Reframing for Non-Linear Dataset Shift

Md Shadman Rafid, Mohammad Mazedul Islam,
Md Naimul Hoque, Chowdhury Farhan Ahmed

Department of CSE, University of Dhaka, Bangladesh
shadmanrafiddeep@gmail.com
mazidmailbox@gmail.com
naimul.et@easternuni.edu.bd
farhan@du.ac.bd

Abstract. Discriminative classification models assume that both training and deployment data have same distributions of data attributes. These models give significantly varied performances when they are deployed under varied circumstances with different data distributions. This phenomenon is called Dataset Shift. In this paper we have provided a method which first determines whether there is a significant shift in the distributions of attributes between the training and deployment datasets. If there exists a shift in the data the proposed method then uses a Hill climbing approach to map this shift irrespective of its nature i.e. (linear or non-linear) to the equation for quadratic transformation. Experimental results on three real life datasets show strong performance gains achieved by the proposed method over previously established methods such as retraining and linear reframing.

1 Introduction

The main concerns of supervised machine learning is to learn a model for classification, regression or any other function using a set of training data and then applying this new learned model to deployment data. While deploying a particular data model it is implicitly assumed that both the training and test data will follow the same distributions. But in real life scenarios it is natural for the distributions of data attributes and decision functions to change especially when the training data is collected in one context while the deployment data is used in a different context e.g.(the training data is collected the summer season while the model is deployed on the data for the season of autumn). Such "Dataset Shift" [Han et al. (2012)] if not compensated for can greatly reduce the efficiency of the results provided by the learned model. One solution is to retrain the entire model on the deployment data. But it is not a feasible option often as collection and labelling of data in deployment may become costly. Another recent approach is reframing the data so as to make the learned model compensate for the shift in deployment data in real time and provide efficient results. We mainly focus on the shifts of continuous attributes of the data from test to deployment dataset. For example the daily food consumption of the residents of a city in North America may vary greatly from that of the residents of a similar city in South America.