Metric learning pairwise kernel for graph inference with SVM

Jean-Philippe Vert$^1$ Jian Qiu$^2$ William Stafford Noble$^3$

$^1$Center for Computational Biology
Ecole des Mines de Paris

$^2$Department of Genome Sciences
University of Washington

$^2$Department of Genome Sciences
Department of Computer Science and Engineering
University of Washington

NIPS workshop “Open problems in computational biology”, Whistler, Canada, December 8th, 2006
Motivation

Data
- Gene expression,
- Gene sequence,
- Protein localization, ...

Graph
- Protein-protein interactions,
- Metabolic pathways,
- Signaling pathways, ...

Vert, Qiu and Noble ()
Strategies

Unsupervised approaches

The graph is completely unknown

- **model-based** approaches: Bayes nets, dynamical systems,..
- **similarity-based**: connect similar nodes

Supervised approaches

Part of the graph is known

- **Undirect approach** (ML): a pair \((a, b)\) is likely to be connected if it is similar to a pair \((c, d)\) which is connected.
- **Direct approach**: \(a\) is likely to be connected to \(b\) if they are similar to each other, for a measure of similarity to be optimized.
Strategies

Unsupervised approaches

The graph is completely unknown
- **model-based** approaches: Bayes nets, dynamical systems,..
- **similarity-based**: connect similar nodes

Supervised approaches

Part of the graph is known
- **Undirect approach** (ML): a pair \((a, b)\) is likely to be connected if it is similar to a pair \((c, d)\) which is connected.
- **Direct approach**: \(a\) is likely to be connected to \(b\) if they are similar to each other, for a measure of similarity to be optimized
Example: supervised undirect approach with SVM

TPPK kernel

- Suppose a kernel $K$ exists for individual genes.
- Construct a tensor product pairwise kernel (TPPK) between pairs (Ben-Hur and Noble, 2005):

$$K_{TPPK} \left((x_1, x_2), (x_3, x_4)\right) = K(x_1, x_3)K(x_2, x_4) + K(x_1, x_4)K(x_2, x_3).$$

- Given a training set of connected and non-connected pairs, train a binary SVM to predict if a new pair is connected or not.
Example: supervised direct approach

Metric learning

- Suppose a kernel $K$ exists for individual genes.
- Given a training set of connected and non-connected pairs, find a new distance metric such that connected genes are closer to each other than non-connected pairs.
- For a new pair, predict an edge is the distance is below a threshold (Vert and Yamanishi, 2005).
Contribution

- A new distance metric learning algorithm for supervised graph inference
- A relaxation equivalent to a SVM with a particular kernel for pairs, which we call metric learning pairwise kernel (MLPK).
- Therefore the possibility to:
  - use out-of-the-box SVM implementation for supervised direct graph inference
  - easily combine direct and undirect approaches by kernel combination.
Objective function

Suppose genes are described as vectors $x \in \mathbb{R}^d$

After a linear mapping $\Phi(x) = Ax$ the square Euclidean distance is:

$$d_M(x, x') = (x - x')^\top M(x - x')$$

$$= tr \left( M(x - x')(x - x')^\top \right),$$

with $M = A^\top A \succ 0$.

Direct edge inference is possible if, for example,

$$d_\phi(x_i, x_j) \begin{cases} 
\leq \gamma - 1 & \text{for } x_i \sim x_j, \\
\geq \gamma + 1 & \text{for } x_i \not\sim x_j.
\end{cases}$$
Large-margin metric learning

Problem formulation

In the spirit of SVM, this suggests the following optimization problem:

Minimize \[ \| M \|_{Fro}^2 + C \sum_{(i,j)} \zeta_{i,j} \]

subject to \[ \zeta_{i,j} \geq 0, \quad \forall (i,j) \in \mathcal{T} \]
\[ d_M(x_i, x_j) \leq \gamma - 1 + \zeta_{i,j}, \quad i \sim j \]
\[ d_M(x_i, x_j) \geq \gamma + 1 - \zeta_{i,j}, \quad i \not\sim j \]
\[ M \succeq 0. \]

Similar to Tsang and Kwok (2003)
Representation of the solution

Lemma

The solution $\hat{M}$ of the previous problem can be expanded as:

$$\hat{M} = \sum_{(i,j) \in T} \alpha_{i,j} (x_i - x_j) (x_i - x_j)^\top,$$

where $\alpha_{i,j}$ are real number associated to the training pairs $(i,j) \in T$.

Remarks

- The “classical” representer theorem can not be applied here because of the $M \succeq 0$ constraint. A slight extension is required.
- Plugging back into the problem leads to a SDP problem.
- The dimension of the SDP is $2 |T| + 1 \implies$ computational issues.
Lemma

The solution $\hat{M}$ of the previous problem can be expanded as:

$$\hat{M} = \sum_{(i,j) \in T} \alpha_{i,j} (x_i - x_j) (x_i - x_j)^\top,$$

where $\alpha_{i,j}$ are real number associated to the training pairs $(i, j) \in T$.

Remarks

- The “classical” representer theorem can not be applied here because of the $M \succeq 0$ constraint. A slight extension is required.
- Plugging back into the problem leads to a SDP problem
- The dimension of the SDP is $2 |T| + 1 \implies$ computational issues.
Relaxation

- **Relax the constraint** $M \succeq 0$

- The problem becomes:

  \[
  \text{Minimize} \quad \|M\|_F^2 + C \sum_{(i,j)} \zeta_{i,j}
  \]

  \[
  \text{subject to} \quad \zeta_{i,j} \geq 0, \quad \forall (i,j)
  \]

  \[
  < M, D_{i,j} >_F - \gamma \leq -1 + \zeta_{i,j}, \quad i \sim j
  \]

  \[
  < M, D_{i,j} >_F - \gamma \geq 1 - \zeta_{i,j}, \quad i \not\sim j.
  \]

  with the notation

  \[
  D_{i,j} = (x_i - x_j) (x_i - x_j)^T.
  \]

- This is a **SVM over** $d \times d$ matrices with training set $\{(D_{i,j})_{(i,j) \in \mathcal{T}}\}$.
Relaxation

- Relax the constraint $M \succeq 0$
- The problem becomes:

$$\begin{align*}
\text{Minimize} & \quad \| M \|_{Fro}^2 + C \sum_{(i,j)} \zeta_{i,j} \\
\text{subject to} & \quad \zeta_{i,j} \geq 0, \quad \forall (i, j) \\
& \quad < M, D_{i,j} >_{Fro} - \gamma \leq -1 + \zeta_{i,j}, \quad i \sim j \\
& \quad < M, D_{i,j} >_{Fro} - \gamma \geq 1 - \zeta_{i,j}, \quad i \not\sim j.
\end{align*}$$

with the notation

$$D_{i,j} = (x_i - x_j)(x_i - x_j)\top.$$ 

This is a SVM over $d \times d$ matrices with training set $(D_{i,j})_{(i,j) \in \mathcal{T}}$. 

Vert, Qiu and Noble ()
Graph inferences with SVM
Relax the constraint $M \succeq 0$

The problem becomes:

Minimize $\| M \|_{\text{Fro}}^2 + C \sum_{(i,j)} \zeta_{i,j}$

subject to $\zeta_{i,j} \geq 0$, $\forall (i, j)$

$< M, D_{i,j} >_{\text{Fro}} - \gamma \leq -1 + \zeta_{i,j}$, $i \sim j$

$< M, D_{i,j} >_{\text{Fro}} - \gamma \geq 1 - \zeta_{i,j}$, $i \not\sim j$.

with the notation

$$D_{i,j} = (x_i - x_j) (x_i - x_j)^\top.$$ 

This is a SVM over $d \times d$ matrices with training set $\{ D_{i,j} \}_{(i,j) \in \mathcal{T}}$. 
Metric learning pairwise kernel

Pairwise inner product

\[
\langle D_{x_1, x_2}, D_{x_3, x_4}\rangle_{\text{Fro}} = \operatorname{Trace}\left( (x_1 - x_2) (x_1 - x_2)^\top (x_3 - x_4) (x_3 - x_4)^\top \right)
\]

\[
= \left( (x_1 - x_2)^\top (x_3 - x_4) \right)^2
\]

\[
= \left( x_1^\top x_3 - x_1^\top x_4 - x_2^\top x_3 + x_2^\top x_4 \right)^2 .
\]

Pairwise kernel

\[
K_{\text{MLPK}} \left( (x_1, x_2), (x_3, x_4) \right)
\]

\[
= \left( K (x_1, x_3) - K (x_1, x_4) - K (x_2, x_3) + K (x_2, x_4) \right)^2 .
\]
## Results

### Metabolic network
- 769 vertices: enzymes
- 3702 edges: catalyze successive reactions
- 3702 negative pairs (randomly sampled)
- 5-fold CV, 3 repeats, parameter optimization over the training set

### Table

<table>
<thead>
<tr>
<th>Data</th>
<th>MLPK</th>
<th>TPPK</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>AUC</td>
</tr>
<tr>
<td>Expression</td>
<td>77.8 ± 0.2</td>
<td>84.5 ± 0.1</td>
</tr>
<tr>
<td>Localization</td>
<td>63.9 ± 0.4</td>
<td>68.2 ± 0.4</td>
</tr>
<tr>
<td>Phylogenetic profile</td>
<td>79.8 ± 0.1</td>
<td>84.9 ± 0.2</td>
</tr>
<tr>
<td>Yeast two-hybrid</td>
<td>76.6 ± 0.2</td>
<td>82.0 ± 0.1</td>
</tr>
<tr>
<td>Sum</td>
<td>83.9 ± 0.4</td>
<td>90.9 ± 0.3</td>
</tr>
</tbody>
</table>

Vert, Qiu and Noble ()

Graph inferences with SVM
## Results

### Co-complex network
- 797 vertices: **proteins**
- 3280 edges: **member of the same complex**
- 3280 negative pairs (randomly sampled)
- 5-fold CV, 3 repeats, parameter optimization over the training set

<table>
<thead>
<tr>
<th>Data</th>
<th>MLPK</th>
<th>TPPK</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>AUC</td>
</tr>
<tr>
<td>Localization</td>
<td>76.5 ± 0.1</td>
<td>76.8 ± 0.1</td>
</tr>
<tr>
<td>Chip-chip</td>
<td>82.4 ± 0.3</td>
<td>89.7 ± 0.1</td>
</tr>
<tr>
<td>Pfam</td>
<td>92.2 ± 0.2</td>
<td>98.2 ± 0.1</td>
</tr>
<tr>
<td>PSI-BLAST</td>
<td>90.0 ± 0.3</td>
<td>97.3 ± 0.1</td>
</tr>
</tbody>
</table>
Conclusion and future work

Summary

- A kernel method for distance metric learning, with an objective function optimized for graph inference
- A relaxation that leads to a SVM with a particular kernel for pairs
- Encouraging experimental results

Future work

- Assess the effect of relaxation
- Integration of multiple pairwise kernels

Reference