==========================================

Bike Sharing Dataset

==========================================

Hadi Fanaee-T

Laboratory of Artificial Intelligence and Decision Support (LIAAD), University of Porto

INESC Porto, Campus da FEUP

Rua Dr. Roberto Frias, 378

4200 - 465 Porto, Portugal

=========================================

Background

=========================================

Bike sharing systems are new generation of traditional bike rentals where whole process from membership, rental and return

back has become automatic. Through these systems, user is able to easily rent a bike from a particular position and return

back at another position. Currently, there are about over 500 bike-sharing programs around the world which is composed of

over 500 thousands bicycles. Today, there exists great interest in these systems due to their important role in traffic,

environmental and health issues.

Apart from interesting real world applications of bike sharing systems, the characteristics of data being generated by

these systems make them attractive for the research. Opposed to other transport services such as bus or subway, the duration

of travel, departure and arrival position is explicitly recorded in these systems. This feature turns bike sharing system into

a virtual sensor network that can be used for sensing mobility in the city. Hence, it is expected that most of important

events in the city could be detected via monitoring these data.

=========================================

Data Set

=========================================

Bike-sharing rental process is highly correlated to the environmental and seasonal settings. For instance, weather conditions,

precipitation, day of week, season, hour of the day, etc. can affect the rental behaviors. The core data set is related to

the two-year historical log corresponding to years 2011 and 2012 from Capital Bikeshare system, Washington D.C., USA which is

publicly available in http://capitalbikeshare.com/system-data. We aggregated the data on two hourly and daily basis and then

extracted and added the corresponding weather and seasonal information. Weather information are extracted from http://www.freemeteo.com.

=========================================

Associated tasks

=========================================

- Regression:

Predication of bike rental count hourly or daily based on the environmental and seasonal settings.

- Event and Anomaly Detection:

Count of rented bikes are also correlated to some events in the town which easily are traceable via search engines.

For instance, query like "2012-10-30 washington d.c." in Google returns related results to Hurricane Sandy. Some of the important events are

identified in [1]. Therefore the data can be used for validation of anomaly or event detection algorithms as well.

=========================================

Files

=========================================

- Readme.txt

- hour.csv : bike sharing counts aggregated on hourly basis. Records: 17379 hours

- day.csv - bike sharing counts aggregated on daily basis. Records: 731 days

=========================================

Dataset characteristics

=========================================

Both hour.csv and day.csv have the following fields, except hr which is not available in day.csv

- instant: record index

- dteday : date

- season : season (1:springer, 2:summer, 3:fall, 4:winter)

- yr : year (0: 2011, 1:2012)

- mnth : month ( 1 to 12)

- hr : hour (0 to 23)

- holiday : weather day is holiday or not (extracted from http://dchr.dc.gov/page/holiday-schedule)

- weekday : day of the week

- workingday : if day is neither weekend nor holiday is 1, otherwise is 0.

+ weathersit :

- 1: Clear, Few clouds, Partly cloudy, Partly cloudy

- 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

- 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

- 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

- temp : Normalized temperature in Celsius. The values are divided to 41 (max)

- atemp: Normalized feeling temperature in Celsius. The values are divided to 50 (max)

- hum: Normalized humidity. The values are divided to 100 (max)

- windspeed: Normalized wind speed. The values are divided to 67 (max)

- casual: count of casual users

- registered: count of registered users

- cnt: count of total rental bikes including both casual and registered

=========================================

License

=========================================

Use of this dataset in publications must be cited to the following publication:

[1] Fanaee-T, Hadi, and Gama, Joao, "Event labeling combining ensemble detectors and background knowledge", Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg, doi:10.1007/s13748-013-0040-3.

@article{

year={2013},

issn={2192-6352},

journal={Progress in Artificial Intelligence},

doi={10.1007/s13748-013-0040-3},

title={Event labeling combining ensemble detectors and background knowledge},

url={http://dx.doi.org/10.1007/s13748-013-0040-3},

publisher={Springer Berlin Heidelberg},

keywords={Event labeling; Event detection; Ensemble learning; Background knowledge},

author={Fanaee-T, Hadi and Gama, Joao},

pages={1-15}

}

=========================================

Contact

=========================================

For further information about this dataset please contact Hadi Fanaee-T (hadi.fanaee@fe.up.pt)