

Abstract

Learning the structures of large undirected graphical models from data is an active research area and has many potential applications in various domains, including molecular biology, social science, marketing data analysis, among others. The estimated structures provide semantic clarity, the possibility of causal interpretation, and ease of integration with a variety of tools. For example, one very important direction in system biology is to discover gene regulatory networks from microarray data (together with other data sources) based on the observed mRNA levels of thousands of genes under various conditions.

Structure learning for undirected graphs is an open challenge in machine learning. Most probabilistic structure learning approaches enforce sparsity on the estimated structure by penalizing the number of edges in the graph, which leads to a non-convex optimization problem. Thus these approaches have to search for locally optimal solutions through the combinatorial space of structures, which makes them unscalable for large graphs. Furthermore, the local optimal solution they find could be far away from the global optimal solution, especially when the number of configuration instances is small compared with the number of nodes in the graph.

This thesis tries to address these issues by developing a novel structure learning approach that can learn large undirected graphs efficiently in a probabilistic framework. We use the Graphical Gaussian Model (GGM) as the underlying model and propose a novel ARD style Wishart prior for the precision matrix of the GGM, which encodes the graph structure we want to learn. With this prior, we can get the MAP estimation of the precision matrix by solving a modified version of Lasso regression and thus achieve a sparse solution. By proposing a generalized version of Lasso regression, which is called the Feature Vector Machine (FVM), our structure learning model is further extended so that it can capture non-linear dependencies between node variables. In particular, the optimization problem in our model remains convex even in non-linear cases, which makes our solution globally optimal. We have also developed a graph-based classification approach for predicting node labels given network structures, either observed or automatically induced. This approach is especially suitable when edges in the networks contain multiple input features.

The contributions of this thesis work can be seen from several aspects. First, it provides a probabilistic framework that allows us to learn optimal undirected graph structures with a low polynomial (quadratic when the graph is sparse) computational cost. Second, the development of Feature Vector Machine, which is both theoretically and practically meaningful, enriches current approaches to feature selection and extends our structure learning model so that the non-linear dependencies among node variables can be captured. Third, a graph-based classification approach is developed for predicting node labels using the observed or learned network structures. Fourth, we provided empirical evidence for the proposed methods in gene regulatory network re-construction and gene function prediction, as well as multi-class text categorization tasks.