Information-Theoretic Network Inference

P. E. Meyer, K. Kontos and G. Bontempi

Machine Learning Group, Université Libre de Bruxelles

CIL contact day
1. Introduction
2. State of the Art
3. MRNet
4. Experiments
5. Results and Conclusion
Example: Gene Network

Gene interaction:

- Biological dogma: gene → RNA → protein
- A protein can block or activate another gene

Network:

- Each node of the network is gene (a variable).
- There is a link between two nodes if there is a direct interaction between them.

Interests:

- Knowledge representation
- Reverse engineering
- Drug discovery
Principle of Network Inference

- **input:** Data, $m \times n$ matrix, where $DATA_{ij}$ is the RNA-concentration of $G_i$ at experiment $Exp_j$ (microarray)
- **output:** Network, $n \times n$ symmetric matrix, where $NET_{ij}$ is the probability of a direct interaction between $G_i$ and $G_j$
Known methods

- Boolean Networks
- Bayesian Networks
- Differential Equation Networks
- Association Networks
  - Partial Correlation
  - Mutual Information (Information-Theoretic Networks)
Relevance Network

Definition (Mutual Information)

The *mutual information* between two random variables $X_i$ and $X_j$ is defined as,

$$I(X_i; X_j) = \sum_{x_i \in \mathcal{X}} \sum_{x_j \in \mathcal{X}} p(x_i, x_j) \log \left( \frac{p(x_i, x_j)}{p(x_i)p(x_j)} \right)$$  \hspace{1cm} (1)$$

Discretize data + compute MI for all couples of genes

<table>
<thead>
<tr>
<th>MIM</th>
<th>$G_1$</th>
<th>$G_2$</th>
<th>...</th>
<th>$G_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G_1$</td>
<td>-</td>
<td>$I(G_1; G_2)$</td>
<td>...</td>
<td>$I(G_1; G_n)$</td>
</tr>
<tr>
<td>$G_2$</td>
<td>$I(G_1; G_2)$</td>
<td>-</td>
<td>...</td>
<td>$I(G_2; G_n)$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$G_n$</td>
<td>$I(G_1; G_n)$</td>
<td>$I(G_2; G_n)$</td>
<td>...</td>
<td>-</td>
</tr>
</tbody>
</table>
RELNET: False Positive Trends

- Normalize the matrix (MIM) and consider it as the inferred network [Butte and Kohane, 2000]

- The method is $O(m \times n^2)$

- False Positive Trends:
  Assume $G_1$ influence $G_3$ through $G_2$

  $$G_1 \rightarrow G_2 \rightarrow G_3$$

  Then $I(G_1; G_2)$ and $I(G_2; G_3)$ will be high
  but also $I(G_1; G_3) \rightarrow$ add false link between $G_1$ and $G_3$
Algorithm for the Reconstruction of Accurate Cellular Network [Margolin et al., 2006]

There are three cases of indirect interaction with three variables:

- $G_1 \rightarrow G_2 \rightarrow G_3$
- $G_1 \leftarrow G_2 \rightarrow G_3$
- $G_1 \rightarrow G_2 \leftarrow G_3$

Whatever the case, $I(G_1; G_3) < I(G_1; G_2)$ and $I(G_1; G_3) < I(G_2; G_3)$ by the data processing inequality

- For all triples of genes suppress the weakest link among them.
ARACNE: False Negative Trends

- Aracne is $O(m \times n^2 + n^3)$
- False Negative Trends:
  Assume a triple interaction

The algorithm will suppress a good link
The minimum redundancy - maximum relevance (MRMR) criterion [Peng and Long, 2004] consists in

- selecting the variable that maximizes $u_i$, the relevance to the output $Y$,

$$u_i = I(X_i; Y)$$  \hspace{1cm} (2)

- and that minimizes the mean redundancy $z_i$ with the already selected variable,

$$z_i = \frac{1}{d} \sum_{X_j \in X_S} I(X_i; X_j)$$  \hspace{1cm} (3)

$$X_{MRMR} = \arg \max_{X_i \in X_{-S}} \{u_i - z_i\}$$  \hspace{1cm} (4)
Mrmr Example

\[ \text{H}(Y) \cap \text{H}(X_1) \cap \text{H}(X_2) \]
#### Mrmr Example

- **H(Y)**
- **H(X1)**
- **H(X3)**
MrMr Example
Minimum Redundancy - Maximum Relevance

- Greedy approach selecting the variable with best trade-off relevance-redundancy
- Selection of a subset of variables (composed of the most independent ones)

Network Inference Algorithm:
- Compute the MIM, $O(m \times n^2)$
- For variable $X_1$, compute the MRMR score of all the other variables, $O(n^2)$
- Repeat the operation for all variables, $O(n^3)$
- Normalize the score matrix, $O(n^2)$
- The method is $O(m \times n^2 + n^3)$
Experimental Framework

Network and Data Generator

Original Network → Artificial Dataset

Entropy Estimator

Mutual Information Matrix → Inferred Network

Validation Procedure

Precision-Recall Curves and F-Scores
Validation

<table>
<thead>
<tr>
<th>edge</th>
<th>actual positive</th>
<th>actual negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>inferred positive</td>
<td>TP</td>
<td>FP</td>
</tr>
<tr>
<td>inferred negative</td>
<td>FN</td>
<td>TN</td>
</tr>
</tbody>
</table>

**Table:** Confusion matrix.

Precision and Recall:

\[
p = \frac{TP}{TP + FP}, \quad r = \frac{TP}{TP + FN}
\]

F-Scores:

\[
F_\beta = \left(1 + \beta^2\right)\frac{pr}{r + \beta^2p},
\]

A weighted harmonic average of precision and recall.
Datasets

**Table:** The six artificial datasets generated, where \( n \) is the number of genes and \( m \) is the number of samples.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Generator</th>
<th>Topology</th>
<th>( n )</th>
<th>( m )</th>
</tr>
</thead>
<tbody>
<tr>
<td>dR1</td>
<td>sRogers</td>
<td>power-law tail</td>
<td>2000</td>
<td>1000</td>
</tr>
<tr>
<td>dR2</td>
<td>sRogers</td>
<td>power-law tail</td>
<td>1000</td>
<td>750</td>
</tr>
<tr>
<td>dR3</td>
<td>sRogers</td>
<td>power-law tail</td>
<td>600</td>
<td>600</td>
</tr>
<tr>
<td>dS1</td>
<td>SynTReN</td>
<td>( E. coli )</td>
<td>500</td>
<td>500</td>
</tr>
<tr>
<td>dS2</td>
<td>SynTReN</td>
<td>( E. coli )</td>
<td>300</td>
<td>300</td>
</tr>
<tr>
<td>dS3</td>
<td>SynTReN</td>
<td>( E. coli )</td>
<td>50</td>
<td>500</td>
</tr>
</tbody>
</table>
F-scores with $\beta = 1$ (precision as important as recall). The best score for each dataset is in boldface.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>RelNet</th>
<th>ARACNE</th>
<th>MRNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.24</td>
<td>0.28</td>
<td>0.26</td>
</tr>
<tr>
<td>2</td>
<td>0.25</td>
<td>0.36</td>
<td>0.29</td>
</tr>
<tr>
<td>3</td>
<td>0.25</td>
<td>0.45</td>
<td>0.45</td>
</tr>
<tr>
<td>4</td>
<td>0.09</td>
<td>0.06</td>
<td>0.10</td>
</tr>
<tr>
<td>5</td>
<td>0.16</td>
<td>0.12</td>
<td>0.19</td>
</tr>
<tr>
<td>6</td>
<td>0.18</td>
<td>0.11</td>
<td>0.24</td>
</tr>
</tbody>
</table>
F-Scores

F-scores with $\beta = 0.5$ (precision more important than recall). The best score for each dataset is in boldface.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>RelNet</th>
<th>ARACNE</th>
<th>MRNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.29</td>
<td>0.37</td>
<td>0.38</td>
</tr>
<tr>
<td>2</td>
<td>0.31</td>
<td>0.38</td>
<td>0.39</td>
</tr>
<tr>
<td>3</td>
<td>0.32</td>
<td>0.49</td>
<td>0.52</td>
</tr>
<tr>
<td>4</td>
<td>0.07</td>
<td>0.08</td>
<td>0.13</td>
</tr>
<tr>
<td>5</td>
<td>0.13</td>
<td>0.14</td>
<td>0.15</td>
</tr>
<tr>
<td>6</td>
<td>0.13</td>
<td>0.15</td>
<td>0.20</td>
</tr>
</tbody>
</table>
Curves: DR3 (600,600)

PR-curve

- MRNET
- RELNET
- ARACNe
Conclusions and Future Works

Further work will focus on:

- the significativity of performances
- the robustness of the inference to noise and to the mutual information estimator
- analyzing real biological datasets


http://www.ulb.ac.be/di/mlg/