Big Data Management and Analytics

Group: Bo Dong, Qiushi Zhang, Yifan Li

Project Description:

Consider a set of independent non-stationary processes (denoted as P), each generating a continuous stream of data instances from a domain D. A data instance is denoted by (x, y), where x ∈ Dy , is a vector of v covariates, and y ∈ Y = {1 . . . k}, is its corresponding value. Here, k is the index of true values. In this setting, assume that P consists of two processes, source and target, denoted as S and T . Throughout the paper, we use the symbol denoting a process interchangeably with its corresponding data stream. In process (or stream) S, both x and y of each data instance are observed. On the contrary, only x of each data instance in stream T is observed. This indicates that S is completely labeled while T is unlabeled. We define the MultiStream Regression problem as follows.

Let XS ∈ D be a set of v-dimensional vectors of covariates and YS be the corresponding values observed on a non-stationary stream S. Similarly, let XT ∈ D be a set of v-dimensional vectors of covariates observed on an independent non-stationary stream T . Construct a regression set M that predicts value of x ∈ XT using XS, YS and XT.

Here since (x, y) ∈ S are used to predict labels of x ∈ T , we call S as the source stream and T as the target stream.