Shortcuts for Unbalanced Classification

by
Mu Zhu
University of Waterloo
200 University Ave. W.
Waterloo, Ontario, Canada N2L 3G1
mzhu@post.harvard.edu

Abstract

We study the "unbalanced classification" or "rare target detection" problem. The training set consists of two classes. The class of interest is rare; most observations belong to a majority, background class. Given a set of unlabeled observations, the goal is to rank those belonging to the rare class ahead of the rest. Due to the "unbalanced" or "rare" nature of the problem, there is often a rather limited amount of useful information in the training set. This means we can — and must — take some shortcuts. I will present two such shortcuts, one being an efficient kernel method that can be viewed as a support vector machine constructed without using any iterative optimization procedure, and the other being a two-stage strategy for dealing with situations where the decision boundary is changing over time. I will describe a number of successful applications in drug discovery, web search, and border management.