Dataset Shift in a Real-Life Dataset

Chowdhury Farhan Ahmed, Nicolas Lachiche, Clement Charnay, Agnes Braud

University of Strasbourg, France

http://www.reframe-d2k.org/

Generalization and reuse of machine learning models over multiple contexts
(LMCE 2014)

Chowdhury Farhan Ahmed
University of Strasbourg
France

Outline

- Introduction and Motivation
- Related Work
- Description of the Bike Sharing Dataset
- Occurrences of Dataset Shift
- The Split of Kaggle
- Bike Sharing for Dataset Shift and More…

Introduction and Motivation

- Dataset shift (Moreno-Torres et al. 2012) refers to the problem where training and testing datasets follow different distributions.
- Although it is natural to observe dataset shift in real-life datasets, unfortunately, existence of clear dataset shift is rarely found in the publicly available real-life datasets. The existing methods used either synthetic or non-publicly available real-life datasets.
- Here we present the existence of remarkable dataset shift in a publicly available dataset called Bike Sharing (Fanaee-T et al. 2014).
- We experimentally analyze how to split the dataset to achieve dataset shift in both input and output variables.
- Future research directions are discussed where this dataset can effectively be used as a real-life benchmark.

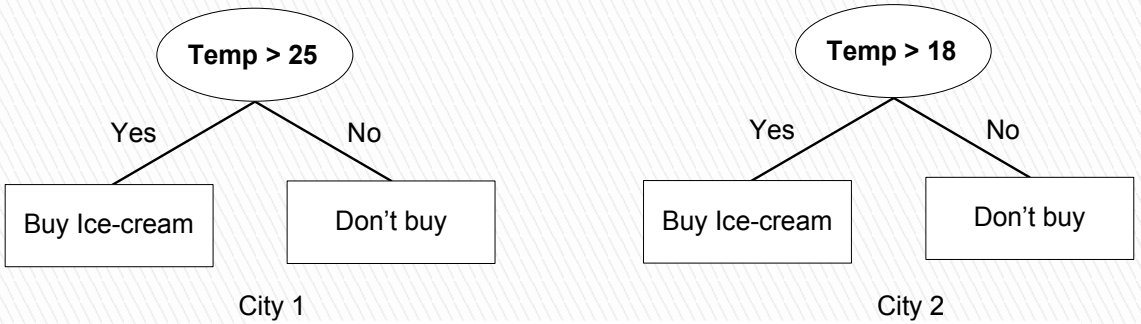


Fig 1: An example of dataset shift.

Related Work

- Some methods have been proposed to perform adjustment on the model output when to be applied over different contexts, such as
  - Tuning multi-class classification problem (Charnay et al. 2013).
  - Cost-sensitive regression model (Zhao et al. 2011).
  - ROC curve for regression (Hernandez-Orallo et al. 2013).
- Input variable shift is most often known as covariate shift in machine learning. Research has been done to tackle covariate shift such as
  - Importance Weighted Cross Validation (IWCV) (Sugiyama et al. 2007).
  - Integrated Optimization Problem (Bickel et al. 2009).
  - Kernel Mean Matching (Gretton et al. 2009).
- An input transformation based method, called GP-RFD (Moreno-Torres et al. 2013) (Genetic Programming-based feature extraction method for the Repairing of Fractures between Data) has been proposed for handling dataset shift.

Description of the Bike Sharing Dataset

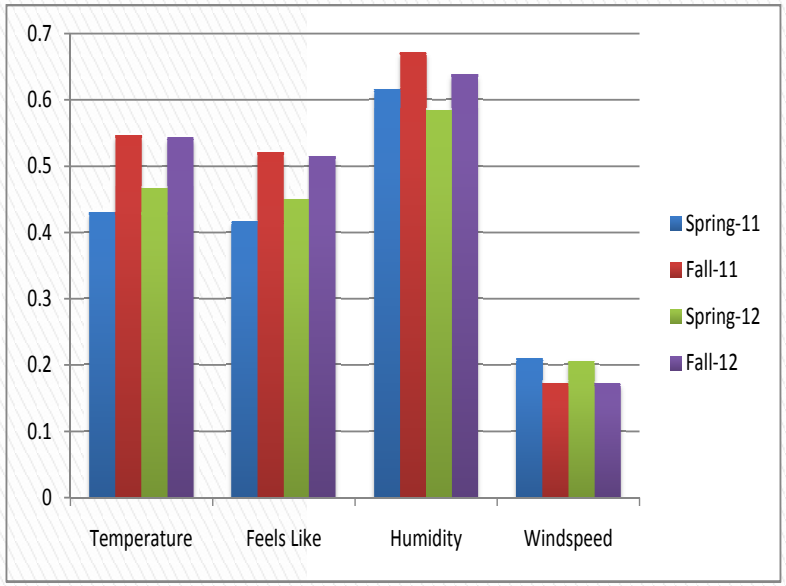
- The Bike Sharing Dataset (Fanaee-T et al. 2014) contains usage logs of a bike sharing system called Capital Bike Sharing (CBS) at Washington, D.C., USA for two years (2011and 2012).
- It is publicly available in UCI Machine Learning Repository.
- It contains bike rental counts in both hourly (17,379 records) and dialy (731records) formats based on environmental and seasonal settings.
- The input variables contain day, hour, season, workday/ holiday and some weather information such as temperature, feels like temperature, humidity and wind speed.
- The original objective of the creators of this dataset was event and anomaly detection.

Occurrences of Dataset Shift

- In real-life weather changes according to the change of seasons. Moreover, the renting behaviour of people may also change according to time.
- We have splitted this dataset into four parts according to four months of sequential time and labelled as Spring-11, Fall-11, Spring-12 and Fall-12 (Spring: January to June, Fall:July-December)
- We have taken four most influential input attributes of this dataset representing weather information. The values of this attributes have been normalized as follows
  - Temperature: The values (Celsius) are normalized by dividing by 41 (max).
  - Feels Like Temperature: The values (Celsius) are normalized by dividing by 50 (max).
  - Humidity: The values are normalized by dividing by 100 (max).
  - Windspeed: The values are normalized by dividing by 67 (max).
- These splits contain remarkable dataset shift. Other possible splits for observing dataset shifts are with respect to months, seasons and years.

Occurrences of Dataset Shift (Contd..)

(a)



(b)

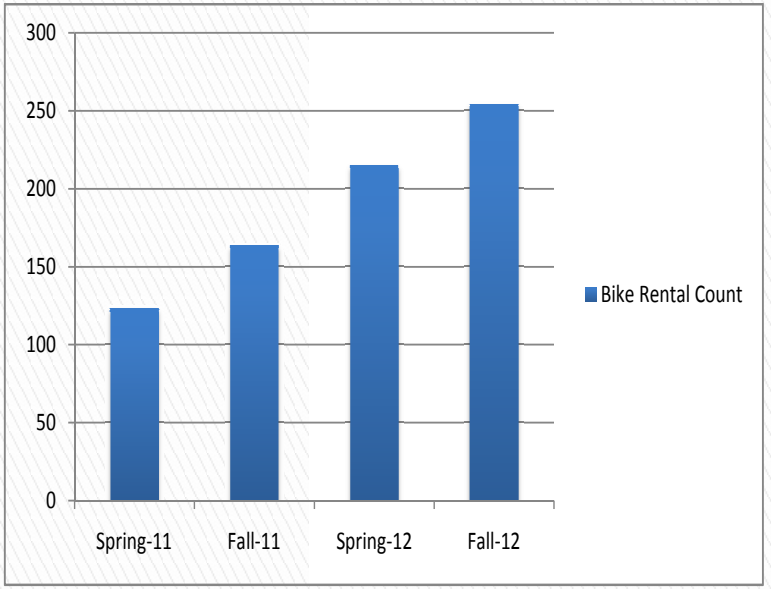


Fig 2: Distribution of the average values of (a) input and (b) output attributes in the semester splits.

Table 1: Performance of the base model Spring-11 in other semesters

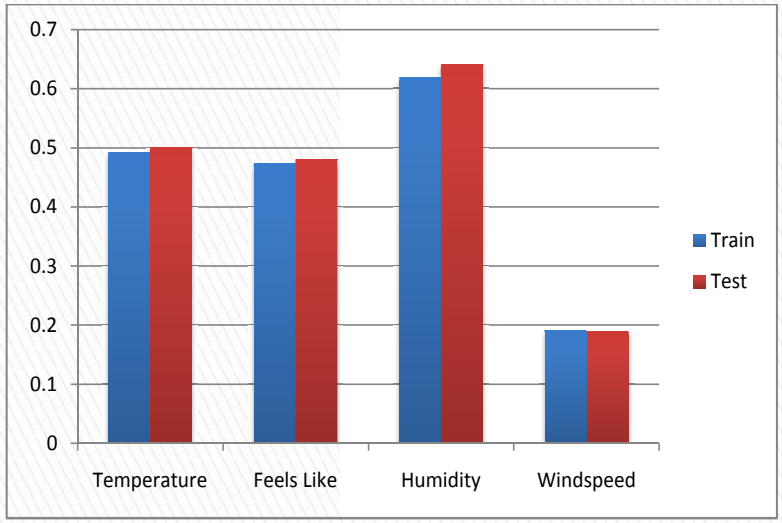
Source	Measure	Deployment			
		Spring-11	Fall-11	Spring-12	Fall-12
Spring-11	MAE	71.789	102.293	132.226	158.884
	RMSE	99.058	135.427	186.459	224.307

The Split of Kaggle

- Recently, Kaggle has provided a problem on Bike Sharing Dataset.
- The original dataset has been divided into two parts called Train and Test.
- The Train part contains data of each month from day 1 to 19, and the Test part contains data from day 20 to the end of a month.
- The problem is to build a regression model with the Train dataset and predict the bike rental counts in the Test dataset.
- Here, dataset shift is absent because of mixing data of every month in the Train and Test datasets.

The Split of Kaggle (Contd..)

(a)



(b)

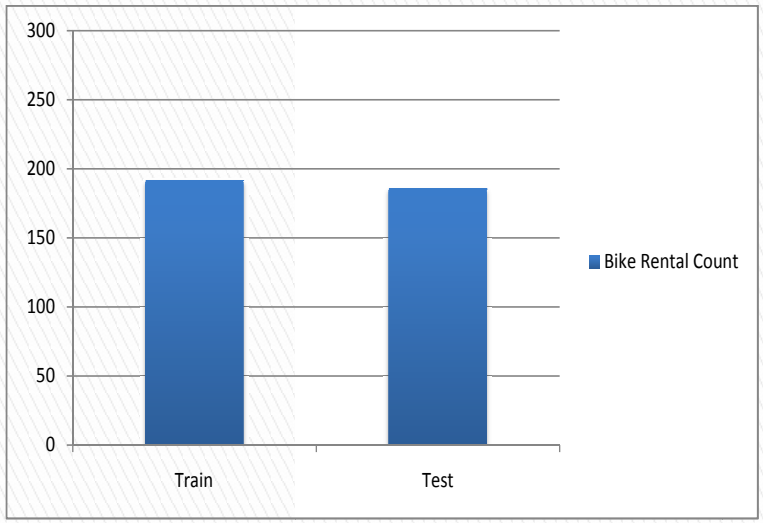
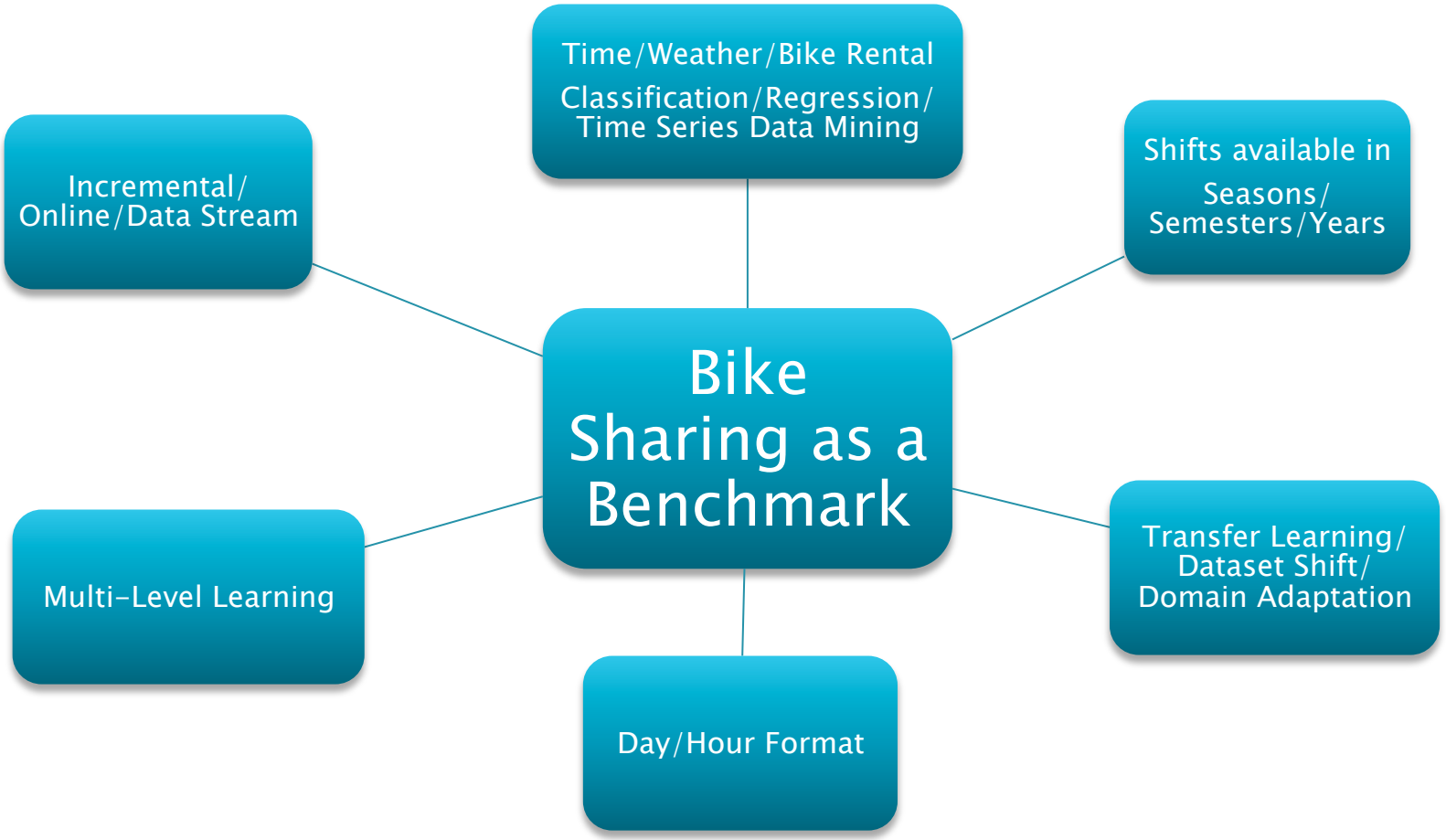


Fig 3: Distribution of the average values of (a) input and (b) output attributes in the Kaggle split.

Table 2: Performance of the base model Train in Test for the Kaggle split

Source	Measure	Deployment	
		Train	Test
Train	MAE	117.595	117.17
	RMSE	158.607	157.517

Bike Sharing for Dataset Shift and More .....



REFRAME

Thank You