Capturing imaging patterns via machine learning and classification

Christos Davatzikos

Professor of Radiology, Electrical and Systems Engineering, and Bioengineering
Director, Section of Biomedical Image Analysis

http://www.rad.upenn.edu/sbia
Focal targets: easy to detect and measure
Detecting spatially complex very subtle anatomical abnormalities

Healthy
Mildly Cognitively impaired: Prodromal stage to Alzheimer’s
Functional activity during truth telling and lying

Lies

Truths
Pathological patterns are overlaid on normative patterns
Challenge: Sensitive and Specific Biomarkers for Individuals rather than groups.

Data from the ADNI study (Davatzikos et.al., NeuroImage, 2007, 2008)
A pattern is sampled by measuring imaging information (structural/functional) in a number of brain regions.
- Robust voxel-based statistics are used to determine candidate features/regions
- **Watershed** obtains a number of spatial clusters (brain regions)
- Forward and reverse *feature elimination* are used to reduce the feature set
- **Bagging** (bootstrap aggregation) is used to obtain robust and generalizable classification
- Support vector machines for final classification

Available under http://www.rad.upenn.edu/sbia

*Cite: Fan et.al., IEEE-TMI, 2007*
ADNI Study: Conversion from MCI to AD

Davatzikos et.al., to appear in Neurobiology of Aging

-- 239 MCI patients (75.2±7.3 y.o.)

-- average 12 months follow-up with a standard deviation of 6 months

-- 69 MCI-C (76.86±6.88y.o) and 170 MCI-NC (74.47±7.35)

-- mean baseline MMSE: MCI-C: 25.75±2.18; MCI-NC: 27.09±1.82.
Spatial Patterns of Atrophy in AD

GM

WM

Ventricles
SPARE-AD Score distribution

MCI Converters (MCI-C)    MCI Non-Converters (MCI-NC)
Trajectories of SPARE-AD Scores
Average baseline MMSE and rate of change of MMSE for MCI-C and three sub-groups of MCI-NC.

<table>
<thead>
<tr>
<th></th>
<th>Average baseline MMSE scores</th>
<th>Average rate of change of MMSE per year</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCI-C</td>
<td>26.45</td>
<td>-2.06</td>
</tr>
<tr>
<td>MCI-NC part 1</td>
<td>26.92</td>
<td>-0.9</td>
</tr>
<tr>
<td>MCI-NC part 2</td>
<td>26.71</td>
<td>-0.3</td>
</tr>
<tr>
<td>MCI-NC part 3</td>
<td>28.67</td>
<td>-0.25</td>
</tr>
</tbody>
</table>
GM differences between MCI-C and MCI-NC (Baseline)

WM differences between MCI-C and MCI-NC (baseline)
Feature Extraction and Selection (Dimensionality reduction)?
Matrix Factorization:

The basis pixels that will be combined into the reconstructed pixel

Weights used to combine basis images into reconstructed pixel

Batmanghelich, Taskar, and Davatzikos, IPMI 2009
How to formulate it

- How far is our estimation from data:
  \[ D_\varphi(X; BC) \]
  \[ D_\varphi(X; Y) := \sum_{i,j} D_\varphi(x_{ij}, y_{ij}) \]
  \[ D_\varphi(x; y) := \varphi(x) - \varphi(y) - \varphi'(y)(x - y) \]

- Pick different \( \varphi \) and get different distance function:
  \[ \varphi(x) = x \log x \quad \Rightarrow \quad \text{KL-divergence} \]
  \[ \varphi(x) = x^2 \quad \Rightarrow \quad \text{Frobenius distance} \]

- But the problem is ill-posed, we need regularization:
  \[ D_\varphi(X; BC') + \alpha(B) + \beta(C) \]
  - sparseness
  - MRF
  - Make NMF supervised!
Some preliminary result

- Two of most discriminative basis after ranking:

\[ b_1 \]

\[ b_2 \]

With only 30 basis from classification pairs (AD vs All, MCI vs. All, NC vs. All), totally 90, and using very small portion of samples, we got 88% for AD vs. NC and 57% for AD vs MCI and 76% for MCI vs NC
Classification
SVM

Classification hyperplane

- "normal"
- "abnormal"

\[ f(x) = w^T \varphi(x) + b \quad \text{- separating function} \]
\[ y_i \in \{-1, +1\} \quad \text{- class labels} \]

\[
\min_{w, b, \xi} \frac{1}{2} w^T w + \beta \sum_{i=1}^{n} \xi_i \\
\text{s.t. } y_i \left( w^T \varphi(x_i) + b \right) \geq 1 - \xi_i, \forall i = 1, \ldots, n
\]
Semi-supervised Classification and Clustering of Medical Images

Roman Filipovych and Christos Davatzikos
Figure 1: Typical scenarios of disease evolution. Scenario 1: While patients are grouped under a common umbrella, in reality they form distinct clinical categories. Scenario 2: Disease evolves gradually, and there are no distinct clinical categories in the data. However, the level of disease progression is different for different individuals.
More Realistic Scenario

Semi-supervised classification and clustering of medical images
Semi-supervised SVM (Transductive SVM)

\[
\min_{y_{l+1}, \ldots, y_n} \min_{w, b, \xi} \frac{1}{2} w^T w + \beta_l \sum_{i=1}^{l} \xi_i + \beta_u \sum_{j=l+1}^{n} \xi_j \\
\text{s.t. } y_i \left( w^T \varphi(x_i) + b \right) \geq 1 - \xi_i, \forall i = 1, \ldots, l \\
y_j \left( w^T \varphi(x_j) + b \right) \geq 1 - \xi_j, \forall j = l + 1, \ldots, n
\]

\( y_i \in \{-1, +1\} \) - known labels, \( i = 1, \ldots, l \)

\( y_j \in \{-1, +1\} \) - unknown labels, \( j = l + 1, \ldots, n \)
Maximum-margin Clustering (MMC)

Classification (clustering) hyperplane

\[
\min_{z_1, \ldots, z_n} \min_{\hat{w}, \hat{b}, \xi} \frac{1}{2} \hat{w}^T \hat{w} + \beta \sum_{i=1}^{n} \hat{\xi}_k \\
\text{s.t. } z_k \left( \hat{w}^T \varphi(x_k) + \hat{b} \right) \geq 1 - \hat{\xi}_k, \forall k = 1, \ldots, n \\
\hat{\xi}_k \geq 0, \forall k = 1, \ldots, n
\]

\((z_1, \ldots, z_n) \in \{-1, +1\}^n - \text{unknown labels}\)
Joint Maximum-Margin Classification and Clustering

\[
\begin{align*}
&\min_{y_{l+1}, \ldots, y_n, z_1, \ldots, z_n} \min_{w, b, \xi} \min_{\hat{w}, \hat{b}, \hat{\xi}} \frac{1}{2} w^T w + \frac{1}{2} \hat{w}^T \hat{w} \\
&\quad + \beta_l \sum_{i=1}^{l} \xi_i + \beta_u \sum_{j=l+1}^{n} \xi_j + \hat{\beta} \sum_{k=1}^{n} \hat{\xi}_k \\
&\text{s.t.} \quad y_i \left(w^T \phi(x_i) + b\right) \geq 1 - \xi_i \\
&\quad y_j \left(w^T \phi(x_j) + b\right) \geq 1 - \xi_j \\
&\quad z_k \left(\hat{w}^T \phi(x_k) + \hat{b}\right) \geq 1 - \hat{\xi}_k, \quad \text{if} \quad \left(w^T \phi(x_k) + b\right) \geq 0 \\
&\quad \xi_i \geq 0, \xi_j \geq 0, \hat{\xi}_k \geq 0
\end{align*}
\]
JointMMCC: Relaxed

\[
\min_{y_{l+1}, \ldots, y_n} \min_{z_1, \ldots, z_n} \min_{w, b, \xi, \hat{w}, \hat{b}, \hat{\xi}} \frac{1}{2} w^T w + \frac{1}{2} \hat{w}^T \hat{w} \\
+ \beta_i \sum_{i=1}^{l} \xi_i + \beta_u \sum_{j=l+1}^{n} \xi_j + \hat{\beta} \sum_{k=1}^{n} \hat{\xi}_k
\]

s.t. \( y_i \left( w^T \varphi(x_i) + b \right) \geq 1 - \xi_i \)
\( y_j \left( w^T \varphi(x_j) + b \right) \geq 1 - \xi_j \)
\( z_k \left( \hat{w}^T \varphi(x_k) + \hat{b} \right) + \gamma R(x_k) \geq 1 - \hat{\xi}_k \)

where \( \gamma \gg 1 \) and \( R(x) \) is the ramp function

\[
R(x) = \begin{cases} 
- (w^T \varphi(x) + b), & \text{if } w^T \varphi(x) + b < 0 \\
0, & \text{otherwise}
\end{cases}
\]

and

\( y_j = \text{sign}(w^T \varphi(x_j) + b) \)
\( z_k = \text{sign}(\hat{w}^T \varphi(x_k) + \hat{b}) \), \text{ if } w^T \varphi(x_k) + b \geq 0
Experiments

Labeled data: *15 subjects with best CVLT scores, and 24 subjects with worst scores*
The “obtained” cognitively declined class contains “obtained” Cluster 1 and Cluster 2

Table 1. Differences in clinical scores between the obtained classes and clusters. p-values of two-sided t-test.

<table>
<thead>
<tr>
<th></th>
<th>Normal vs. Cognitively declined</th>
<th>Cluster 1 vs Cluster 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>CVLT List A Sum</td>
<td>p=0.0000</td>
<td>p=0.0000</td>
</tr>
<tr>
<td>CVLT Long Delay Free</td>
<td>p=0.0000</td>
<td>p=0.0000</td>
</tr>
<tr>
<td>BVRT Errors</td>
<td>p=0.0034</td>
<td>p=0.3105</td>
</tr>
<tr>
<td>MMSE</td>
<td>p=0.0001</td>
<td>p=0.0125</td>
</tr>
</tbody>
</table>
Solving JointMMCC for a given constraint set

Figure 3. Voxel-based analysis of the obtained classes and clusters
High-Dimensional Pattern Regression: from Images to Clinical Variables

Ying Wang
Christos Davatzikos
## Regression Method: RVR

<table>
<thead>
<tr>
<th></th>
<th>Relevance Vector Machine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduced</td>
<td>2001 Tipping</td>
</tr>
<tr>
<td>Theory</td>
<td>Bayesian formulation</td>
</tr>
<tr>
<td>Basis Function</td>
<td>Arbitrary kernel</td>
</tr>
<tr>
<td>Regularization</td>
<td>L1 Norm</td>
</tr>
<tr>
<td>Sparse Representation</td>
<td>Fewer relevance vectors needed</td>
</tr>
<tr>
<td>Regression</td>
<td>Probabilistic prediction</td>
</tr>
</tbody>
</table>

- **Regression Method**: RVR
- **Introduced**: 2001 Tipping
- **Theory**: Bayesian formulation
- **Basis Function**: Arbitrary kernel
- **Regularization**: L1 Norm
- **Sparse Representation**: Fewer relevance vectors needed
- **Regression**: Probabilistic prediction
Regression Method: RVR

- Input vectors \( \{x_n\}_{n=1}^N \) and targets \( \{t_n\}_{n=1}^N \)

\[
t_n = y(x_n) + \zeta_n
\]

- Model:

\[
y(x, w) = \sum_{m=0}^{M} \omega_m \phi_m(x) = w^T \phi
\]

- Regression:

\[
P(t|x) = \mathcal{N}(t|y(x), \sigma^2)
\]
Regression Method: RVR

- Gaussian prior for $\omega_m$ with hyper-parameters $\alpha_m$:

$$p(w|\alpha) = \prod_{m=0}^{N} \mathcal{N}(\omega_m | O, \alpha_m^{-1}) \alpha_0$$

- Hyper-priors over $\sigma^2$:

$$\tau = \sigma^{-2}$$
Real Data: ADNI

- **MMSE:** Mini Mental State Examination
- **BNT:** Boston Naming Testing

<table>
<thead>
<tr>
<th>Score</th>
<th>Information</th>
<th>AD</th>
<th>MCI</th>
<th>CN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No. of subjects</td>
<td>23</td>
<td>74</td>
<td>22</td>
</tr>
<tr>
<td>MMSE</td>
<td>Age, mean ± std</td>
<td>78.45 ± 6.02</td>
<td>75.87 ± 7.28</td>
<td>73.25 ± 5.45</td>
</tr>
<tr>
<td></td>
<td>Mean ± std</td>
<td>15.06 ± 2.43</td>
<td>25.75 ± 2.44</td>
<td>28.54 ± 0.98</td>
</tr>
<tr>
<td>BNT</td>
<td>No. of subjects</td>
<td>22</td>
<td>31</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>Age, mean ± std</td>
<td>77.60 ± 5.95</td>
<td>74.75 ± 8.63</td>
<td>71.36 ± 7.04</td>
</tr>
<tr>
<td></td>
<td>Mean ± std</td>
<td>16.55 ± 1.17</td>
<td>23 ± 2.60</td>
<td>28.88 ± 0.78</td>
</tr>
</tbody>
</table>
Experiments on ADNI Data - MMSE

RVR Results For Regional Features

MSE is 10.8338; Correlation is 0.75775
Experiments on ADNI Data - BNT

RVR Results For Regional Features

MSE is 22.7281; Correlation is 0.58908
Experiments on ADNI Data – Rate of Change

<table>
<thead>
<tr>
<th>Measured Change of MMSE</th>
<th>Predicted Change of MMSE</th>
</tr>
</thead>
</table>

RVR Results For Regional Features

MSE is 1.4223; Correlation is 0.53746
Regions Most Representative of the Group Difference

A: MMSE regression by using GM only; B: MMSE regression by using GM, WM and CSF; C: BNT regression by using GM; D: MMSE Change regression by using GM.
Conclusions

- Multi-parametric imaging patterns carry more information than individual measurements viewed in isolation.
- Fully-supervised (trained) and semi-supervised classification can produce valuable diagnostic and prognostic biomarkers.
- Pattern regression methods can estimate continuous variables.
Thank you