UCLA C.S. Dept. Tech. Report TR060027 December 2006

Load Shedding in Classifying Multi-Source Streaming Data: A Bayes Risk Approach

Yijian Bai
UCLA
bai@cs.ucla.edu

Haixun Wang
IBM T. J. Watson
haixun@us.ibm.com

Carlo Zaniolo
UCLA
zaniolo@cs.ucla.edu

Abstract

In many applications, we monitor data obtained from multiple streaming sources for collective decision making. The task presents several challenges. First, data in sensor networks, satellite transmissions, and many other fields are often of large volume, fast speed, and highly bursty nature. Second, because data are collected from multiple sources, it is impossible to offload classification decisions to individual data sources. Hence, the central classifier responsible for decision making is constantly under overloaded situations. In this paper, we study intelligent load shedding for classifying multi-source data. We aim at maximizing classification quality under resource (CPU and bandwidth) constraints. We use a Markov model to predict the distribution of feature values over time. Then, leveraging Bayesian decision theory, we use Bayes risk analysis to model the variances among different data sources in their contributions to classification quality. We adopt an Expected Observational Risk criterion to quantify the loss of classification quality due to load shedding, and propose a Best Feature First (BFF) algorithm that greedily minimizes such a risk. We also introduce an approximate BFF algorithm that reduces computation complexity. The effectiveness of the approach proposed is confirmed by several experiments on both synthetic and real-life data.