hopkins University

DARPA ISP
October 13-14, 2004
Alexandria, Virginia
• sensing —

\[ s : \mathcal{R} \rightarrow \mathbb{R}^p, \ p >> 1 \]

• dimension reduction —

\[ r : \mathbb{R}^p \rightarrow \mathbb{R}^q, \ q << p \]

• classification/clustering —

\[ g : \mathbb{R}^q \rightarrow [0, 1]^C \]
Trunk (1979)

![Graph showing the relationship between log_10(Dimension) and Probability of Error for Bayes Optimal Classifier (o = Theoretical; + = Monte Carlo) and Linear Classifier, n=2 (o = Theoretical; + = Monte Carlo).]
Fisher’s Conditionality Principle

Foundations of Statistical Inference:
Likelihood Principle, Sufficiency Principle, Conditionality Principle

Fisher’s Conditionality Principle (1950,1956)

But: Welch (1939)?
Consider Amari’s 1985 statement of the Conditionality Principle:

"When there exists an exact ancillary statistic \( r \), the conditionality principle requires that the statistical inference should be performed by conditioning on \( r \). ..."

Shun-ichi Amari,
Differential Geometric Methods in Statistics,
Amari continues ...

(inference about \( u \) is the goal)

"... A statistical problem then is decomposed into subproblems in each of which \( r \) is fixed at its observed value, thus dividing the whole set of the possible data points into subclasses. It is expected that each subclass consists of relatively homogeneous points with respect to the informativeness about \( u \). We can then evaluate our conclusion about \( u \) based on \( r \), and it gives a better evaluation than the overall average one. This is a way of utilizing information which ancillary \( r \) conditionally carries."
Most of the basic work in classifier construction (Random Forests, Support Vector Machines, etc.) focuses single-mindedly on classification at all times. That is, a feature’s value is in direct proportion with its utility in estimating the true but unknown class label. Little or no use is made of features ancillary to this task.

See *The X-property* (D-G-L, 1996, p. 319) for a general exception:

“[T]he form of the tree is determined by the [feature vectors] only, that is, the labels . . . do not play a role in constructing the partition, but they are used in voting.”

Example
(from Priebe et al., PAMI ‘04)

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim$$

$$\begin{pmatrix} X_3 \\ X_4 \end{pmatrix} \sim$$

$$\begin{pmatrix} X_5 \\ X_6 \end{pmatrix} \sim$$
Example
(from Priebe et al., PAMI ‘04)

\[
\begin{pmatrix}
X_1 \\
X_2
\end{pmatrix} \sim
\begin{pmatrix}
X_3 \\
X_4
\end{pmatrix} \sim
\begin{pmatrix}
X_5 \\
X_6
\end{pmatrix}
\]

\begin{align*}
Y &= 0 \\
Y &= 1
\end{align*}

\frac{1}{2} \quad \frac{1}{2}
Example
(from Priebe et al., PAMI ‘04)
In this pedagogical example, a willingness to consider partitioning based on *ancillary features* when constructing the tree yields a superior solution: the iterative denoising tree.

Theorem

Given $\epsilon > 0$, we construct $F_{XY}$ with feature vectors $X = [X_1, \cdots, X_d]' \in \mathbb{R}^{d=d(\epsilon)}$ and class labels $Y \in \{0, 1\}$ such that for $(X, Y) \sim F_{XY}$

$$
\min_{g: \mathbb{R}^d \to \{0,1\}} \min_{i,j} P[g(X_i, X_j) \neq Y] \geq \frac{1}{2} - \epsilon
$$

while

$$
\exists g \text{ with } P[g(X_1, X_{X_1}) \neq Y] = 0,
$$

and $X_1$ is ancillary for the classification task at hand.

That is, there is no pair of features $X_i, X_j$ which work, while conditioning on the ancillary $X_1$ — using $X_1, X_{X_1}$ — works.

Theorem

Given $\epsilon > 0$, we construct $F_{XY}$ with feature vectors $X = [X_1, \cdots, X_d]' \in \mathbb{R}^{d=d(\epsilon)}$ and class labels $Y \in \{0, 1\}$ such that for $(X, Y) \sim F_{XY}$

$$\min_{g: \mathbb{R}^d \to \{0,1\}} \min_{i,j} P[g(X_i, X_j) \neq Y] \geq \frac{1}{2} - \epsilon$$

while

$$\exists g \text{ with } P[g(X_1, X_{X_1}) \neq Y] = 0,$$

and $X_1$ is ancillary for the classification task at hand.

That is, there is no pair of features $X_i, X_j$ which work, while conditioning on the ancillary $X_1$ — using $X_1, X_{X_1}$ — works.

This theorem illustrates the potential benefit of a willingness to consider conditionality / partitioning / $X$-property methods.
But there is an added benefit ...

(huge bags of cash, only $2.98 each)
Corpus-Dependent Feature Extraction

CDFE:
The features —
    the features sensed, as with an adaptive sensor, or
    the features extracted, as with dimensionality reduction —
depend on the collection of entities under consideration.

Appropriate denoising can allow the new “local” features,
extracted at internal nodes of the tree,
to be considerably more valuable than their “global” brethren.
A Hyperspectral Imaging Example
HYMAP airborne hyperspectral scanner

Characteristics

Number of Bands: 126
Wavelength: 450 - 2500 nm
Spatial resolution: varies
Spectral resolution: 15 - 20 nm
a HYMAP hyperspectral image
a HYMAP hyperspectral image
HYMAP hyperspectral imaging example I

black: runway, red: pine, green: scrub
## HYMAP hyperspectral imaging example II

<table>
<thead>
<tr>
<th></th>
<th>original</th>
<th>denoised</th>
</tr>
</thead>
<tbody>
<tr>
<td>misclassification of scrub as pine</td>
<td>367</td>
<td>284</td>
</tr>
<tr>
<td>misclassification of pine as scrub</td>
<td>363</td>
<td>293</td>
</tr>
<tr>
<td>sum</td>
<td>730</td>
<td>577 (+ 10)</td>
</tr>
</tbody>
</table>

**HYMAP Misclassification Errors**

|pine| = 2236 ; |scrub| = 2284

**Note:** 10 scrub observations are misclassified (as runway) via the denoising partition.
HYMAP hyperspectral imaging example III

Distance dimension vs. misclassification rate.

- Original
- Denoised

 cep@jhu.edu
 DARPA ISP, Oct 2004  -- p.25/50
Glimpses of Genius

Mathematicians and historians piece together a puzzle that Archimedes pondered

Erica Klarreich

At the start of the 20th century, a Danish mathematical historian named Johan Ludvig Heiberg made a once-in-a-lifetime find. Tucked away in the library of a monastery in Istanbul was a medieval parchment containing copies of the works of the ancient Greek mathematician Archimedes, including two never-before-seen essays. To mathematicians’ astonishment, one of the new essays contained many of the key ideas of calculus, a subject supposedly invented two millennia after Archimedes' time. The essay caused a sensation and landed Heiberg’s discovery on the front page of a 1907 New York Times.

The other new essay, by contrast, mystified mathematicians. A fragment of a treatise called the Stomachion, it appeared to be nothing more than a description of a puzzle that might have been a children’s toy. Mathematicians wondered why Archimedes, whose other works were so monumental, should have spent his time on something so frivolous.

The Stomachion fragment offered only tantalizing glimpses into Archimedes thinking. The parchment, probably first written on in the 10th century in Constantinople, is a palimpsest—a document whose surface has been scraped and reused. In 1204, the Fourth Crusade sacked Constantinople, and shortly thereafter monks unbound the Archimedes’s parchment, did their best to erase the mathematical text, and recycled the volume as a Christian prayer book. Only the beginning pages of the Stomachion made it into the new book, and on those pages, the underlying math text was faint and hard to read.

Armed with only a magnifying glass, Heiberg managed to make out a large portion of the palimpsest. What he read, however, offered few clues to Archimedes’ interest in the puzzle. And before other scholars could examine the faded script, and perhaps catch something Heiberg had missed, the parchment was stolen. It vanished into obscurity for more than 8 decades.

In 1998, the Archimedes palimpsest suddenly resurfaced from a private collection. An anonymous U.S. billionaire bought it for $2 million at Christie’s auction house in New York and made it available to scholars. Mathematical historians have now come up with a guess about what Archimedes was thinking 2,200 years ago.

Mathematicians are also studying the puzzle for its own sake. Its underlying structure, they are finding, is anything but trivial.
Example: Lin & Pantel Mutual Information Feature

\[ \mathcal{L}_C(\cdot) : \text{DocumentSpace} \rightarrow [\text{Mutual Information Feature}]^{d_L(C)}. \]

Both the features themselves and the number of features \( d_L(C) \) depend on the corpus \( C \). Thus \( \mathcal{L}_C(C) \) is a \( |C| \times d_L(C) \) \textit{mutual information feature matrix}. Each of the features is associated with a word (after stemming and removal of stopper words), as follows. For document \( x \) in corpus \( C \), and associated word \( w \), the mutual information between \( x \) and \( w \) is given by

\[
m_{x,w} = \log \left( \frac{f_{x,w}}{\sum_{\xi \in C} f_{\xi,w} \sum_{\omega} f_{x,\omega}} \right).
\]

Here \( f_{x,w} = c_{x,w}/N \) where \( c_{x,w} \) is the number of times word \( w \) appears in document \( x \) and \( N \) is the total number of words in the corpus \( C \).
Example: Lin & Pantel Mutual Information Feature

\[ \mathcal{L}_C(\cdot) : \text{DocumentSpace} \to [\text{MutualInformationFeature}]^{d_L(C)}. \]

Both the features themselves and the number of features \( d_L(C) \) depend on the corpus \( C \). Thus \( \mathcal{L}_C(C) \) is a \( |C| \times d_L(C) \) mutual information feature matrix. Each of the features is associated with a word (after stemming and removal of stopper words), as follows. For document \( x \) in corpus \( C \), and associated word \( w \), the mutual information between \( x \) and \( w \) is given by

\[
m_{x,w} = \log \left( \frac{f_{x,w}}{\sum_{\xi \in C} f_{\xi,w} \sum_{\omega} f_{x,\omega}} \right).
\]

Here \( f_{x,w} = c_{x,w}/N \) where \( c_{x,w} \) is the number of times word \( w \) appears in document \( x \) and \( N \) is the total number of words in the corpus \( C \).
black: Astronomy, red: Mathematics, green: Medicine, blue: Physics
### Initial Science News Denoising Partition

<table>
<thead>
<tr>
<th>branch</th>
<th>Astronomy</th>
<th>Math</th>
<th>Medicine</th>
<th>Physics</th>
</tr>
</thead>
<tbody>
<tr>
<td>left branch</td>
<td>0</td>
<td>2</td>
<td>272</td>
<td>2</td>
</tr>
<tr>
<td>center branch</td>
<td>16</td>
<td>58</td>
<td>8</td>
<td>109</td>
</tr>
<tr>
<td>right branch</td>
<td>105</td>
<td>0</td>
<td>0</td>
<td>7</td>
</tr>
</tbody>
</table>

|Math|=60 ; |Physics|=118

Note: 7 Physics documents are misclassified as Astronomy; 2 Physics and 2 Math documents are misclassified as Medicine
### Science News Misclassification Errors

<table>
<thead>
<tr>
<th>misclassification</th>
<th>original</th>
<th>denoised</th>
</tr>
</thead>
<tbody>
<tr>
<td>misclassification of Physics as Math</td>
<td>42</td>
<td>9</td>
</tr>
<tr>
<td>misclassification of Math as Physics</td>
<td>35</td>
<td>11</td>
</tr>
<tr>
<td>sum</td>
<td>77</td>
<td>20 (+11)</td>
</tr>
</tbody>
</table>

| |Math|=60 ; |Physics|=118
An information-theoretic criterion
for unsupervised iterative denoising tree construction . . .

(Joint work with Damianos Karakos & Sanjeev Kudanpur.)
Total Score: \( I(Y; Z) = I(Y; Z_1) + I(Y; Z_2|Z_1) + I(Y; Z_3|Z_2, Z_1) \)
Information-Theoretic Iterative Denoising

Mutual Information as the Goodness Criterion for ISPDT Construction

\( N \) objects: \( X^n(1), \ldots, X^n(N) \)

Transform to \( \tilde{X}^n(1), \ldots, \tilde{X}^n(N) \).

... and cluster.

Transformation & Clustering Score:

\[
S(c, c_0, c_1) = \frac{N_0}{N} D(c_0 || c) + \frac{N_1}{N} D(c_1 || c)
\]
Optimality of Mutual Information as the Goodness Criterion

\[
S(c, c_0, c_1) = \frac{N_0}{N} D(c_0 \| c) + \frac{N_1}{N} D(c_1 \| c)
\]

\[N, n \rightarrow \infty \quad I(Y; Z)\]

(for stationary and ergodic data)
Information-Theoretic Iterative Denoising

Total Score:
\[ S(C, A_0, A_1) + S(A_1, A_{10}, A_{11}) \times \frac{N_1}{N} + S(A_{11}, A_{110}, A_{111}) \times \frac{N_{11}}{N} \]

Goal: Maximize Total Score
Information-Theoretic Iterative Denoising

Total Score:
\[ S(C, A_0, A_1) + S(A_1, A_{10}, A_{11}) \times \frac{N_1}{N} + S(A_{11}, A_{110}, A_{111}) \times \frac{N_{11}}{N} \]

Nice Property: Iterative Computation
Information-Theoretic Iterative Denoising

Total Score:
\[ S(C, A_0, A_1) + S(A_1, A_{10}, A_{11}) \times \frac{N_1}{N} + S(A_{11}, A_{110}, A_{111}) \times \frac{N_{11}}{N} \]

Score depends only on local features: CDFE
Information-Theoretic Iterative Denoising Example
In the first example (hyperspectral imaging),
the gain from cdfe was (simply) due to dimension reduction.

In the text document examples,
new corpus-dependent features were indeed extracted.

But there was no *re-sensing* involved.

An adaptive sensor example?
   Real-time tunable hyperspectral imagers!
Architecture of the Digital Micro-Mirror Array (DMA) used as a switching device for combining spatio-spectral features

Mirrors rotate along their diagonal axis by exactly ±10°, which is what makes the DMA a digital devise

- 0 = no light reflected (-10°)
- 1 = all light reflected (+10°)

A complete DMA contains 848 columns and 600 rows of mirrors and measures 10.2 mm x 13.6 mm. Here, a DMA is shown with its glass cover removed.
Runway: 100%

Pine: 3%
Scrub: 96%

Pine: 97%
Scrub: 4%

Prof. Richard Granger
Director: Brain Engineering Laboratory
Computer Science and Cognitive Science
University of California, Irvine
Key features of anatomical organization of thalamocortical circuits to be modeled.
Sequential operations at three successive time steps (t = 1, 2, 3) in middle (M), superficial (S), and deep (D) cortical layers, core and matrix thalamic projections (Ct & Mt), and inhibitory nucleus reticularis thalami (NRt-), in response to a static input.
“The wealth of your practical experience with sane and interesting problems will give to mathematics a new direction and a new impetus.”

– Leopold Kronecker to Hermann von Helmholtz –