MIT Dataset

資料格式介紹

data/friendship_matrix.txt

第一個row和column代表subject的標號,譬如說(row,colum):(7,4)=1 代表7號把4號當成好朋友,好友關係是有方向性的,(4,7)=NaN代表4並沒有把7當好友,另外要注意有(7,7)=1這樣的關係,請忽略,這是原始dataset的問題

data/processed.friendship_matrix.txt

格式同data/friendship_matrix.txt。差別在於拿掉了self-loop，以及把原本檔案轉換為undirected graph。

data/location_record_sna.dat

每筆資料是 (device_id,日期,地點id),代表該device_id在某年某月某天某時某分某秒在哪個地點,地點以地點id表示

data/calling_record_sna_new_v2.dat

每筆資料是 (device_id,日期,Type of communication,direction,duration,calling_device_id),

(1)Type of communication有兩種分別:分別是Short message(簡訊),Voice call(語音)和Packet Data(封包),Packet Data作者並沒有特別提是什麼,
可能可以當一般的資料傳輸
(2)direction: incoming(接到電話) or outgoing(打電話過去)
(3)duration:通話時間,如果0通常代表電話沒有接
(4)calling_device_id:通話對象是誰

** 12/28更新: 移除noise data
** 12/30更新: 移除四行noise data

data/encounter_record_sna.dat

每筆資料是 (device_id,encounter_device_id,開始偵測到對方在範圍內的時間,最後一次偵測到對方在範圍內的時間)
(3,4,24-Nov-2004 12:35:33,24-Nov-2004 12:38:02)這筆資料可以代表12:35:33~12:38:02,subject 3和4都是待在一起的(5-10公尺)

Note:有時會出現friendship_matrix裡面沒見過的device_id,這是正常的,因為不是全部的device_id我們都知道他們的交友關係,所以我們只要predict friendship_matrix.txt中有個device_id就好

train_and_test/mit.$id.{train/test}.txt

“$v1\t$v2\t $is_connected”，中間是用\t隔開。
$v1、$v2是pair的node id。為了簡化問題，不考慮方向性。
$is_connected代表$v1、$v2是否相連。如果相連，則label為1；相反，則label為0。

MIT資料切分training/testing方式

我們把原本的friendship network，依照下列方式處理，切成10份train/testing。

1. 首先，我們做了以下的前處理：
    (a)把 friendship network轉成**undirected graph**。
    (b)**移除self-loop**的edge。
    (c)移除點“107”，因為點”107”沒有資料。

    處理過的network檔請見”processed.friendship_matrix.txt”。

2. 把graph的nodes切分成**10**等份，每一次取其中的一等份來產生testing data。以下暫稱這一等份為TestNS(testing nodes set) ，剩下的九等份稱為TrainNS(training node set)。

Training Data：

將TrainNS的node不考慮方向性地兩兩做pair (v1,v2)，產生一筆資料。如果 v1 和v2相連，則label為1；若不相連，則label為0。

Testing Data：

同樣，先將TestNS的node不考慮方向性地兩兩做pair (v1,v2)，產生一筆資料。如果 v1 和v2相連，則label為1；若不相連，則label為0。

接著，將TrainNS和TestNS的點兩兩做pair：先從TrainNS取出一個點v1，再從TestNS取出一個點v2，產生一筆資料。如果 v1 和v2相連，則label為1；若不相連，則label為0。

重複上述步驟十次，就得到10份對應的Training Data和Testing Data。

使用方法

1. 每一次請使用編號相同的Training Data和Testing Data。請以Training Data，搭配其他資料，去預測Testing Data每一個pair是否相連。

2. 使用的evaluation metrics： Precision、Recall、F1-score

3. 把每一次evaluation metrics的結果平均，以求得到平均表現的分數。

參考資料

Precision、Recall、F1-score
http://en.wikipedia.org/wiki/Precision_and_recall