Active Learning Procedure via Sequential Experimental Design and Uncertainty Sampling

Eunsik Park*
Chonnam National University, Gwangju, Republic of Korea - espark02@gmail.com

Jing Wang
Chonnam National University, Gwangju, Republic of Korea

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 data sets containing large amounts of unlabeled samples. Because labeling each of them is usually costly and inefficient, the way 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; 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: classification rule; stochastic approximation; D-optimal design; Bayesian estimation.