

Pot-luck challenge: FACT SHEET.

(for a task solved)

Title: Using mutual information to infer causal relationships

Participant name, address, email and website:

Catharina Olsen: colsen@ulb.ac.be
Patrick E. Meyer: pmeyer@ulb.ac.be
Gianluca Bontempi: gbonte@ulb.ac.be

Machine Learning Group, Computer Science Department
Université Libre de Bruxelles, Brussels, Belgium

Task solved: LOCANET

References

- C. Ding and H. Peng. Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology*, 3:185–205, 2005.
- P. E. Meyer, K. Kontos, F. Lafitte, and G. Bontempi. Information-theoretic inference of large transcriptional regulatory networks. *EURASIP Journal on Bioinformatics and Systems Biology*, 2007.
- P. E. Meyer, F. Lafitte, and G. Bontempi. An open source R/Bioconductor package for mutual information based network inference. *BMC Bioinformatics*, 2008.
- C. Olsen, P. E. Meyer, and G. Bontempi. On the impact of noise and missing values on transcriptional regulatory network inference based on mutual information. *EURASIP Journal on Bioinformatics and Systems Biology*, 2008.

Method:

The methods we applied to the given problems are based on the notion of mutual information. In a first step we applied the MRNET inference method to the datasets in order to obtain the base structure of the graph.

The MRNET method (Meyer et al. (2007)) is based on the maximum relevance/ minimum redundancy (MRMR) feature selection technique (Ding and Peng (2005)). This iterative selection technique chooses at each step, among the least redundant variables, the one having the highest mutual information with the target.

The method ranks the set of inputs according to a score which is the difference between the mutual information with the output variable Y (maximum relevance) and the average mutual information with the previously ranked variables (minimum redundancy). The network is inferred by deleting all edges whose score lies below a given threshold.

Direct interactions should be well ranked whereas indirect interactions should be badly ranked. In the first step, the variable X_i which has the highest mutual information to the

target Y is selected. The second selected variable X_j will be the one with a high information $I(X_j; Y)$ to the target and at the same time a low information $I(X_j; X_i)$ to the previously selected variable.

In the next steps, given a set \mathbf{X}_S of selected variables, the criterion updates \mathbf{X}_S by choosing the variable that maximizes the score

$$s_j = I(X_j; Y) - \frac{1}{|\mathbf{S}|} \sum_{X_k \in \mathbf{X}_S} I(X_j; X_k) \quad (1)$$

which can be described as a relevance term minus a redundancy term.

For each pair $\{X_i, X_j\}$ the algorithm returns two scores s_i and s_j and computes the maximum of the two. All edges with a score below a given threshold are then deleted. For further information see for example (Olsen et al. (2008))

The second step consisted in the edge orientation in the graph. In order to achieve this task we used the interaction information which is characterized by the mutual information of two variables X_i and X_j conditioned on a third one Y minus the mutual information of X_i and X_j .

$$I(X_i; X_j|Y) - I(X_i; X_j) \quad (2)$$

In contrast to mutual information this quantity also assumes negative values and leads to the following criteria for edge orientation:

$$\left. \begin{array}{l} X_i \leftrightarrow Y \leftrightarrow X_j \\ I(X_i; X_j|Y) > I(X_i; X_j) \end{array} \right\} \Rightarrow X_i \rightarrow Y \leftarrow X_j \quad (3)$$

$$\left. \begin{array}{l} X_i \rightarrow Y \leftrightarrow X_j \\ I(X_i; X_j|Y) \leq I(X_i; X_j) \end{array} \right\} \Rightarrow X_i \rightarrow Y \rightarrow X_j \quad (4)$$

Results: The summary of our results can be found in the following table

Dataset	Score
REGED	0.52
MARTI	0.21
CINA	3.31

Table 1: Result table.

- The main advantage of our method is its speed. It can deal with large networks (up to 20000 variables) in a reasonable amount of time
- The method is theoretically motivated by the inference of transcriptional networks from microarray data. The coupling of a simple information-theoretic arc orientation algorithm with MRNET is a novel approach to the causal detection problem.

The first part of our solution was carried out using the recently released R-package minet (Meyer et al. (2008)) which can be downloaded from:
<http://cran.r-project.org/web/packages/minet>

Keywords: interaction information, filter, MINET