2014 – 01

Active learning procedure via sequential experimental design and uncertainty sampling

Jing Wang, Eunsik Park, Yuan-chin Ivan Chang

Jun. 19, 2014

Active learning procedure via sequential experimental design and uncertainty sampling

Jing Wang[a], Eunsik Park[a], Yuan-chin Ivan Chang[b]

[a]*Department of Statistics, Chonnam National University, Gwangju 500-757, Korea*
[b]*Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan*

Abstract

Classification is an important task in many fields including biomedical research and machine learning. Traditionally, a classification rule is constructed based on a bunch of labeled data. Recently, due to technological innovation and automatic data collection schemes, we easily encounter with data sets containing large amounts of unlabeled samples. Because to label each of them is usually costly and inefficient, how to utilize these unlabeled data in a classifier construction process becomes an important problem. In machine learning literature, active learning or semi-supervised learning are popular concepts discussed under this situation, where classification algorithms recruit new unlabeled subjects sequentially based on the information learned from previous stages of its learning process, and these new subjects are then labeled and included as new training samples. From a statistical aspect, these methods can be recognized as a hybrid of the sequential design and stochastic approximation procedure. In this paper, we study sequential learning procedures for building efficient and effective classifiers, where only the selected subjects are labeled and included in its learning stage. The proposed algorithm combines the ideas of Bayesian sequential optimal design and uncertainty sampling. Computational issues of the algorithm are discussed. Numerical results using both synthesized data and real examples are reported.

Keywords: Active learning, Uncertainty sampling, Sequential experimental design, D-optimal design, Bayes rule