Active learning: mild assumptions, tractability and cost-sensitive learning

Boris Ndjia Njike

Abstract

The paradigm of supervised learning consists in learning an arbitrary mapping $f : \mathcal{X} \longrightarrow \mathcal{Y}$ based on a labelled data set of (X, Y) pairs in which $X \in \mathcal{X}$ represents an unlabelled data, and $Y \in \mathcal{Y}$ its corresponding label. In the standard approach of supervised learning called *passive learning*, a starting point involves gathering a substantial amount of labelled data that have been randomly selected from the underlying population distribution. Afterward, based on these data, we proceed to develop a procedure to approximate the mapping f. However, in many applications, the volume of unlabelled data X is very huge, and then the gathering process can be quite tiresome and time-consuming as each unlabelled data has to be manually labelled. This constrains us to look beyond the passive learning approach. In this context, one of the most studied techniques is the iterative supervised learning called *active learning* that aims at reducing the data labelling effort. In active learning, the algorithm (sometimes called "learner") sequentially selects (according to some criteria) unlabelled data from a source of data, and decide to request (to a so-called "oracle") the value of the corresponding label one at time. In such scheme, the number of labelled data required to learn the mapping f can be sometimes lower than the one obtained in standard passive learning.

This thesis presents various procedures in active learning that show the benefits over the passive learning counterpart in terms of the number of labelled data needed to learn the mapping f. These procedures have strong theoretical guarantees and are mostly computationally tractable on a broad spectrum of real-world data, as illustrated by numerical simulations.