Latent forests to represent high-dimensional and spatially correlated data and validation of the ability to detect causal genetic risk factors of diseases

Christine Sinoquet (Ecole Polythechnique de l'Université de Nantes)

November 13, 2013

Genome-wide association studies have revolutionized the search for genetic influences on common genetic diseases such as diabetes, obesity, asthma, cardio-vascular diseases and some cancers. In particular, together with the population aging concern, increasing health care costs require that further investigations are pursued to design scalable and efficient tools. The high dimensionality and complexity of genetic data hinder the detection of genetic associations. To decrease the risks of missing the causal factor and discovering spurious associations, machine learning offers an attractive framework alternative to classical statistical approaches. We have proposed a novel class of probabilistic graphical models (PGMs) - the forest of latent tree models (FLTMs) - , to reach a trade-off between faithful modeling of data dependences and tractability. In the work presented here, we assess the ability of the FLTM model to detect genotype-disease associations. This talk will first put our contribution into the perspective of PGM-based works meant to model the dependences in genetic data; second, our contribution will be considered from the technical viewpoint of LTM learning, with the vital objective of scalability in mind. Thirdly, we will present the systematic and comprehensive evaluation conducted to assess the ability of the FLTM model to detect genetic associations through latent variables. Realistic simulations were performed under various controlled conditions. We will also show and discuss results obtained on real data. Beside guaranteeing data dimension reduction through latent variables, the FLTM model is empirically proven able to capture indirect genetic associations with the disease: strong associations are evidenced between the disease and the ancestor nodes of the causal genetic marker node, in the forest; in contrast, very weak associations are obtained for other latent variables. Finally, we will discuss the prospects of the model for association detection at genome scale.

Keywords: Probabilistic Graphical Model, Bayesian Network, Latent TreeModel, Detection of Genetic Association, Latent Variable, Data Dimension Reduction, Scalability.

Webpage: `http://pagesperso.lina.univ-nantes.fr/info/perso/permanents/sinoquet/`