

Information-Theoretic Network Inference

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

Machine Learning Group,
Université Libre de Bruxelles

CIL contact day

Outline

Outline

Introduction

State of the Art

MRNet

Experiments

Results and Conclusion

- 1 Outline
- 2 Introduction
- 3 State of the Art
- 4 MRNet
- 5 Experiments
- 6 Results and Conclusion

Example: Gene Network

Outline

Introduction

State of the Art

MRNet

Experiments

Results and Conclusion

Gene interaction:

- Biological dogma: gene \rightarrow RNA \rightarrow 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

Outline

Introduction

State of the Art

MRNet

Experiments

Results and Conclusion

- 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

| DATA | G_1 | G_2 | ... | G_n |
|-------|-------|-------|-----|-------|
| Exp 1 | 0.1 | 0.9 | ... | 0.5 |
| Exp 2 | 0.4 | 0.7 | ... | 0.1 |
| Exp 3 | 0.6 | 0.2 | ... | 0.7 |
| ... | ... | ... | ... | ... |
| Exp m | 0.2 | 0.3 | ... | 0.8 |



| NET | G_1 | G_2 | ... | G_n |
|-------|-------|-------|-----|-------|
| G_1 | - | 0.3 | ... | 0.8 |
| G_2 | 0.3 | - | ... | 0.6 |
| ... | ... | ... | - | ... |
| G_n | 0.8 | 0.6 | ... | - |

Known methods

Outline

Introduction

State of the Art

MRNet

Experiments

Results and Conclusion

- 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,

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

Discretize data + compute MI for all couples of genes

| MIM | G_1 | G_2 | ... | G_n |
|-------|---------------|---------------|-----|---------------|
| G_1 | - | $I(G_1; G_2)$ | ... | $I(G_1; G_n)$ |
| G_2 | $I(G_1; G_2)$ | - | ... | $I(G_2; G_n)$ |
| ... | ... | ... | - | ... |
| G_n | $I(G_1; G_n)$ | $I(G_2; G_n)$ | ... | - |

RELNET: False Positive Trends

Outline

Introduction

State of the Art

MRNet

Experiments

Results and Conclusion

- 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

Outline

Introduction

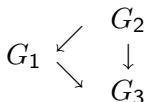
State of the Art

MRNet

Experiments

Results and Conclusion

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



The algorithm will suppress a good link

Minimum Redundancy - Maximum Relevance

Outline

Introduction

State of the Art

MRNet

Experiments

Results and Conclusion

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 ,

$$u_i = I(X_i; Y) \quad (2)$$

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

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

$$X_{MRMR} = \arg \max_{X_i \in X_{-S}} \{u_i - z_i\} \quad (4)$$

Mrmr Example

Outline

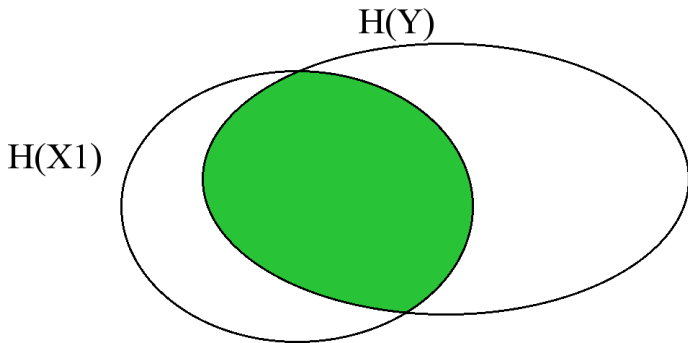
Introduction

State of the Art

MRNet

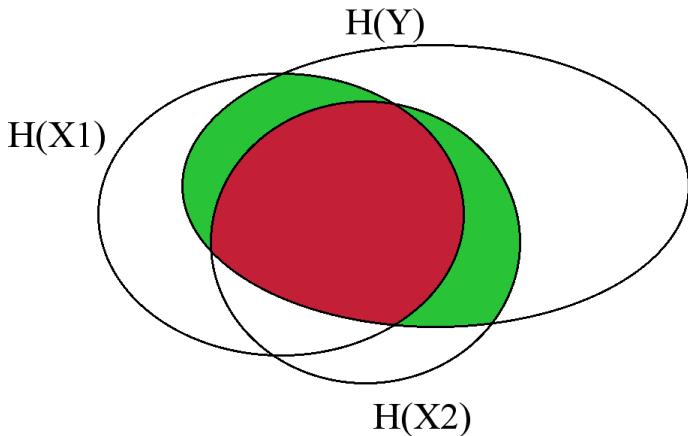
Experiments

Results and Conclusion

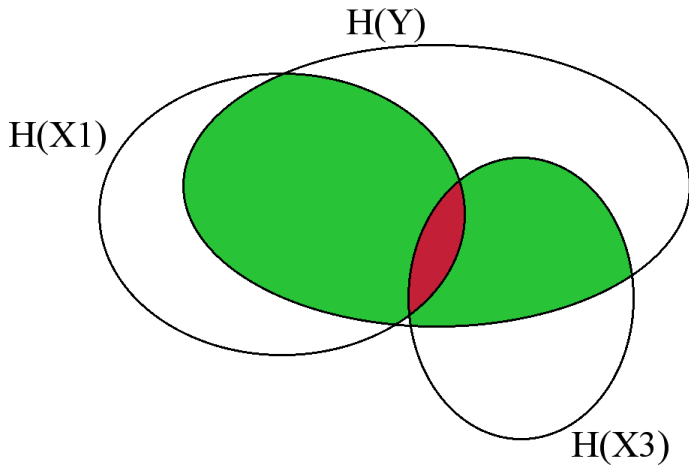


Mrmr Example

- Outline
- Introduction
- State of the Art
- MRNet
- Experiments
- Results and Conclusion



Mrmr Example



Outline

Introduction

State of the Art

MRNet

Experiments

Results and Conclusion

Minimum Redundancy - Maximum Relevance

Outline

Introduction

State of the Art

MRNet

Experiments

Results and Conclusion

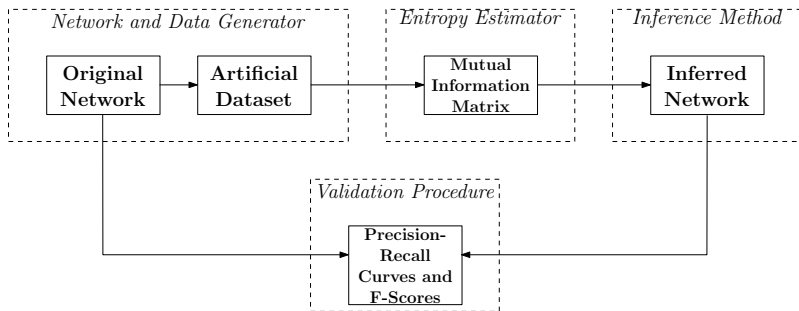
- Greedy approach selecting the variable with best trade-off relevance-redundancy
- Selection of a subset of variables (composed of the most independent ones)

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

- Outline
- Introduction
- State of the Art
- MRNet
- Experiments
- Results and Conclusion



Validation

Outline

Introduction

State of the Art

MRNet

Experiments

Results and Conclusion

Table: Confusion matrix.

| edge | actual positive | actual negative |
|-------------------|-----------------|-----------------|
| inferred positive | TP | FP |
| inferred negative | FN | TN |

Precision and Recall:

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

F-Scores:

$$F_{\beta} = (1 + \beta^2) \frac{pr}{r + \beta^2 p},$$

A weighted harmonic average of precision and recall.

Datasets

Outline

Introduction

State of the Art

MRNet

Experiments

Results and Conclusion

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

| Dataset | Generator | Topology | n | m |
|---------|-----------|----------------|------|------|
| dR1 | sRogers | power-law tail | 2000 | 1000 |
| dR2 | sRogers | power-law tail | 1000 | 750 |
| dR3 | sRogers | power-law tail | 600 | 600 |
| dS1 | SynTReN | <i>E. coli</i> | 500 | 500 |
| dS2 | SynTReN | <i>E. coli</i> | 300 | 300 |
| dS3 | SynTReN | <i>E. coli</i> | 50 | 500 |

F-Scores

Outline

Introduction

State of the Art

MRNet

Experiments

Results and Conclusion

F-scores with $\beta = 1$ (precision as important as recall). The best score for each dataset is in boldface.

| Dataset | RelNet | ARACNE | MRNet |
|---------|--------|-------------|-------------|
| 1 | 0.24 | 0.28 | 0.26 |
| 2 | 0.25 | 0.36 | 0.29 |
| 3 | 0.25 | 0.45 | 0.45 |
| 4 | 0.09 | 0.06 | 0.10 |
| 5 | 0.16 | 0.12 | 0.19 |
| 6 | 0.18 | 0.11 | 0.24 |

F-Scores

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

| Dataset | RelNet | ARACNE | MRNet |
|---------|--------|--------|-------------|
| 1 | 0.29 | 0.37 | 0.38 |
| 2 | 0.31 | 0.38 | 0.39 |
| 3 | 0.32 | 0.49 | 0.52 |
| 4 | 0.07 | 0.08 | 0.13 |
| 5 | 0.13 | 0.14 | 0.15 |
| 6 | 0.13 | 0.15 | 0.20 |

Outline

Introduction

State of the Art

MRNet

Experiments

Results and Conclusion

Curves: DR3 (600,600)

Outline

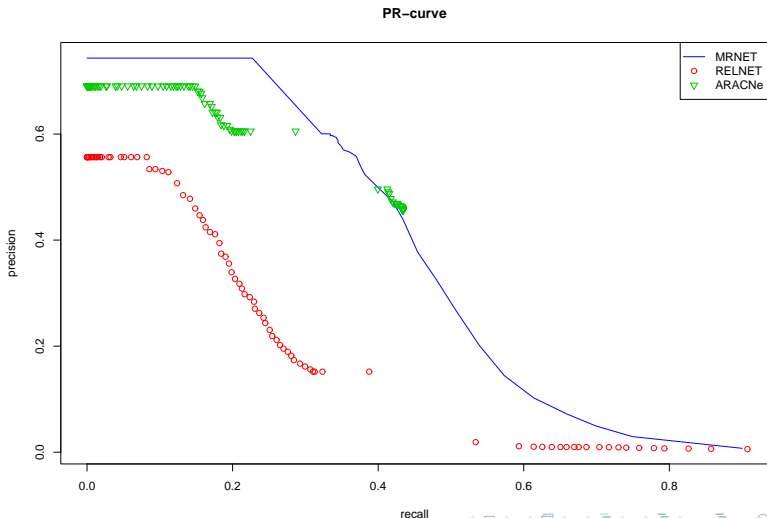
Introduction

State of the Art

MRNet

Experiments

Results and Conclusion



Conclusions and Future Works

Outline

Introduction

State of the Art

MRNet

Experiments

Results and Conclusion

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

Bibliography

Outline

Introduction

State of the Art

MRNet

Experiments

Results and Conclusion



Butte, A. J. and Kohane, I. S. (2000).

Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements. *Pacific Symposium on Biocomputing*, 5:415–426.



Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D., and Califano, A. (2006).

ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7.



Peng, H. and Long, F. (2004).

An efficient max-dependency algorithm for gene selection. In *36th Symposium on the Interface: Computational Biology and Bioinformatics*.