

“Inference of Patterns and Associations Using Dictionary Models”

Jun Liu

Department of Statistics

Harvard University

Pattern discovery is a ubiquitous problem in many disciplines. It is especially prominent in recent years due to our greatly improved data-generation capabilities in science and technologies. The method I present here is motivated by the "motif-finding" and "module-finding" problems in biology, i.e., to find sequence patterns (i.e., "words") that seem to appear more frequent than usual in a given set of text sequences (i.e., sentences) and to find which of these "words" tend to co-occur in a sentence. A challenge in the motif-finding problem is that there are no spacing's and punctuations between the words and the dictionary of "words" is unknown to us. Existing methods are mostly "bottom-up" approaches, i.e., to build up the dictionary starting with single-letter words and then concatenate some existing words that appear to occur next to each other in sentences more frequently than chance. Our new approach is a top-down strategy, which uses a tree structure to represent the relationship among all possible existing words and uses the EM algorithm to estimate the usage frequency of each word. It automatically trims down most of the incorrect "words" by letting their usage frequencies converge to zero.

The module-finding problem is closely related to the well-known "market basket" problem, in which one attempts to mine association rules among the items in a supermarket based on customers' transaction records. It is also related to the two-way clustering problem. In this problem, we assume that the words are given, and our goal is to find subsets of words that tend to co-occur in a sentence.

We call the set of co-occurring words (not necessarily orderly) a "theme" or a "module". We can generalize the dictionary model to the "theme"-model and use a similar EM-strategy to infer these themes. I will demonstrate its applications in a few examples including an analysis of Chinese medicine prescriptions and an analysis of a Chinese novel.

This is based on a joint work with Ke Deng and Zhi Geng.