Perspectives of feature selection in bioinformatics: from relevance to causal inference

Gianluca Bontempi

Machine Learning Group
Interuniversity Institute of Bioinformatics in Brussels (IB)^2
Université libre de Bruxelles, Bld de Triomphe
1050 Brussels, Belgium
mlg.ulb.ac.be

A major goal of the scientific activity is to model real phenomena by studying the dependency between entities, objects or more in general variables. Sometimes the goal of the modeling activity is simply predicting future behaviors. Sometimes the goal is to understand the causes of a phenomenon (e.g. a disease). Finding causes from data is particular challenging in bioinformatics where often the number of features (e.g. number of probes) is huge with respect to the number of samples [5]. In this context, even when experimental interventions are possible, performing thousands of experiments to discover causal relationships between thousands of variables is not practical. Dimensionality reduction techniques have been largely discussed and used in bioinformatics to deal with the curse of dimensionality. However, most of the time these techniques focus on improving prediction accuracy, neglecting causal aspects. This tutorial will introduction some basics of causal inference and will discuss some open issues: may feature selection techniques be useful also for causal feature selection? Is prediction accuracy compatible with causal discovery [2]? How to deal with Markov indistinguishable settings [1]? Recent results based on information theory [3], and some learned lessons from a recent Kaggle competition [4] will be used to illustrate the issue.

References

1. G. Bontempi and M. Flauder. From dependency to causality: A machine learning approach. *JMLR*, 15(16):2437–2457, 2015.
2. G. Bontempi, B. Haibe-Kains, C. Desmedt, C. Sotiriou, and J. Quackenbush. Multiple-input multiple-output causal strategies for gene selection. *BMC Bioinformatics*, 12(1):458, 2011.
3. G. Bontempi and P.E. Meyer. Causal filter selection in microarray data. In *Proceedings of ICML*, 2010.
4. G. Bontempi, C. Olsen, and M. Flauder. *D2C: Predicting Causal Direction from Dependency Features*, 2014. R package version 1.1.
5. P.E. Meyer and G. Bontempi. *Biological Knowledge Discovery Handbook*, chapter Information-theoretic gene selection in expression data. IEEE Computer Society, 2014.