Mining MR Image Data by Discriminative Methods for the Diagnosis of Dementia

Ceyhun Burak Akgül
Former Marie Curie Postdoctoral Fellow @ Philips Research

www.cba-research.com
cb.akgul@gmail.com
Motivation – 1/2

Diagnose dementia (e.g., Alzheimer’s disease) from MR Images

Standard medical practice:
- patient history, collateral history from relatives
- clinical observations: neurological/neuropsychological features

**BUT:** does not often lead to an early diagnosis

An emerging trend: Exploit imaging data

**HOW?**
Brain Atrophy?

- requires longitudinal data: MR scans at different time stamps
- requires complex mathematical modeling and algorithms
- should quantify minute changes (that human eye can’t see)

Or something else...
Data Mining Framework

**Representation**
learn an image representation from data: analyze images
- at each location
- at several scales
- with several patterns

**Selection**
discover image features using labeled data

**Classification**
characterize patient groups discriminatively

**Information Fusion**
combine multiple (visual or non-visual) information sources
Data Mining Framework: Overview

Dataset (view, slice, labels)

Training Set (63 subjects)

PCA-based Image Description

Feature Ranking and Selection

Descriptors

SVM Model Selection

Model parameters

SVM Classifier Training

Descriptors

Feature Computation

Intensity templates

Feature indices

Classification

SVM decision values

Classifier model

Non-visual patient Information

Probabilistic Information Fusion

Image-based decisions

Combined decisions
(1) Image Representation

**Training set**

**Training set**

**IMAGE SET**

**Support size** $w_s$

**Collect Patches**

**Basis vectors** $B_s = \{b_{s,k}\}$

**Image patches** $(w_s \times w_s)$

**PCA**

**XCORR**

**Feature Maps at scale $w$**

$FM_{s,d'}$

$FM_{s,k}$

$FM_{s,1}$

Learning image patterns

- Basis vectors are common intensity patterns
- Use these patterns as templates
- Each window size induces a basis at a different scale
- Repeat the analysis by varying the window size
(2) Feature Selection by Ranking – 1/3

- Each image is described by $S \times K_s$ feature maps
- At each pixel location, there are $S \times K_s$ feature values
- Each feature $\leftrightarrow$ a distinct (scale, template)-pair
- At each location:
  - rank the features based on their “usefulness”
  - pick the most “useful” feature for description

**“Usefulness” $\leftrightarrow$ Mutual information between feature and diagnostic label**

$$MI(x,y) = \sum_{y \in \{-1, +1\}} \int p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

$x \in [0,1]:$ normalized feature value at location $(i,j)$

$y \in \{-1, +1\}:$ diagnostic label of the image
(2) Feature Selection by Ranking – 2/3

Maximum Mutual Information Maps\(^(*)\) at Different Scales

Scale 1
\(w = 5\)

Scale 2
\(w = 11\)

Scale 3
\(w = 21\)

Scale 4
\(w = 31\)

Scale 5
\(w = 41\)

Maximum Mutual Information Map\(^(*)\) Combined over Scales

\(^(*)\) Map size: 87 ×70
(2) Feature Selection by Ranking – 3/3

- Each $\lambda$-value induces a subset of “surviving” locations
- Remember
  - Each location is associated with a (scale, basis)-pair whose feature value gives the maximum MI.

$\rightarrow$ By varying $\lambda$, one obtains **nested subsets of features**

- **Or alternatively, retain top $k$ features**
Amount of Data Processed: Some Facts

• 100 slices per subject  ~ **400 Megabytes/subject**
• 121 subjects  ~ **50 Gigabytes** TOTAL AMOUNT OF DATA PROCESSED

• 100 informative features/(subject×slice) selected as the descriptor  
  < **1 kilobyte/(subject×slice)**
(3) Classification: SVM Basics

A non-linear SVM classifier $F$ is indexed by two parameters $(C,\gamma)$:

- The parameter $C$ trades off training error vs. classifier complexity
- The kernel parameter $\gamma$ determines the class of functions $F$ and affects class separation
  (in some sense, it also determines the classifier complexity)

One has to specify the “best” $(C,\gamma)$-pair before testing the classifier.

**A good empirical option**

$$(C,\gamma)^* = \arg\min C V \text{Err}(F(C,\gamma))$$

$Err_{CV}$: Cross validation error
(3) Classification: Model Selection

- Leave-One-Out (LOO) cross-validation
- Initial search for the \((C,\gamma)\)-parameters on a coarse grid

\[ \{(C,\gamma) \text{ such that } \text{ACC}>\text{THRESH}\} \]

- Search on a finer grid
- Further heuristics – Look at:
  - Sensitivity
  - Fraction of SVs (model parsimony)
  - Specificity
(4) Probabilistic Information Fusion

**Bayesian Theory:** The decision on the class label should be made on the conditional probability of the class label given all other relevant information.

\[
P(\text{label} \mid \text{info}) = P(\text{label} \mid \text{visual, non-visual})
\]

\[
P(\text{label} \mid \text{info}) \propto P(\text{label}) \times P(\text{visual, non-visual} \mid \text{label})
\]

\[
= P(\text{label}) \times P(\text{visual} \mid \text{label}) \times P(\text{non-visual} \mid \text{label})
\]

\[
\propto P(\text{label} \mid \text{visual}) \times P(\text{non-visual} \mid \text{label})
\]

derived from SVM outputs  class-conditional distributions

estimated from training data
Experiments: Dataset

OASIS Dataset
121 Subjects

Training set
63 Subjects
16 AD (CDR=1)
47 Control (CDR=0)
• PCA-based feature learning
• Feature ranking and filtering
• Discriminative learning (SVM)
  - SVM model selection

Test set
58 Subjects
14 AD (CDR=1)
44 Control (CDR=0)
• Not seen during none of the training stages
• Reserved only for performance evaluation

• CDR: Clinical Dementia Rating: normal \(\rightarrow\) CDR = 0  moderate dementia \(\rightarrow\) CDR =1
• Stratified split keeps the class proportions the same in both sets (Control/AD \(\approx\) 3)
Experiments: MR Data

- 26 Axial + 46 Sagittal + 28 Coronal = 100 MR slices processed separately
- Each slice described by 100 informative image features
Experiments: Discriminative Slices – 1/2

Axial 6: Acc = 70.7%, Sens = 64.3%, Spec = 72.7%

Axial 10: Acc = 79.9%, Sens = 71.4%, Spec = 81.8%

Axial 12: Acc = 84.5%, Sens = 78.6%, Spec = 86.4%

Axial 15: Acc = 72.4%, Sens = 64.3%, Spec = 75.0%

Axial 26: Acc = 81.0%, Sens = 64.3%, Spec = 86.4%

Sagittal 26: Acc = 67.2%, Sens = 57.1%, Spec = 70.5%

Sagittal 32: Acc = 79.3%, Sens = 71.4%, Spec = 81.8%

Sagittal 33: Acc = 77.6%, Sens = 64.3%, Spec = 81.8%

Sagittal 35: Acc = 84.5%, Sens = 64.3%, Spec = 90.9%

Sagittal 37: Acc = 75.9%, Sens = 57.1%, Spec = 81.8%

Coronal 15: Acc = 65.5%, Sens = 57.1%, Spec = 68.2%

Coronal 25: Acc = 72.4%, Sens = 57.1%, Spec = 77.3%

Coronal 26: Acc = 81.0%, Sens = 57.1%, Spec = 88.6%
Experiments: Discriminative Slices – 2/2

Axial 12
Acc = 84.5%
Sens = 78.6%
Spec = 86.4%

Coronal 26
Acc = 81.0%
Sens = 57.1%
Spec = 88.6%

Sagittal 32
Acc = 79.3%
Sens = 71.4%
Spec = 81.8%

Axial 12 > Sagittal 32 > Coronal 26
Experiments: ROC vs. Descriptor Size

### Area Under the Curve (AUC) vs Descriptor Size

<table>
<thead>
<tr>
<th>Descriptor Size</th>
<th>Accuracy (%)</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Axial 12</td>
<td>100</td>
<td>84.5 (49/58)</td>
<td>84.1 (37/44)</td>
</tr>
<tr>
<td>Sagittal 32</td>
<td>160</td>
<td>65.5 (38/58)</td>
<td>65.9 (29/44)</td>
</tr>
<tr>
<td>Coronal 25</td>
<td>160</td>
<td>77.6 (45/58)</td>
<td>77.3 (34/44)</td>
</tr>
</tbody>
</table>

**ROC**: Receiver Operating Characteristic: TPR vs. FPR

**AUC**: Area under the ROC curve

**EER**: Equal error rate (sensitivity = specificity)
Experiments: Information Fusion – 1/2

**SVM-only**: Image-based decisions gleaned from SVM outputs

**MMSE-only**: MMSE-based decisions: *if MMSE<Thresh, then decide ill*

**SVM+MMSE-OASIS**: statistics estimated from OASIS training set (63 subjects)

**SVM+MMSE-ADNI**: statistics estimated from ADNI dataset (322 subjects)
Experiments: Information Fusion – 2/2

**ROC Summary**

<table>
<thead>
<tr>
<th></th>
<th>AUC</th>
<th>EER (%)</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM only</td>
<td>0.8260</td>
<td>15.3</td>
<td>84.5</td>
</tr>
<tr>
<td>MMSE only</td>
<td>0.9798</td>
<td>13.3</td>
<td>86.7</td>
</tr>
<tr>
<td>SVM+MMSE-OASIS</td>
<td>0.9798</td>
<td>8.7</td>
<td>91.3</td>
</tr>
<tr>
<td>SVM+MMSE-ADNI</td>
<td>0.9871</td>
<td>8.4</td>
<td>91.6</td>
</tr>
</tbody>
</table>

**SVM+MMSE-ADNI > SVM+MMSE-OASIS > MMSE-only > SVM-only**

- Information fusion is very useful indeed
- **Reliable statistics!!!** ADNI (322 subjects) > OASIS (63 subjects)
  - 229 controls
  - 93 positives
  - 47 controls
  - 16 positives
Summary

• Data-driven image representation
  – Unsupervised learning of local image patterns via PCA
  – Localized, at several scales, with several patterns

• Feature ranking and filtering
  – Supervised: based on MI between scalar features and class labels

• Discriminative learning
  – SVM model selection via cross-validation and further heuristics

• Information fusion
  – Leverage image-only decisions by non-visual information
  – Generic: works with any kind of meta-data as long as statistics available

Proof of concept:
A promising data-driven framework for the diagnosis of dementia with high predictive performance
What’s Next?

**Practical**
Go validate these results clinically
Do these slices, locations, scales, patterns make sense?
Acquire larger sets of labeled data
Allocate higher computational resources

**Methodological**
Other sparser image representations: ICA-based? NNMF-based?
Multivariate feature selection
Model selection: Don’t use one, average multiple models
Other classification schemes: AdaBoost

**Theoretical ...**
What’s Next? – *Theoretical*

Data → **Representations**
- **Features**
- **Classifiers**

? → Diagnosis

convolutional networks?
To conclude...

There’s nothing more practical than a good theory.

Lewin, 1952