Label Space Dimension Reduction for Multi-label Classification

林軒田
臺灣大學資訊工程學系

Abstract

Multiclass classification is an important problem in machine learning. It can be used in a variety of applications, such as organizing documents to different categories automatically. Multi-label classification is an extension of multi-class classification --- the former allows a set of labels to be associated with an instance while the latter allows only one. For instance, a document may belong to both the "politics" and "health" class if it is about the National Health Insurance. Many other similar applications arise in domains like text mining, vision, or bio-informatics.

In this talk, we discuss a coding view about the output (label) space of multi-label classification. The view represents each set of possible labels as a (fixed-length) binary string. We discuss the close connection between the binary-string representation and the coding theory. In particular, we demonstrate two novel research directions based on the connection: data compression (source coding) and learnable data compression (conditional source coding). The directions lead to two algorithms that systematically compresses the label space for more efficient computation.