Bayesian sparse neural networks for biomarker discovery with omics data

Recent advances in high-throughput biotechnologies, such as microarray and sequencing technologies, have provided an unprecedented opportunity for biomarker discovery, which potentially enhances the development of precision medicine. From a statistical point of view, the discovery of biomarkers can be best cast as variable selection, where the variable refers to the molecular attributes under investigation, e.g., genes or SNPs. Variable selection for omics data is challenging. First, the omics data is high-dimensional, whose sample size is typically much smaller than the number of variables (a.k.a. small-n-large-p). Second, certain processes of the cell are interconnected in complex patterns due to cellular regulations or environmental factors influencing the cell. This results in an unknown, complex, nonlinear relationship among variables. However, most of the current variable selection methods are developed based on linear or generalized linear models, rendering imprecise discovery of useful biomarkers.

Motivated by the universal approximation ability of feed-forward neural networks, we propose a Bayesian sparse neural network method to tackle this problem, where the nonlinearity of the system is modeled via repeated compositions of nonlinear mappings. In the proposed method, a spike-and-slab prior is imposed on each connection of the neural network to induce its sparsity, a large number of sparse neural networks with different structures are sampled from its posterior distribution using a parallel Markov chain Monte Carlo (MCMC) algorithm, and relevant variables are selected according to their marginal inclusion probabilities, i.e., how often they appear in the sampled sparse neural networks. Traditionally, feed-forward neural networks are viewed as a black-box approximator to an objective function. However, our work indicates that this view is not completely correct for Bayesian sparse neural networks: The structures of the sparse neural networks sampled from the posterior distribution indicate the relevance of the selected covariates to output variables. We have provided a theoretical justification for variable selection consistency of the proposed method for high-dimensional nonlinear systems, where the number of variables is allowed to be much larger than the training sample size. In general, the proposed method can be employed to learn a sparse neural network for the problems with a small number of training samples.

We have applied the proposed method to the cancer cell line encyclopedia (CCLE) dataset for identification of the genes sensitive to chemical compounds, which is fundamental to elucidate the response mechanism for anticancer drugs towards precision medicine. The CCLE dataset consisted of the dose-response data for 24 chemical compounds across over 400 cell lines. For each cell line, it consisted of the expression data of 18,926 genes, which underwent a variable screening procedure before being analyzed by the proposed method. For many chemical compounds, the genes selected by the proposed method are highly consistent with our existing knowledge. For example, for the compounds Topotecan, Irinotecan, 17-AAG and Paclitaxel, the proposed method selected SLFN11, SLFN11, NQO1 and BCL2L1 as their respective top drug-

sensitive genes. These results have been validated by the existing literature that the selective genes are predictive of treatment response for the respective chemical compounds. In addition, for the compounds AEW541, Erlotinib, and Nilotinib, the proposed method selected their target genes as the top drug-sensitive genes.

The proposed method can be extended in various ways. For example, it can work with deep neural networks, it can work with different prior distributions such as the mixture Gaussian prior, and its computation can be accelerated by the stochastic gradient MCMC algorithms and many other Monte Carlo algorithms with the use of the strategy of mini-batch data.

Faming Liang
Department of Statistics, Purdue University, West Lafayette, IN 47906, USA

Publication

Bayesian Neural Networks for Selection of Drug Sensitive Genes
Liang F, Li Q, Zhou L
J Am Stat Assoc. 2018/i>