Broad Learning for Healthcare

Bokai Cao
03/08/2018

Defense Committee:
Philip S. Yu (Department of Computer Science), Chair and Advisor
Bing Liu (Department of Computer Science)
Piotr Gmytrasiewicz (Department of Computer Science)
Alex D. Leow (Department of Psychiatry)
Olusola Ajilore (Department of Psychiatry)
BIG DATA

HEALTHCARE DATA

Volume
AMOUNT OF DATA

Variety
TYPES OF DATA

Velocity
FREQUENCY OF DATA

Veracity
QUALITY OF DATA

Volume

Variety

Velocity

Veracity
Broad Learning

• Mining heterogeneous data sources and information fusion.

• Multi-view learning (the same entity).

• Transfer learning (similar entities).

• Learning in heterogeneous information networks (linked entities).
Model capacity

Data capacity

Data-driven model
Healthcare Problems

• (P1) How can we determine important biomarkers from measures collected from multiple medical examinations?
Healthcare Problems

- (P2) How can we identify connectivity patterns in the brain that are associated with brain injury using side information?
Healthcare Problems

• (P3) How can we obtain effective brain network representations with the guidance of side information and other constraints?
Healthcare Problems

• (P4) How can we monitor the mental health using time series data collected from multiple sensors on a mobile and wearable device?

Accelerometer

Gyroscope

GPS

Microphone

Mood detection

Keyboard
A Mind Map

Multi-view measures

Brain networks

Keyboard dynamics

(P1) Multi-view feature selection

(P2) Subgraph pattern mining

(P3) Brain network embedding

(P4) Multi-view sequence prediction

EEG

fMRI, DTI

HIV infection on brain

Anxiety disorder

Bipolar disorder

Computer-aided diagnosis

Mobile health
(P1) Multi-view Feature Selection
Tensor-based Multi-view Feature Selection with Applications to Brain Diseases,
B. Cao et al., ICDM 2014.
Multi-view Data in Healthcare

Limited subjects available yet introducing a large number of measures

Multi-view learning + feature selection
Modeling Multi-View Data

• Vector-based method

Ignoring cross-view difference
Modeling Multi-View Data

- Tensor-based method

Multi-view data

View 1

View 2

View 3

Modeling in the tensor space

Feature selection in the tensor space

Redundancy and irrelevance
Modeling Multi-View Data

• Input space
  • Preserve the interpretability of original data.
  • Fail to explore feature interactions across different views.

• Tensor space
  • Explore feature interactions across different views.
  • Introduce redundant and irrelevant features.

It motivates us to build a dual mapping between the *input space* and *tensor space*.
Modeling Multi-View Data

• Dual method

Simplicity of *input space* + Abundance of *tensor space*

Multi-view data

Modeling in the tensor space

Feature selection in the input space
Wrapper-based Feature Selection

![Diagram showing the process of wrapper-based feature selection.]

1. **Training set** flows to the **Classification algorithm**.
2. The **Feature set** is then selected.
3. The selected **Feature set** and **Hypothesis** are used to train the **Classification algorithm**.
4. The **Classification evaluation** provides the estimated performance.
5. The process is repeated with the **Test set** and the selected **Feature set**.
Recursive Feature Elimination

- Feature ranking with **Support Vector Machines**

  **Train a SVM classifier**
  
  Weight vector \( w = \sum_i \alpha_i y_i x_i \)
  
  Ranking criteria \( r_i = (w_i)^2 \)
  
  Eliminate feature \( \text{arg min}(r) \)


SVM-RFE
Recursive Feature Elimination

- Feature ranking with **Support Tensor Machines**

Train a STM classifier

\[ \mathcal{X} = \prod_{v=1}^{m} \circ x^{(v)} \]

Weight tensor

\[ \mathcal{W} = \sum \alpha_i y_i \mathcal{X}_i \]

Ranking criteria

\[ r_{i_v}^{(v)} = \sum_{i_1}^{i} \cdots \sum_{i_{v-1}}^{i_{v-1}} \sum_{i_{v+1}}^{i_{v+1}} \cdots \sum_{i_m}^{i_m} (w_{i_1, \ldots, i_m})^2 \]

Eliminate feature

\[ \text{arg min}(r^{(v)}) \]
Tensor-based Multi-view Recursive Feature Elimination

Tensor product

Support Tensor Machines

Feature selection

Feature evaluation
Linear Kernel

Optimization objective

$$\min_{w^{(v)}, b, \xi} \frac{1}{2} \prod_{v=1}^{m} \left\| w^{(v)} \right\|_F^2 + C \sum_{i=1}^{n} \xi_i$$

s.t. $$y_i \left( \prod_{v=1}^{m} \langle w^{(v)}, x^{(v)}_i \rangle + b \right) \geq 1 - \xi_i$$

$$\xi_i \geq 0, \forall i = 1, \ldots, n.$$  

$$f(\mathcal{X}) = \text{sign} \left( \prod_{v=1}^{m} \langle w^{(v)}, x^{(v)} \rangle + b \right)$$

$$\mathcal{W} = \prod_{v=1}^{m} w^{(v)}$$

Feature selection

$$\text{argmin}_{i_v} \left( r^{(v)}_{i_v} \right)$$

$$r^{(v)}_{i_v} = \left( w^{(v)}_{i_v} \right)^2$$

$$= \sum_{i_1} \cdots \sum_{i_{v-1}} \sum_{i_{v+1}} \cdots \sum_{i_m} \left( w^{(1)}_{i_1} \cdots w^{(m)}_{i_m} \right)^2$$

$$= \left( w^{(v)}_{i_v} \right)^2 \prod_{1 \leq j \leq m} \left\| w^{(j)} \right\|_F^2$$

$$= P(-v) \left( w^{(v)}_{i_v} \right)^2$$
Extension to Nonlinear Kernels

**Optimization objective**

$$\min_{\alpha} \frac{1}{2} \alpha^{(v)T} H \alpha^{(v)} - \alpha^{(v)T} 1$$

s.t. \( \sum_{i=1}^{n} \alpha_i^{(v)} y_i = 0 \)

\( 0 \leq \alpha_i^{(v)} \leq C, \forall i = 1, \ldots, n. \)

where \( H_{p,q} = y_p y_q \kappa(x_p^{(v)'}, x_q^{(v)'}) \)

**Feature selection**

$$\arg\min_{i_v} \left( r_{i_v}^{(v)} \right)$$

\( r_{i_v}^{(v)} = \frac{1}{2} (\alpha^{(v)T} H \alpha^{(v)} - \alpha^{(v)T} H \alpha^{(v)}) \)

where

\( H_{p,q}(-i_v) = y_p y_q \kappa(x_p^{(v)'}, x_q^{(v)'}, x_{-i_v}^{(v)}') \)

**Equations**

\( P^{(-v)} = \prod_{1 \leq j \leq m} \| w^{(j)} \|_F^2 \)

\( Q_i^{(-v)} = \prod_{1 \leq j \leq m} \langle w^{(j)}, x_i^{(j)} \rangle \)

\( x_i^{(v)'} = (Q_i^{(-v)}/\sqrt{P^{(-v)}}) x_i^{(v)} \)
Algorithm tMVFS

Input: \( \{X^{(v)}\}_{v=1}^{m} \) (multi-view training samples), \( y \) (class labels), \( \{J_v\}_{v=1}^{m} \) (number of features to be selected in each view)

Output: \( \{s^{(v)}\}_{v=1}^{m} \) (selected multi-view features)

1. for \( v = 1 \) to \( m \) do
2.   Initialize the subset of surviving features: \( s^{(v)} = [1, \ldots, I_v] \)
3.   repeat
4.     Restrict training samples to good feature indices: \( X^{(v)*} = X^{(v)}(s^{(v)},:) \)
5.     Train the classifier: \( \alpha = \text{SVM\text{-}train}(X^{(v)*}, y) \) as in Equation 2.16
6.     Compute the weight vector \( w^{(v)} \) according to Equation 2.17
7.     Compute the ranking criteria \( r^{(v)} \) according to Equation 2.13
8.     Find the bad feature index: \( f = \arg \min(r^{(v)}) \)
9.     Eliminate the feature: \( s^{(v)}(f) = [] \)
10. until length(\( s^{(v)} \)) \( \leq J_v \)
11. end for
Data Collection

• Chicago Early HIV Infection Study

56 HIV and 21 seronegative controls.
Baselines

- **tMVFS**: the proposed dual method of tensor-based multi-view feature selection.
- **MIQP**: iterative tensor product feature selection with mixed-integer quadratic programming.
- **STM-RFE**: recursive feature elimination with the tensor product features.
- **SVM-RFE**: recursive feature elimination with the concatenated multi-view features.
- **MKL**: multi-kernel SVM.
- **STM**: direct modeling with the tensor product features.
- **SVM**: direct modeling with the concatenated multi-view features.

https://github.com/caobokai/tMVFS
### Two Views

Classification performance with two views in the linear case.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Methods</th>
<th>Evaluation metrics</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Accuracy</td>
<td>Precision</td>
<td>Recall</td>
<td>$F_1$</td>
</tr>
<tr>
<td>D2.1</td>
<td>tMVFS</td>
<td>0.762 (1)</td>
<td>0.784 (1)</td>
<td>0.792 (4)</td>
<td>0.786 (1)</td>
</tr>
<tr>
<td></td>
<td>MIQP</td>
<td>0.667 (4)</td>
<td>0.667 (5)</td>
<td>0.833 (1)</td>
<td>0.737 (3)</td>
</tr>
<tr>
<td></td>
<td>STM-RFE</td>
<td>0.643 (5)</td>
<td>0.657 (6)</td>
<td>0.792 (4)</td>
<td>0.713 (6)</td>
</tr>
<tr>
<td></td>
<td>SVM-RFE</td>
<td>0.690 (3)</td>
<td>0.730 (3)</td>
<td>0.750 (7)</td>
<td>0.736 (4)</td>
</tr>
<tr>
<td></td>
<td>MKL</td>
<td>0.643 (5)</td>
<td>0.710 (4)</td>
<td>0.792 (4)</td>
<td>0.730 (5)</td>
</tr>
<tr>
<td></td>
<td>STM</td>
<td>0.595 (7)</td>
<td>0.623 (7)</td>
<td>0.833 (1)</td>
<td>0.698 (7)</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.738 (2)</td>
<td>0.747 (2)</td>
<td>0.833 (1)</td>
<td>0.782 (2)</td>
</tr>
<tr>
<td>D2.2</td>
<td>tMVFS</td>
<td>0.714 (1)</td>
<td>0.769 (1)</td>
<td>0.792 (1)</td>
<td>0.759 (1)</td>
</tr>
<tr>
<td></td>
<td>MIQP</td>
<td>0.595 (5)</td>
<td>0.638 (6)</td>
<td>0.667 (4)</td>
<td>0.648 (5)</td>
</tr>
<tr>
<td></td>
<td>STM-RFE</td>
<td>0.619 (3)</td>
<td>0.656 (4)</td>
<td>0.750 (2)</td>
<td>0.690 (3)</td>
</tr>
<tr>
<td></td>
<td>SVM-RFE</td>
<td>0.524 (7)</td>
<td>0.607 (7)</td>
<td>0.500 (6)</td>
<td>0.540 (7)</td>
</tr>
<tr>
<td></td>
<td>MKL</td>
<td>0.667 (2)</td>
<td>0.730 (2)</td>
<td>0.667 (4)</td>
<td>0.692 (2)</td>
</tr>
<tr>
<td></td>
<td>STM</td>
<td>0.619 (3)</td>
<td>0.639 (5)</td>
<td>0.750 (2)</td>
<td>0.683 (4)</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.548 (6)</td>
<td>0.657 (3)</td>
<td>0.500 (6)</td>
<td>0.553 (6)</td>
</tr>
</tbody>
</table>
More than Two Views

Classification performance with many views in the linear case.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Methods</th>
<th>Evaluation metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Accuracy</td>
</tr>
<tr>
<td>D3.1</td>
<td>tNVFS</td>
<td>0.833 (1)</td>
</tr>
<tr>
<td></td>
<td>SVM-RFE</td>
<td>0.714 (4)</td>
</tr>
<tr>
<td></td>
<td>MEL</td>
<td>0.738 (3)</td>
</tr>
<tr>
<td></td>
<td>STM</td>
<td>0.690 (5)</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.762 (2)</td>
</tr>
<tr>
<td>D3.2</td>
<td>tNVFS</td>
<td>0.810 (1)</td>
</tr>
<tr>
<td></td>
<td>SVM-RFE</td>
<td>0.667 (5)</td>
</tr>
<tr>
<td></td>
<td>MEL</td>
<td>0.714 (2)</td>
</tr>
<tr>
<td></td>
<td>STM</td>
<td>0.714 (2)</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.690 (4)</td>
</tr>
<tr>
<td>D4.1</td>
<td>tNVFS</td>
<td>0.929 (1)</td>
</tr>
<tr>
<td></td>
<td>SVM-RFE</td>
<td>0.881 (3)</td>
</tr>
<tr>
<td></td>
<td>MEL</td>
<td>0.905 (2)</td>
</tr>
<tr>
<td></td>
<td>STM</td>
<td>0.810 (5)</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.857 (4)</td>
</tr>
<tr>
<td>D4.2</td>
<td>tNVFS</td>
<td>0.920 (1)</td>
</tr>
<tr>
<td></td>
<td>SVM-RFE</td>
<td>0.833 (4)</td>
</tr>
<tr>
<td></td>
<td>MEL</td>
<td>0.905 (2)</td>
</tr>
<tr>
<td></td>
<td>STM</td>
<td>0.810 (5)</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.857 (3)</td>
</tr>
<tr>
<td>D5.1</td>
<td>tNVFS</td>
<td>0.952 (1)</td>
</tr>
<tr>
<td></td>
<td>SVM-RFE</td>
<td>0.905 (2)</td>
</tr>
<tr>
<td></td>
<td>MEL</td>
<td>0.905 (2)</td>
</tr>
<tr>
<td></td>
<td>STM</td>
<td>0.810 (5)</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.905 (2)</td>
</tr>
<tr>
<td>D5.2</td>
<td>tNVFS</td>
<td>0.905 (1)</td>
</tr>
<tr>
<td></td>
<td>SVM-RFE</td>
<td>0.857 (4)</td>
</tr>
<tr>
<td></td>
<td>MEL</td>
<td>0.905 (1)</td>
</tr>
<tr>
<td></td>
<td>STM</td>
<td>0.714 (5)</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.881 (3)</td>
</tr>
<tr>
<td>D6.1</td>
<td>tNVFS</td>
<td>0.952 (1)</td>
</tr>
<tr>
<td></td>
<td>SVM-RFE</td>
<td>0.905 (2)</td>
</tr>
<tr>
<td></td>
<td>MEL</td>
<td>0.905 (2)</td>
</tr>
<tr>
<td></td>
<td>STM</td>
<td>0.810 (5)</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.881 (4)</td>
</tr>
<tr>
<td>D6.2</td>
<td>tNVFS</td>
<td>0.929 (1)</td>
</tr>
<tr>
<td></td>
<td>SVM-RFE</td>
<td>0.905 (2)</td>
</tr>
<tr>
<td></td>
<td>MEL</td>
<td>0.905 (2)</td>
</tr>
<tr>
<td></td>
<td>STM</td>
<td>0.810 (5)</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.905 (2)</td>
</tr>
</tbody>
</table>
RBF Kernel

Accuracy | Precision | Recall | F1
---|---|---|---
0 | 0 | 0 | 0
0 | 0 | 0 | 0
0 | 0 | 0 | 0
0 | 0 | 0 | 0
0 | 0 | 0 | 0
0 | 0 | 0 | 0
0 | 0 | 0 | 0
0 | 0 | 0 | 0

SVM | SVM-RFE | tMVFS
Selected Features

• Neuropsychological tests:
  Karnofsky Performance Scale, NART FSIQ, Rey Trial
• Flow cytometry:
  T cells 4+8-, 3+56-16+NKT Cells 4+8-, Lymphocytes
• Plasma luminex:
  MMP-2, GRO, TGFα
• Freesurfer:
  Cerebral Cortex, Thalamus Proper, CC_Mid_Posterior
• Overall brain microstructure:
  CC, Hippocampus, Cerebral-White-Matter
• Localized brain microstructure:
  CC_Mid_Anterior, CC_Anterior, CC_Central
• Brain volumetry:
  Peripheral Gray Matter, Brain Parenchyma Volume
Conclusions

• We propose a dual method that (1) effectively explores multi-view feature interactions in the tensor space and (2) efficiently conducts feature selection in the input space.

• Empirical studies demonstrate that the selected features (1) yield good classification performance and (2) are relevant to disease diagnosis.

• Extension to higher rank Support Tensor Machines: (P4).

$$\mathcal{W} = \prod_{v=1}^{m} \circ \mathbf{w}^{(v)}$$

$$\mathcal{W} = \sum_{r=1}^{R} \prod_{v=1}^{m} \circ \mathbf{w}_{r}^{(v)}$$
(P2) Subgraph Pattern Mining

Mining Brain Networks using Multiple Side Views for Neurological Disorder Identification,
B. Cao et al., ICDM 2015.

Identifying HIV-induced Subgraph Patterns in Brain Networks with Side Information,
B. Cao et al., Brain Informatics 2015.
Brain Network

Neuroimaging

Brain network
Brain Region Connectivity

Healthy brain

Good family
Brain Region Connectivity

Diseased brain

Connectivity problem
Subgraph Patterns

+ Brain disorder

- Normal control

+ Brain disorder

- Normal control

+ Brain disorder

- Normal control
Multiple Side Views

Side view 1: clinical measures

Side view 2: cognitive measures

Side view 3: immunologic measures

Side view 4: serologic measures

Side view 5: MRI sequence A

Side view 6: MRI sequence B
Data Fusion

• The primary view in graph representation.

• The side views in vector representations.

fMRI, DTI brain networks

clinical, serologic and cognitive measures, etc.
Data Fusion

Late fusion

Early fusion
Side Information Consistency

• The similarity of side view features between instances with the same label is likely to be larger than that with different labels.

\[ A_s^{(p)} = \{ \kappa^{(p)}(i, j) | y_i y_j = 1 \} \]
\[ A_d^{(p)} = \{ \kappa^{(p)}(i, j) | y_i y_j = -1 \} \]

\[ H_0 : \mu_s^{(p)} - \mu_d^{(p)} \leq 0 \]
\[ H_1 : \mu_s^{(p)} - \mu_d^{(p)} > 0 \]
Selection Criterion

- **Side view guidance**

\[
\arg\min_{T \subseteq S} \frac{1}{2} \sum_{p=1}^{v} \lambda^{(p)} \sum_{i,j=1}^{n} \left\| I_T x_i - I_T x_j \right\|_F^2 \Theta^{(p)}(i, j)
\]

\[
\Theta^{(p)}(i, j) = \begin{cases} 
\frac{1}{|H^{(p)}|} & (i, j) \in H^{(p)} \\
-\frac{1}{|L^{(p)}|} & (i, j) \in L^{(p)} 
\end{cases}
\]

- **Direct supervision**

\[
\arg\min_{T \subseteq S} \frac{1}{2} \sum_{i,j=1}^{n} \left\| I_T x_i - I_T x_j \right\|_F^2 \Omega(i, j)
\]

\[
\Omega(i, j) = \begin{cases} 
\frac{1}{|M|} & (i, j) \in M \\
-\frac{1}{|C|} & (i, j) \in C \\
0 & \text{otherwise}
\end{cases}
\]

H is the set of instance pairs with large similarity and L is that with small similarity. M is the set of must-links and C is the set of cannot-links.
Selection Criterion

- Integration

\[
\Phi(i, j) = \Omega(i, j) + \sum_{p=1}^{v} \lambda^{(p)} \Theta^{(p)}(i, j)
\]

\[
\mathcal{F}(\mathcal{T}) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \|I_{\mathcal{T}} x_i - I_{\mathcal{T}} x_j\|_F^2 \Phi(i, j)
\]

\[
= \text{tr}(I_{\mathcal{T}}^\top X (D - \Phi) X^\top I_{\mathcal{T}})
\]

\[
= \text{tr}(I_{\mathcal{T}}^\top X L X^\top I_{\mathcal{T}})
\]

\[
= \sum_{g_i \in \mathcal{T}} f_i^\top L f_i
\]

\[
g\text{Side score} \quad q(g_i) = f_i^\top L f_i
\]
Subgraph Mining

Xifeng Yan and Jiawei Han. "gspan: Graph-based substructure pattern mining." ICDM 2002.

Lower bound of gSide for super-graphs

\[ \hat{q}(g_i) = f_i^T \hat{L} f_i \]

\[ \hat{L}(p, q) = \min(0, L(p, q)) \]
Algorithm 2 gMSV

Input: $\mathcal{D}$ (graph dataset), $\mathcal{Z}$ (side views), $\{\lambda^{(p)}\}_{p=1}^{v}$ (view coefficients), $k$ (number of subgraph patterns to be selected), $\text{min\_sup}$ (minimum support value)

Output: $\mathcal{T}$ (selected subgraph patterns)

1: Initialize $\mathcal{T} = \emptyset$, $\theta = \text{Inf}$
2: while unexplored nodes in DFS code trees $\neq \emptyset$ do
3:   $g =$ currently explored node in DFS code trees
4:   if frequency($g$) $\geq$ min\_sup then
5:     if $|\mathcal{T}| < k$ or $q(g) < \theta$ then
6:       $\mathcal{T} = \mathcal{T} \cup \{g\}$
7:     end if
8:     if $|\mathcal{T}| > k$ then
9:       $g_{\text{max}} = \text{argmax}_{g' \in \mathcal{T}} q(g')$
10:      $\mathcal{T} = \mathcal{T} / \{g_{\text{max}}\}$
11:     end if
12:     $\theta = \text{max}_{g' \in \mathcal{T}} q(g')$
13:   end if
14:   if $\hat{q}(g) < \theta$ then
15:     Depth-first search the subtree rooted from $g$
16:   end if
17: end while
Data Collection

- Chicago Early HIV Infection Study

56 HIV and 21 seronegative controls.

- fMRI and DTI brain networks.

- Neuropsychological tests, flow cytometry, plasma luminex, freesurfer, overall brain microstructure, localized brain microstructure, brain volumetry.
gMSV: the proposed discriminative subgraph selection method using multiple side views.

gSSC: a semi-supervised feature selection method using both labeled and unlabeled graphs.

Conf, Ratio, Gtest, and HSIC: supervised subgraph selection methods based on confidence, frequency ratio, G-test score, and Hilbert-Schmidt Independence Criterion.

Freq: an unsupervised subgraph selection method based on frequency.
Time and Space Complexity

CPU Time Cost (se)

fMRI

DTI

Unpruning
Pruning

min_sup%

10^0 10^1 10^2 10^3

5 10 15

12.5 15.0 17.5 20.0 22.5 25.0

CPU Time Cost (se)

Unpruning
Pruning

min_sup%

10^0 10^1 10^2

5 10 15

12.5 15.0 17.5 20.0 22.5 25.0

# Subgraphs Explored

fMRI

DTI

Unpruning
Pruning

min_sup%

10^4 10^5 10^6 10^7

5 10 15

12.5 15.0 17.5 20.0 22.5 25.0

# Subgraphs Explored

Unpruning
Pruning

min_sup%

10^4 10^5 10^6 10^7

5 10 15

12.5 15.0 17.5 20.0 22.5 25.0
Effects of Side Views

Classification performance with different side views.

<table>
<thead>
<tr>
<th>Brain networks</th>
<th>Side views</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>$F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>fMRI</td>
<td><strong>neuropsychological tests</strong></td>
<td>0.743</td>
<td>0.851</td>
<td>0.679</td>
<td>0.734</td>
</tr>
<tr>
<td></td>
<td><strong>flow cytometry</strong></td>
<td>0.887</td>
<td>0.919</td>
<td>0.872</td>
<td>0.892</td>
</tr>
<tr>
<td></td>
<td><strong>plasma luminex</strong></td>
<td>0.715</td>
<td>0.769</td>
<td>0.682</td>
<td>0.710</td>
</tr>
<tr>
<td></td>
<td><strong>freesurfer</strong></td>
<td>0.786</td>
<td>0.851</td>
<td>0.737</td>
<td>0.785</td>
</tr>
<tr>
<td></td>
<td><strong>overall brain microstructure</strong></td>
<td>0.672</td>
<td>0.824</td>
<td>0.500</td>
<td>0.618</td>
</tr>
<tr>
<td></td>
<td><strong>localized brain microstructure</strong></td>
<td>0.628</td>
<td>0.686</td>
<td>0.605</td>
<td>0.637</td>
</tr>
<tr>
<td></td>
<td><strong>brain volumetry</strong></td>
<td>0.701</td>
<td>0.739</td>
<td>0.737</td>
<td>0.731</td>
</tr>
<tr>
<td></td>
<td><strong>All side views</strong></td>
<td>0.972</td>
<td>1.000</td>
<td>0.949</td>
<td>0.973</td>
</tr>
</tbody>
</table>

| DTI            | **neuropsychological tests**    | 0.616    | 0.630     | 0.705  | 0.662  |
|                | **flow cytometry**              | 0.815    | 0.847     | 0.808  | 0.822  |
|                | **plasma luminex**              | 0.736    | 0.801     | 0.705  | 0.744  |
|                | **freesurfer**                  | 0.631    | 0.664     | 0.632  | 0.644  |
|                | **overall brain microstructure**| 0.604    | 0.626     | 0.679  | 0.647  |
|                | **localized brain microstructure**| 0.723    | 0.717     | 0.775  | 0.741  |
|                | **brain volumetry**             | 0.605    | 0.616     | 0.679  | 0.644  |
|                | **All side views**              | 0.973    | 1.000     | 0.951  | 0.974  |
Selected Subgraph Patterns

Occipital_Sup_L  Occipital_Sup_R

Calcarine_L  Calcarine_R

Lingual_L  Lingual_R
Conclusions

• We propose an efficient subgraph selection method that can effectively explore complementary information in multiple side views.

• Experimental results show that the selected subgraph patterns perform well on disease diagnosis and could facilitate better understanding of brain injury.

• Extension to latent representation learning: (P3).
(P3) Brain Network Embedding

Semi-supervised Tensor Factorization for Brain Network Analysis,
B. Cao et al., ECML/PKDD 2016.
t-BNE: Tensor-based Brain Network Embedding,
B. Cao et al., SDM 2017.
Graph Classification

- Existing works: subgraph patterns and graph-theoretical measures (explicit).
  - Pros: interpretable.
  - Cons: feature engineering and domain knowledge needed.
- This work: an end-to-end approach to learning latent representations (implicit).
Tensor Modeling

- [1] How to preserve the graph property in the tensor factorization process?

Brain network data → Tensor modeling → Partially symmetric tensor factorization

\[ \chi \approx B_{:,1} S_{:,1} + \cdots + B_{:,k} S_{:,k} \]
Tensor Modeling

• [2] How to leverage the side information associated with brain networks?
Tensor Modeling

• [3] How to fuse the classifier training and the representation learning procedures?
Tensor Modeling

- Partially coupled matrix and tensor factorization
Tensor-based Framework

\[
\min_{B,S,W} \left\{ \|X - C \times_1 B \times_2 B \times_3 S\|_F^2 \right\}
\]

\[
+ \alpha \text{tr}(S^T L_Z S) + \beta \|DSW - Y\|_F^2 + \gamma \|W\|_F^2
\]

(3.7)

\[\text{s.t. } S^T S = I\]

[1] Partially symmetric tensor factorization

[2] Side information guidance

[3] Partially coupled matrix and tensor factorization

orthogonality
Optimization

• Alternating Direction Method of Multipliers (ADMM)
  
  • Update the node factor matrix

\[
\mathcal{L}(\mathbf{B}, \mathbf{P}) = \| \mathbf{X} - \mathbf{C} \times_1 \mathbf{B} \times_2 \mathbf{P} \times_3 \mathbf{S} \|_F^2 + \text{tr}(\mathbf{U}^\top(\mathbf{P} - \mathbf{B})) + \frac{\mu}{2} \| \mathbf{P} - \mathbf{B} \|_F^2
\]

• Update the subject factor matrix

\[
\mathcal{L}(\mathbf{S}) = \| \mathbf{S}\mathbf{G}^\top - \mathbf{X}_{(3)} \|_F^2 + \alpha \text{tr}(\mathbf{S}^\top \mathbf{L}_\mathbf{Z} \mathbf{S}) + \beta \| \mathbf{DSW} - \mathbf{Y} \|_F^2
\]

\[
\text{s.t. } \mathbf{S}^\top \mathbf{S} = \mathbf{I}
\]

• Update the weight matrix

\[
\mathcal{L}(\mathbf{W}) = \| \mathbf{DSW} - \mathbf{Y} \|_F^2 + \gamma \| \mathbf{W} \|_F^2
\]
Algorithm 3 tBNE

Input: $\mathcal{X}$ (stacked brain networks), $Z$ (auxiliary features), $Y$ (class labels), $k$ (rank of tensor factorization), $\alpha$ (weight of side information), $\beta$ (weight of classification loss), $\gamma$ (weight of regularization)

Output: $B$ (node factors), $S$ (subject factors), $W$ (classification weights)

1: $\mu_{\text{max}} = 10^6$, $\rho = 1.15$
2: Initialize $B, S, W \sim N(0, 1)$, $U = 0$, $\mu = 10^{-6}$
3: repeat
4: Update $B$ and $P$ by Equation 4.10 and Equation 4.12
5: Update $U$ by Equation 4.13
6: Update $\mu$ by $\mu \leftarrow \min(\rho \mu, \mu_{\text{max}})$
7: Update $S$ by Equation 4.15 with the curvilinear search
8: Update $W$ by Equation 4.17
9: until convergence
Data Collection

- 3 emotion regulation tasks: *neutral*, *maintain*, and *reappraise*.

- 37 patients with anxiety disorder and 32 healthy participants.

- 34 scalp channels (nodes).
Classification Performance

![Graph classification accuracy](https://github.com/caobokai/tBNE)

<table>
<thead>
<tr>
<th>Methods</th>
<th>Datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Neutral</td>
</tr>
<tr>
<td>tBNE</td>
<td>0.7833</td>
</tr>
<tr>
<td>CMTF</td>
<td>0.5810</td>
</tr>
<tr>
<td>Rubik</td>
<td>0.6405</td>
</tr>
<tr>
<td>ALS</td>
<td>0.6119</td>
</tr>
<tr>
<td>gMSV</td>
<td>0.6500</td>
</tr>
<tr>
<td>CC</td>
<td>0.5357</td>
</tr>
</tbody>
</table>

- **tBNE**: the proposed tensor factorization model for brain network embedding.
- **CMTF**: coupled matrix and tensor factorization where brain networks and side information are coupled in the subject mode.
- **Rubik**: tensor factorization with orthogonality and sparsity constraints.
- **ALS**: tensor factorization using alternating least squares without any constraint.
- **gMSV**: a discriminative subgraph selection approach using side information.
- **CC**: extracting local clustering coefficients as features.
Parameter Sensitivity

- **Neutral**
- **Maintain**
- **Reappraise**

**k**

**γ**

**α**

**β**

Accuracy vs. Parameter Values
Factor Analysis

- Neutral

- Maintain

- Reappraise
Extensions

• **Guidance**: column-wise guidance from community information [Wang et al., KDD’15].

• **Supervision**: must-link, cannot-link, and separability [Kong and Yu, KDD’10].

• **Multimodality**: joint tensor factorization to capture consensus information across multiple modalities (e.g., EEG, fMRI, and DTI) [Liu et al., AAAI’18].
Conclusions

• We propose a tensor-based framework to learn the brain network embedding that (1) preserves the symmetric graph property, (2) incorporates side information guidance, and (3) fuses the classifier learning procedure and tensor factorization.

• Experimental results show that the obtained latent factors are discriminative and could facilitate better understanding of brain mechanism with anxiety disorder under different emotion regulations.
(P4) Multi-view Sequence Prediction

Multi-view Machines,
B. Cao et al., WSDM 2016.

DeepMood: Modeling Mobile Phone Typing Dynamics for Mood Detection,
B. Cao et al., KDD 2017.
Mobile Health

- Bipolar disorder affects approximately 5.7 million adult Americans, or about 2.6% of the U.S. population age 18 and older every year. (NIMH)

- Mobile phones, in particular, "smartphones" have become near ubiquitous with 2 billion smartphone users worldwide.

Keyboard  Accelerometer  Gyroscope  GPS  Microphone
Challenges

• **Unaligned views**: Features defined in one view are usually missing for data points collected in another view.

• **Dominant views**: Dense views may dominate a concatenated feature space and override the effects of other important views.

• **View interactions**: Multi-view time series from typing dynamics contains complementary information about a person’s mental health.
Late Fusion

How to fuse?

Alphanumeric characters

Special characters

Accelerometer values
Support Vector Machines (SVM)

- [Vapnik 1995] The linear SVM model (and LR etc.) is limited to the first-order feature interactions.
Support Tensor Machines (STM)

- [Cao et al. 2014] The STM model explores only the highest-order feature interactions.
Factorization Machines (FM)

- [Rendle 2010] The FM model uses separate parameters to approximate interactions in different orders.
Multi-view Machines (MVM)

- [Cao et al. 2016] The MVM model includes the full-order interactions across multi-view features and jointly factorizes the interaction parameters in different orders.
\[ \hat{y} = \beta_0 + \sum_{p=1}^{m} \sum_{i_p=1}^{I_p} \beta^{(p)}_{i_p} x^{(p)}_{i_p} \]

\text{global bias}

\[ + \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \beta^{(1,2)}_{i_1,i_2} x^{(1)}_{i_1} x^{(2)}_{i_2} + \cdots + \sum_{i_{m-1}=1}^{I_{m-1}} \sum_{i_m=1}^{I_m} \beta^{(m-1,m)}_{i_{m-1},i_m} x^{(m-1)}_{i_{m-1}} x^{(m)}_{i_m} \]

\text{first-order interactions}

\[ + \cdots + \sum_{i_1=1}^{I_1} \cdots \sum_{i_m=1}^{I_m} \beta_{i_1,\ldots,i_m} \left( \prod_{p=1}^{m} x^{(p)}_{i_p} \right) \]

\text{second-order interactions}

\[ + \cdots + \sum_{i_1=1}^{I_1+1} \cdots \sum_{i_m=1}^{I_m+1} w_{i_1,\ldots,i_m} \left( \prod_{p=1}^{m} z^{(p)}_{i_p} \right) \]

\text{mth-order interactions}

\[ \hat{y} = \sum_{i_1=1}^{I_1+1} \cdots \sum_{i_m=1}^{I_m+1} \left( \prod_{p=1}^{m} z^{(p)}_{i_p} \right) \left( \sum_{f=1}^{k} \prod_{p=1}^{m} a^{(p)}_{i_p,f} \right) \]

\[ \mathcal{W} = C \times_1 A^{(1)} \times_2 \cdots \times_m A^{(m)} \]
FC Layer

• Using a fully connected layer for data fusion.

\[ f(x) = W^{(2)}x \]

\[ f(x) = \text{relu}(x) \]

\[ f(x) = W^{(1)}x \]

Intermediate representations, e.g., GRU outputs
• Using a Factorization Machine layer for data fusion.

$$f(x) = \text{sum}(x)$$

$$f(x) = x \odot x$$

$$f(x) = Wx$$
MVM Layer

• Using a Multi-view Machine layer for data fusion.

\[ f(x) = \text{sum}(x) \]

\[ f(\{x\}_{i=1}^m) = \bigoplus_{i=1}^m x_i \]

\[ f(x) = Wx \]
DeepMood Framework

Alphanumeric characters

Special characters

Accelerometer values

First-order interactions

Second-order interactions

Third-order interactions

Global bias

Mood score

Dropout

GRU

Dropout

GRU

Dropout

GRU

Dropout
Data Collection

- **Participants**: 7 bipolar I, 5 bipolar II, and 8 controls.
- **Features**: keypress metadata and accelerometer movement.
- **Labels**: Hamilton Depression Rating Scale (HDRS) and Young Mania Rating Scale (YMRS).

### Statistics of the BiAffect dataset.

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Alphanum</th>
<th>Special</th>
<th>Accel</th>
</tr>
</thead>
<tbody>
<tr>
<td># data points</td>
<td>836,027</td>
<td>538,520</td>
<td>14,237,503</td>
</tr>
<tr>
<td># sessions</td>
<td>34,993</td>
<td>33,385</td>
<td>37,647</td>
</tr>
<tr>
<td>Mean length</td>
<td>24</td>
<td>16</td>
<td>378</td>
</tr>
<tr>
<td>Median length</td>
<td>14</td>
<td>9</td>
<td>259</td>
</tr>
<tr>
<td>Maximum length</td>
<td>538</td>
<td>437</td>
<td>90,193</td>
</tr>
</tbody>
</table>
Alphanumeric Characters
Special Characters
Accelerometer Values
• **DeepMood-MVM, DeepMood-FM, and DeepMood-FC**: the proposed DeepMood model with an MVM layer, an FM layer, and an FC layer for data fusion.

• **XGBoost**: a tree boosting system.

• **SVM**: Linear Support Vector Classification/Regression.

• **LR**: Logistic/Ridge Regression.
Convergence Efficiency

![Convergence Efficiency Graph]

Test accuracy

Training epoch

- Red: DeepMood-MVM
- Blue: DeepMood-FM
- Purple: DeepMood-FC
- Grey: XGBoost
# Importance of Different Views

## Prediction performance with different views of typing dynamics.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Classification</th>
<th>Regression</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>$F_1$</td>
<td>RMSE</td>
</tr>
<tr>
<td>DeepMood-MVM w/o Alphanumeric</td>
<td>0.8125</td>
<td>0.8164</td>
<td>3.9833</td>
</tr>
<tr>
<td>DeepMood-MVM w/o Special</td>
<td>0.9008</td>
<td>0.9034</td>
<td>3.8166</td>
</tr>
<tr>
<td>DeepMood-MVM w/o Accel</td>
<td>0.8318</td>
<td>0.8253</td>
<td>3.9499</td>
</tr>
<tr>
<td>DeepMood-MVM w/ all</td>
<td>0.9031</td>
<td>0.9070</td>
<td>3.5664</td>
</tr>
<tr>
<td>DeepMood-FM w/ Alphanumeric</td>
<td>0.8322</td>
<td>0.8224</td>
<td>3.9515</td>
</tr>
<tr>
<td>DeepMood-FM w/ Special</td>
<td>0.6260</td>
<td>0.5676</td>
<td>4.1040</td>
</tr>
<tr>
<td>DeepMood-FM w/ Accel</td>
<td>0.8015</td>
<td>0.8089</td>
<td>3.9722</td>
</tr>
<tr>
<td>DeepMood-FM w/ all</td>
<td>0.9021</td>
<td>0.9011</td>
<td>3.6767</td>
</tr>
</tbody>
</table>
Conclusions

• We validate that mobile phone metadata could be used to predict the presence of mood disorders, and DeepMood achieves 90.31% prediction accuracy.

• The ability to passively collect data that can be used to infer the presence and severity of mood disturbances may enable us to provide interventions to more patients earlier in their mood episodes.
Summary

• (P1) **Multi-view feature selection**: to determine important biomarkers from measures collected from multiple medical examinations.

• (P2) **Subgraph pattern mining**: to identify connectivity patterns in the brain that are associated with brain injury using side information.

• (P3) **Brain network embedding**: to obtain effective brain network representations with the guidance of side information and other constraints.

• (P4) **Multi-view sequence prediction**: to monitor the mental health using time series data collected from multiple sensors on a mobile and wearable device.

• **Drug discovery in heterogeneous bioinformatics networks** (KDD’13, ICDM’14).

• **Brain network clustering** (ECML/PKDD’16, AAAI’18).

• **Deep learning on brain network and neuroimaging data** (SDM’16, KDD’17, SAC’17).
A Final Note

• **Federated learning**: deploying intelligent healthcare services on mobile devices while preserving user privacy.

• **Meta learning**: transferring and sharing medical knowledge across users.

• **Reinforcement learning**: providing effective interventions to users.

• **Broad learning**: fusing heterogeneous signals for synergistic knowledge discovery.
References

- Guixiang Ma, Lifang He, Bokai Cao, Jiawei Zhang, and Philip S. Yu. Multi-graph clustering based on interior-node topology with applications to brain networks. ECML/PKDD, 2016.
Broad Learning for Healthcare

Q & A

Bokai Cao
caobokai.why@gmail.com