2017 Fall Data Mining

Homework 2

Clustering

Due Date: 23:59, December 1st, Friday, 2017
TA: 蔡子豪 stu8978@gmail.com

In this homework, the major task is turn data into different clusters,and explain what the cluster means.

There is a Hackmd for us to discuss this homework.(except answers)

Dataset:

We‘ll use “201707-citibike-tripdata.csv.zip” (after preprocessed in HW0)

Schema:

The preprocessed data should contains the following

Every station’s information
id, name, lat, lng
Every stations’ flow data
id, time, in-flow, out-flow

Tasks:

You are ask to to use listed algorithms to do the following tasks, for each parameter you should try at least 2 values.

Spatial clustering:

Using stations’ geo-information to do clustering,try the following algorithm and try different parameters and explain the results.

- Kmeans(k=?)
    - elbow method
- DBscan(eps=?,min_sample=?,metric=?)

Temporal clustering:

Using the in-flow and out-flow data in the first week(7 days * 48 segment * 2=672 data points for each sensor) try the following methods.
(or you can use any period of time as a unit.)

- Agglomerative Clustering(affinity =?)
- PCA(n_components=?) => Agglomerative Clustering(affinity =?)

Other

Try to combine spatial and temporal information to do clustering,which is more important? How about giving them different weight and see the result.
(You can try any clustering algorithms you like.)

Report:

In this homework, you should try the previous methods and make some **observation**, **compare** different method and parameters, explain the result and see if the output meet your expect.

Your code should be submitted with the report

If you have any questions or suggestions,fell free to contact me:)