Estudi i Classificació de Modes de Transport en Traces de Mobilitat Humana

Jordi Xhafa Daci

Resum– L'estudi de la mobilitat humana ha sigut sempre d'interès per les ciutats i models de transports. No obstant, les noves tecnologies, i molt especialment, els sensors i telèfons mòbils, han obert moltes oportunitats per estudiar la mobilitat humana, ja que permeten recollir grans quantitats de dades de mobilitat, de forma massiva, i pràcticament sense cost. En els darrers anys s'han presentat varis estudis de mobilitat, que processen i analitzen dades de mobilitat en diferents contexts. En la mateixa línia, en aquest treball es presenten els resultats d'un estudi comparatiu de diversos mètodes d'aprenentatge computacional per classificar els modes de transport empleats per les persones. El conjunt de dades és el *GeoLife* –de referència internacional– que conté traces de mobilitat de 182 persones en la ciutat de Pequín, Xina. Amb aquest estudi, s'han pogut entendre els conceptes i models de mobilitat humana així com l'aplicació dels mètodes de classificació adients.

Paraules clau– Mobilitat Humana, Modes de Transport, GeoLife Dataset, Traces de Mobilitat, GPS, Classificació, Random Forest, Logistic Regression, SVM, Xarxes Neuronals, LSTM, GRU.

Abstract– The study of human mobility has always been of interest to cities and transport models. However, new technologies, and especially sensors and mobile phones, have opened up many opportunities to study human mobility as they allow large amounts of mobility data to be collected massively and at virtually no cost. In recent years, several mobility studies have been presented, which process and analyze mobility data in different contexts. In the same vein, this work presents the results of a comparative study of various machine learning methods to classify the modes of transport used by people in their mobility. The data set is *GeoLife* –of international reference– which contains traces of mobility of 182 people in the city of Beijing, China. With this study, it has been possible to understand the concepts and models of human mobility as well as the application of appropriate classification methods.

Keywords– Human Mobility, Modes of Transport, GeoLife Dataset, Mobility Traces, GPS, Classification, Random Forest, Logistic Regression, SVM, Neural Networks, LSTM, GRU.

1 Introducció – Context del treball

Aquest treball se situa en el context de processament i anàlisi de dades de traces reals de mobilitat humana amb mètodes d'aprenentatge computacional (*Machine Learning*). L'estudi de la mobilitat humana ha sigut un problema d'interès des de fa molt de temps i des de diferents vessants. Des de la perspectiva social l'interès ha sigut en entendre el comportament de la mobilitat humana relatiu a grups demogràfics diferents i la seva relació amb les activitats quotidianes, especialment, les de caire social, tant per la construcció d'edificis com pel màrqueting i el comerç. Des de la perspectiva de gestió de transport, l'interès ha sigut en poder utilitzar models de mobilitat humana per millorar el transport, i fer-lo més eficient, sostenible i de qualitat pels usuaris.

Es pot dir, per tant, que l'estudi de la mobilitat humana és un camp multi-disciplinar i versàtil. És per això que la rellevància dels estudis de la mobilitat humana han rebut un impuls significatiu durant les darreres dècades degut a l'Internet i les noves tecnologies. En efecte, amb el ràpid desenvolupament de les tecnologies mòbils i Cloud s'ha incrementat notablement la capacitat de capturar i analitzar traces reals de mobilitat urbana en zones d'interès d'una ciutat, d'una regió, país o fins i tot a escala global. Abans,

• E-mail de contacte: jordi.xhafa@e-campus.uab.cat
• Menció realitzada: Computació
• Treball tutoritzat per: Prof. Ramón Baldrich (Departament de les Ciències de la Computació, UAB)
• Curs 2020/21

realitzar un experiment social per estudiar la mobilitat d'un grup de persones era complicat, necessitava molt de temps, era costós i de petita escala. Avui en dia, amb els sensors i telèfons mòbils, s'ha fet molt més fàcil recollir traces de mobilitat, pràcticament sense cost i a gran escala.

D'aquí que ha esdevingut rellevant estudiar els patrons de mobilitat a partir d'aquestes traces per obtenir informació valuosa. Aquesta informació pot ser utilitzada per diversos agents governamentals, econòmics i socials per prendre decisions sobre mobilitat i infraestructura.

L'objectiu del treball és, per una banda, entendre els diferents models de la mobilitat humana, i per l'altra, fer un estudi comparatiu de diferents mètodes d'anàlisi i classificació d'aprenentatge computacional pels modes de transport.

Actualment existeixen conjunts de dades que recullen la mobilitat en l'espai i en el temps d'un conjunt de persones com a seqüència de posicions geogràfiques i marques de temps. Aquests conjunts es poden analitzar amb vàries metodologies per aconseguir patrons de mobilitat, i també es poden aplicar xarxes neuronals per fer prediccions de tipus de mobilitat en el futur. Com a tal, s'ha escollit el conjunt de dades GeoLife [14], un conjunt de traces reals, de referència internacional, desenvolupat amb el suport de Microsoft, a Pequín (Xina).

Estructura del document: La resta d'aquest document s'estructura com segueix. A la Secció 2, es presenta el resultat de la cerca d'informació sobre el problema de la mobilitat humana i es presenten els trets principals dels estudis analitzats. Se segueix a la Secció 3 amb una breu descripció i definició del problema de la mobilitat humana considerat en aquest treball. Es marca així l'àmbit de l'estudi, ja que existeixen molts models i problemes que es poden estudiar en aquest context. A continuació, a la Secció 4, es presenta la metodologia i el desenvolupament de l'estudi. Els resultats i la seva avaluació és presenten a la Secció 5. Al final, a la Secció 6, es presenten les conclusions i possibles línies de continuació d'aquest treball.

2 L'ESTAT DE L'ART EN ESTUDIS DE LA MOBILITAT HUMANA

Per començar, s'ha estudiat l'estat de l'art amb la literatura existent de fonts científiques, per poder entendre els conceptes i els diferents models que existeixen. Per això, es van identificar varis articles d'interès, i al final, es van seleccionar articles dels últims 5 anys: tres articles de Springer, dos d'Elsevier, i tres d'IEEE per incloure en l'estudi. En base a aquesta cerca/estudi, s'han tret aquestes conclusions sobre els trets principals de la mobilitat humana (es poden trobar els detalls particulars per cada font d'informació en l'Apèndix A.1):

- Tipus de mobilitat: Vianant, transport, mixt.
- Tecnologies de localització: GPS, Wi-Fi, Bluetooth, 3G/4G.
- Nombre d'usuaris: Entre desenes i milions.
- Temps de seguiment: Entre una setmana i varis anys.
- Focus de l'estudi: Patrons de mobilitat, patrons de comportament i prediccions.

També es van cercar, principalment a Github, conjunts de dades, projectes i llibreries útils per l'estudi. Es van analitzar particularment els desenvolupats en llenguatge Python.

3 FORMULACIÓ DEL PROBLEMA

Dins del context de l'estudi de la mobilitat humana es poden formular molts problemes. En efecte, la mobilitat humana és complexa per poder ser analitzada en la seva totalitat, ja que té moltes aplicacions. Per tant, típicament, els estudis consideren algunes particularitats de la mobilitat que siguin d'interès per entendre-la o modelar-la. En qualsevol cas, tots els problemes que es puguin formular es basen en característiques d'espai-temps (*spatio-temporal*) de la mobilitat.

El problema de l'estudi de la mobilitat humana, per tant, es pot formular com: *"donat un conjunt de traces de mobilitat d'un grup de persones en una àrea geogràfica i durant un temps suficient llarg, processar, analitzar i extreure informació sobre la mobilitat, patrons de comportament, caracterització de les traces, classificació de punts d'interès, etc., considerant la dimensió d'espai i la del temps, separadament o de forma conjunta"*.

Una traça o trajectòria de mobilitat es defineix com una seqüència de punts per on ha passat una persona durant la seva mobilitat, normalment formada per una seqüència de triples de format $< timestamp, latitude, longitude >$. Algunes consideracions importants sobre les traces són:

1. Les traces individuals generades per persones diferents són independents –cada persona es mou amb independència de la resta de les persones que participen en l'estudi. No es considera aquí el moviment en grup, si bé té sentit i s'estudia en varis treballs.

2. Una persona genera moltes traces durant el temps de monitorització (de l'odre de centenars). Per tant, en conjunt, les persones que formen part de la recollida de dades poden generar grans quantitats de dades (de l'orde de GBs, com és el cas del conjunt de dades d'aquest treball).

3. Les traces no tenen per què ser iguals; algunes són més llargues o curtes que altres. A més, la longitud de la traça com a quantitat de segments no té perquè tenir relació directa amb la longitud de temps de la traça; depèn del mitjà de transport, la velocitat, etc.

4. Les característiques temporals de les traces són diverses degut al mitjà de mobilitat, de la velocitat del mateix, del temps d'estada en els punts, així com de les franges horàries (per exemple, dies laborals vs. cap de setmana), etc. Això dóna lloc a estudiar els hàbits de mobilitat de les persones entre punts d'interès en una franja d'interès.

El cas més típic, que s'ha considerat també en aquest treball, és el de dades GPS recollides a partir dels telèfons mòbils, ja que permeten descriure les traces de moviment d'individus amb alta precisió i freqüència temporal. Les

dades GPS són per tant considerades fonts de dades importants per analitzar els patrons de la mobilitat humana.

Com es pot apreciar, el problema de processar, analitzar i classificar patrons de comportament, trajectòries, punts d'interès, etc., és un problema complex, no només des de la perspectiva algorísmica, si no que també des de la perspectiva computacional (requeriments sobre l'eficiència d'execució i de memòria). Per això, sovint, es veu necessari l'ús de llibreries específiques de paral·lelisme com és el *joblib* de Python.

Delimitació del problema

En aquest estudi el problema de la mobilitat s'ha delimitat com segueix:

- Processament de les dades per identificar els anomenats *"stop locations"* –que no són punts concrets de les traces sinó que compleixen certes propietats d'àrea (radi d'un cercle) i temps (temps d'estada).
- Processament del resultat de les *"stop locations"* (*clustering*) per extreure les anomenades *"destinations"* que són petites àrees visitades per les persones durant un període de temps prefixat.
- Classificació dels modes de transport i combinacions d'ells, empleats per les persones en les seves trajectòries.
- Predicció de modes de transport a utilitzar per les persones.
- Estudi comparatiu entre les diferents metodologies de classificació utilitzades.

4 Metodologia i Desenvolupament de la Proposta

Per a la realització de l'estudi, un cop analitzada la literatura i entesos els trets principal del problema, s'ha procedit a escollir el conjunt de dades (entre varis disponibles) i acte seguit, s'han escollit els mètodes d'anàlisi (també entre varis disponibles).

4.1 Conjunt de dades: el conjunt GeoLife

Per escollir el conjunt de dades, s'han seguit aquests criteris, gràcies a l'anàlisi de l'estat de l'art realitzat abans:

- Que sigui un conjunt de referència en els estudis de la mobilitat, disponible i obert a tothom.
- Que sigui de traces reals (sense dades simulades, que també formen part d'alguns treballs analitzats a la literatura).
- Que sigui un conjunt de dades riques per la participació d'un nombre gran de persones i durant un temps raonablement llarg.
- Que sigui multi-modal, és a dir, que inclogui varis formes de transport alhora (vianant, vehicle, etc.).
- Que tingui (almenys una part de) les dades etiquetades (*labeled*).

En base d'aquests criteris, s'ha escollit el conjunt de dades GeoLife ja que satisfà tots els criteris de selecció. Aquest conjunt de dades de trajectòries GPS va ser recollit al projecte GeoLife[1] (Microsoft Research Asia) per 182 usuaris en un període de més de tres anys (d'abril de 2007 a agost de 2012). El conjunt de dades consta de 17.621 trajectòries enregistrades pels dispositius mòbils habilitats per GPS.

4.1.1 Freqüència de recollida de dades de la mobilitat

Quan es tracta d'estudis basats en la recollida de dades sensorials, com és el cas de la mobilitat humana, la freqüència de mesura és molt important. Per una banda, una alta freqüència garanteix fidelitat a la realitat, permet cobrir una espectre de temps gairebé continu, i permet cobrir més punts geogràfics que no es detectarien amb una freqüència baixa. Per això, el conjunt de Geolife té una elevada escala espacial i temporal. Això s'ha pogut comprovar empíricament, observant que la freqüència de registrament de posicions del conjunt és de cada 1-5s, amb la majoria de cops sent 2 o 3 segons, com es pot observar en la Fig. 1.

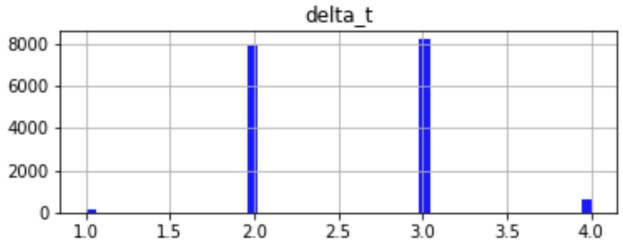

Fig. 1: Freqüència de recollida de dades (extreta de les traces d'una persona). L'eix x representa els segons que han passat entre mesures, i l'eix y el nombre de mesures.

4.2 Mapeig de les traces en l'àrea de Pequín (Xina)

Utilitzant la llibreria *pptk* de Python s'han pogut mapejar els punts de les traces en l'àrea concreta per poder apreciar visualment la densitat dels punts de les traces, com es pot veure de la Fig. 2.

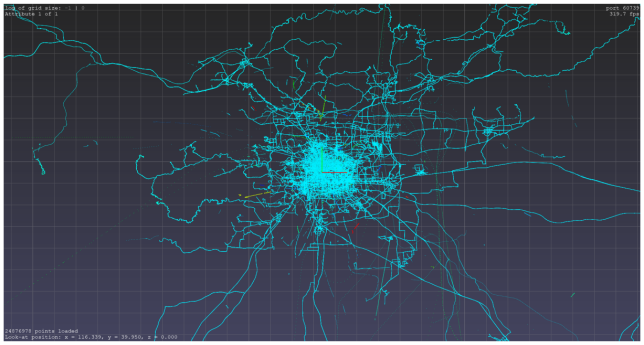

Fig. 2: Mapeig de les traces en l'àrea de Pequín (Xina).

[1] https://www.microsoft.com/en-us/download/details.aspx?id=52367

4.2.1 Dades sense etiquetar i etiquetades

Val a dir que el conjunt de dades GeoLife té una tercera part (aprox.) de les dades etiquetades amb el mode de transport. Aquest fet ens permet aplicar mètodes per classificació / predicció de modes de transport en les traces.

4.2.2 Nota sobre la privacitat de les dades de mobilitat

Qualsevol estudi que implica o recull dades de persones pot necessitar garantir criteris de privacitat, d'ètica, etc. En el cas de la mobilitat humana, es tractaria, per exemple, de no permetre la identificació individual de les persones involucrades en l'estudi. Per això, en el cas de GeoLife, com també d'altres conjunts estudiats, el conjunt de dades no inclou cap dada personal i per tant les trajectòries no es poden associar a cap persona en concret.

4.3 Mètodes d'anàlisi i classificació

Per fer l'estudi comparatiu, s'han escollit varis mètodes d'aprenentatge computacional amb l'objectiu que siguin mètodes de naturalesa diversa: mètodes per la classificació i mètodes de predicció. Concretament, s'han estudiat aquests mètodes: Random Forest, Logistic Regression, SVM (Support Vector Machines), LSTM (Long Short-Term Memory) - un tipus de xarxa neuronal recurrent - i la seva variant GRU (Gated Recurrent Unit).

Abans però, s'han processat les traces de mobilitat per calcular primer els anomenats *stop locations* i després, en base d'ells, les destinacions.

4.3.1 *Stop locations* –Punts de parada

Els stop locations, o punts de parada, es defineixen de la següent manera:

> *Un stop location és un lloc en l'àrea de mobilitat (un cercle) amb radi de 'x' metres, on un o més usuaris han estat durant més de 'y' minuts.*

El motiu per primer calcular els punts de parada a partir de totes les traces de les persones és, per una banda, poder descartar punts per on hi passen usuaris sense parar i, per l'altra, poder identificar els llocs en els quals els usuaris hi paren i passen un temps. Per exemple, un punt de parada podria ser una botiga, un hotspot turístic, etc. Cal notar, per tant, que es crea com un nou conjunt de punts, que és més reduït que el conjunt original.

El nombre resultant de punts de parada depèn dels paràmetres radi i temps d'estada, de manera que tindríem més punts de parada per radis grans i temps d'estada petits, i per altra banda tindríem menys punts de parada per radis petits i temps d'estada grans. D'aquesta manera es pot ajustar el nombre de punts d'estada en funció de les zones (algunes són més denses que altres, tenen més activitat, etc.). Aquest mètode s'implementa fent un filtre dels punts en les traces de l'usuari.

En la Fig. 3 es pot apreciar la definició d'un stop location per radi = 50m i temps d'estada mínim de 15 minuts.

S'ha utilitzat el codi font d'un projecte existent [2]. Casualment, aquest projecte també s'ha desenvolupat sobre un dataset situat a Pequín, però diferent del GeoLife. Per això, s'han hagut de fer varies modificacions per tal d'adaptar el

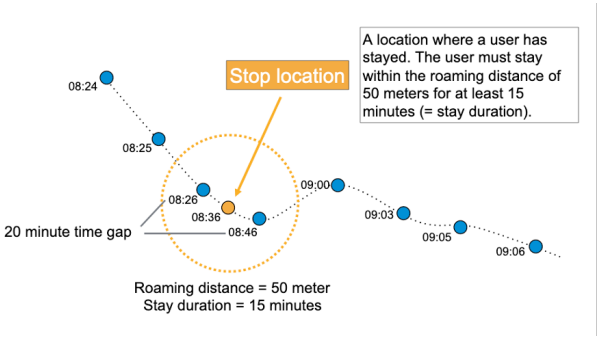

Fig. 3: Un exemple de punt de parada.

codi al dataset de GeoLife, i també s'han hagut d'arreglar errors, o implementar de nou, ja que vàries de les funcions que s'utilitzaven al codi estaven obsoletes.

A la Fig. 4 es poden observar els stop-locations obtinguts de les traces de 5 usuaris.

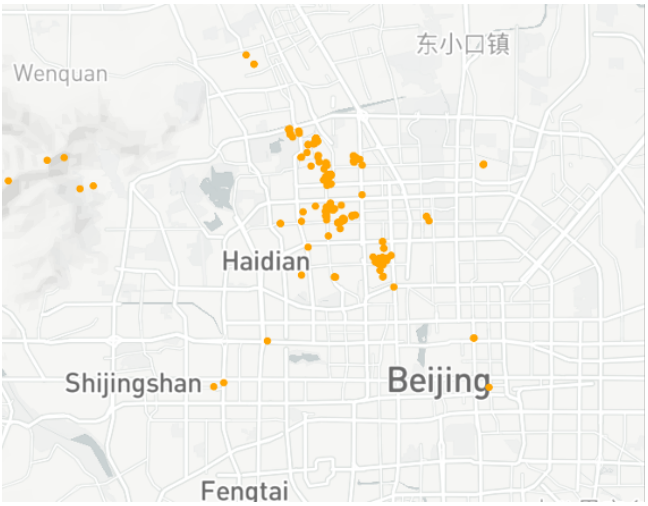

Fig. 4: Punts de parada de GeoLife dataset (extret del processament de les traces d'una mostra usuaris).

4.3.2 *Destinations* –Clustering de punts de parada

Tot i que els punts de parada ja representen punts d'interès en la zona de mobilitat, en les grans ciutats - com és el cas de Pequín - el nombre de punts de parada resultaria molt gran. Es pot definir un tercer nivell de "punts", anomenats destinacions, que similarment representen un clustering dels punts de parada trobats. Un bon exemple de destinació seria un centre comercial, on les botigues visitades serien els punts de parada.

Per cada clúster de punts de parada, el seu centre (anomenat *medoid*), és el punt central de destinació. Clustering, en aquest cas, es fa per distància euclidiana, fixant un radi que engloba punts de parada, però sense relació amb el temps d'estada, ja que la dimensió temporal prové dels punts de parada.

En la Fig. 5 es pot apreciar la definició de destinacions per un radi de 50m.

En canvi, a la Fig. 6 es poden observar les destinations obtingudes fent clustering del conjunt dels punts de parada de les traces de 5 usuaris (representats a la Fig. 4). El tamany dels punts en el mapa indiquen la freqüència amb la

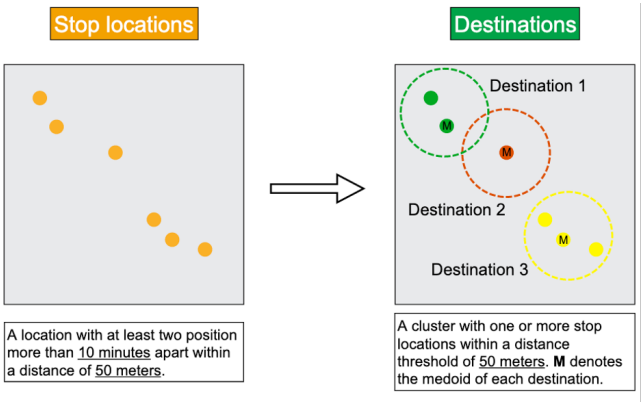

Fig. 5: Un exemple de destinacions.

que s'han visitat aquestes destinacions. S'utilitzen diferents colors per distingir-los visualment.

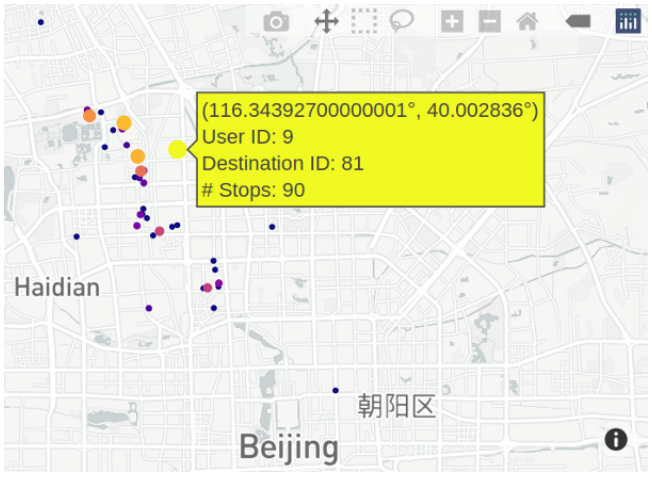

Fig. 6: Destinacions de GeoLife dataset (extret del clustering dels punts de parada d'una mostra usuaris).

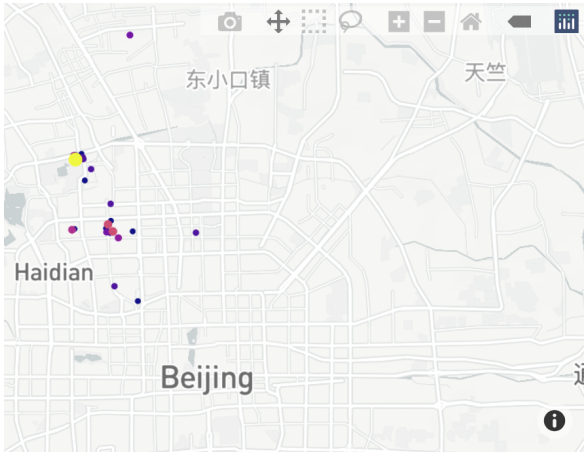

Fig. 7: Destinations pels dies laborals (per una mostra d'usuaris).

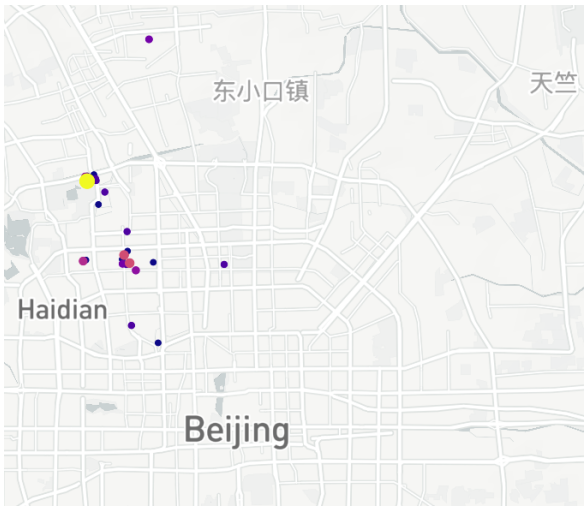

Fig. 8: Destinations pel cap de setmana (la mateixa mostra d'usuaris).

4.3.3 Filtratges de traces per temps i top ranking de destinacions

S'han implementat unes funcions per filtrar el dataset original segons dies de la setmana (laborables/cap de setmana) i hores del dia (00:00-08:00 / 08:00-18:00 / 18:00-00:00). Aquests subsets del dataset original poden ser útils per analitzar el comportament dels usuaris segons el moment del dia i de la setmana en que es trobin. Així, del conjunt original s'han obtingut, primer dos grans subconjunts per dies laborals i cap de setmana, respectivament, i després per cadascun d'ells, es poden obtenir 3 conjunts, un per cada franja horària (un total de 6 conjunts).

En la Fig. 7 podem veure una representació en el mapa de les destinations pels dies laborals (DL-DV) i la Fig. 8 pel cap de setmana.

Temps de processament: Pel que fa el temps de processament, el que pren més temps és el de stop locations, que per la mostra dels usuaris, s'executa en l'ordre d'una hora. En canvi, un cop calculat els stop locations, el càlcul dels destinations és ràpid (ordre d'un minut).

Finalment, s'ha desenvolupat una funció que crea un ranking de les destinacions més visitades (tant d'un usuari com de tots els usuaris). Aquestes destinacions, però, estan donades en coordenades (latitud, longitud); per això, s'ha fet servir la llibreria geopy [10], que retorna el nom de l'edifici o l'adreça que es pot trobar en cadascuna de les coordenades. Això també pot ser útil per analitzar quines destinacions són més freqüents depenent del temps.

4.4 Classificació de modes de transport

Tal i com s'ha comentat, el conjunt de dades GeoLife conté traces etiquetades de 69 usuaris (més de una tercera part del total de traces), indicant els modes de transport en cada segment de la trajectòria. Les etiquetes utilitzades són: *walking, bus, train, car, bike, taxi, subway*. Aquest conjunt etiquetat possibilita aplicar mètodes d'aprenentatge supervisats (*supervised learning*).

En base d'aquest conjunt etiquetat, el problema plantejat és la classificació del mode de transport escollit per un usuari al llarg d'una traça. Per la seva implementació, es defineixen els *features* en base dels quals es fa el *training* i *testing* del classificador. Val a dir que es poden definir molts features: a nivell local (nivell de trajectòria, local features) o a nivell global (de totes les trajectòries, global features).

En ambdós casos els features són típicament calculats amb mètodes estadístics, i hi ha autors que consideren fins a un dotzena en base dels càlculs de mitjana, desviació, valors esperats, top valors (segons un ranking), valors per diferent quartils, etc. [11].

En aquest cas, els features definits, i després calculats del conjunt de dades, són els següents 6 features relacionats amb la velocitat i l'acceleració, calculats en base dels timestamps en les traces:

- Velocitat (`v_ave, v_med, v_max`)
- Acceleració (`a_ave, a_med, a_max`)

Per l'estudi, s'han considerat els següents classificadors ([7]): Random Forest (RF), Logistic Regression (RL), Support Vector Machine (SVM), Long Short-Term Memory (LSTM), i la seva variant Gated Recurrent Unit (GRU).

4.4.1 Random Forest (RF)

RF és un mètode d'aprenentatge per a la classificació, que combina el resultat de múltiples arbres de decisió per aconseguir un resultat únic, essent així la sortida la classe seleccionada per la majoria dels arbres. Varis treballs han demostrat que el rendiment de RF és sensible a les característiques de les dades.

Els RF tenen tres paràmetres que s'han de fixar abans de l'entrenament: la mida del node, el nombre d'arbres i el nombre de característiques.

Paràmetres i resultats: S'ha utilitzat la funció *RandomForestClassifier()* de la llibreria *sklearn*. S'ha indicat com a paràmetre d'entrada del model un nombre d'estimadors (nombre d'arbres a ser utilitzats en el *forest*) igual a 18, com a nombre òptim a partir del qual ja no creix l'*accuracy*. La resta de paràmetres són els valors per defecte de la llibreria.

Els resultats d'*accuracy* obtinguts són:

```
Score on training set: 99,59%
Score on test set: 75,14%
```

Aquests han sigut els millors resultats obtinguts dels 3 mètodes de classificació (excloent LSTM i GRU). Com a tal, es poden analitzar més a fons els resultats observant els valors de *precision*, *recall* i *f1-score* per a cada un dels modes de transport:

	precision	recall	f1-score
bike	0.89	0.87	0.88
bus	0.60	0.63	0.61
car	0.83	0.84	0.84
subway	0.74	0.64	0.68
taxi	0.50	0.03	0.06
train	0.67	0.60	0.63
walk	0.72	0.85	0.78
avg/total	0.77	0.78	0.76

Es poden millorar encara més els resultats si es generalitzen els modes de transport, ajuntant *bus*, *car* i *taxi* en una sola classe, anomenada *vehicle*, ja que tenen comportaments similars.

```
Score on training set: 100%
Score on test set: 84,59%
```

	precision	recall	f1-score
bike	0.86	0.90	0.88
subway	0.84	0.80	0.82
vehicle	0.91	0.88	0.90
walk	0.85	0.85	0.85
avg/total	0.87	0.87	0.87

4.4.2 Logistic Regression (LR)

LR és una generalització de Linear Regression. És un algorisme predictiu, que utilitza variables independents (els features explicats més amunt) per predir la variable dependent (mitjà de transport), igual que la regressió lineal, però amb la diferència que la variable dependent ha de ser variable categòrica, com és el cas del mode de transport. Es pot utilitzar tant per classificació binària com multi-class. Val a dir que LR funciona bé per conjunts que no tenen valors atípics a les dades (outliers) ni correlació entre les variables independents. La presència dels outliers o de correlacions és motiu perquè LR doni resultats pobres. En el cas de GeoLife s'ha pogut comprovar que la presència d'algunes traces d'usuari fa que es degradi la precisió de LR.

Paràmetres i resultats: S'ha utilitzat la funció *LogisticRegression()* de la llibreria *sklearn*. Aquest model s'ha executat sense paràmetres (valors per defecte).

Els resultats obtinguts han sigut els següents:

```
Score on training set: 66,58%
Score on test set: 63,83%
```

Com es pot observar, els resultats mostren que LR no és el mètode més adequat a utilitzar, donat el seu *accuracy* relativament baix.

4.4.3 Support Vector Machines (SVM)

El principi de funcionament de SVM és que per un conjunt de *features* $x_1, x_2, \ldots, x_n$, i el resultat de la classificació y, calcular els pesos de cada *feature* de manera que la combinació lineal $w_1x_1 + w_2x_2 + \cdots + w_nx_n$ predigui el valor de sortida y. SVM, però, utilitza mètodes d'optimització més sofisticats per trobar els pesos i calcular prediccions precises.

Paràmetres i resultats: S'ha utilitzat la funció *SVC()* de la llibreria *sklearn*. Aquest model s'ha executat sense paràmetres (valors per defecte).

Els resultats obtinguts han sigut els següents:

```
Score on training set: 87,52%
Score on test set: 69,61%
```

4.4.4 Long Short-Term Memory (LSTM)

LSTM és una variant de xarxa neuronal recurrent (RNN). El seu tret principal és recordar informació durant períodes relativament llargs de temps. Degut a la utilització de

'memòria' i la dimensió de temps, LSTM ha mostrat avantatges comparat a les xarxes neuronals tradicionals en la classificació de seqüències de dades, com és el cas de les traces de mobilitat [8, 9].

La idea central de LSTM és mantenir informació sobre els inputs passats (en l'anomenat *remember vector*) durant un cert temps (per tant ha de tenir capacitat per guardar i eliminar informació en els anomenats *cell states*) i utilitzarlos per la tasca de l'estat present.

Paràmetres i resultats: Els següents són alguns dels paràmetres i els seus valors utilitzats per la fase de **Training**:

```
batch_size = 16
output_size = 5
hidden_dim = 128
trip_dim = 7
n_layers = 2
drop_prob = 0.2
lr= 0.001 #learning rate
loss_function = CrossEntropyLoss
optimizer = torch.optim.Adam
epochs = 10
```

Per fixar aquests valors s'han fet múltiples execucions per veure l'efecte de valors diferents tant en l'accuracy com en el temps d'execució i així afinar-los. Per exemple, el nombre d'*epochs* és important ja que si s'executa amb un nombre d'epochs baix la xarxa no convergeix a una precisió alta, mentre que si es posa un nombre d'epochs massa alt no necessàriament comportaria millora però sí trigaria molt més per finalitzar.

A la Fig. 9 es pot apreciar que el millor nombre d'epochs és igual a 9, ja que fins a 9 epochs la precisió va incrementant, però a partir de 10 epochs s'estanca.

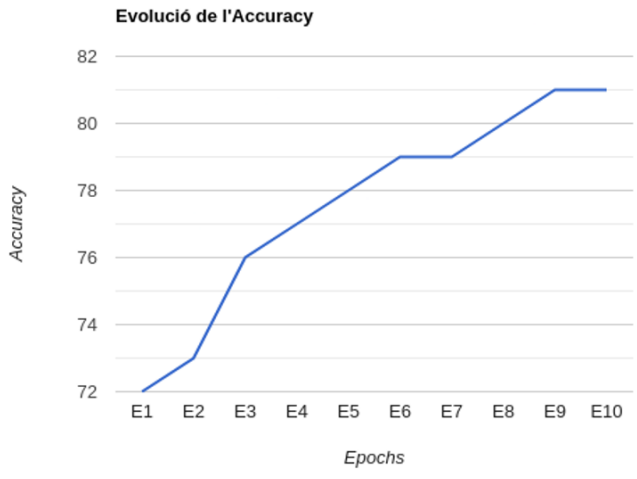

Fig. 9: Estudi del valor del nombre d'epochs per la LSTM.

A part d'això, els dos paràmetres més importants són `n_layers`, que especifica el nombre de *hidden layers* que hi haurà a la xarxa, i `hidden_dim`, que especifica el nombre de nodes a cada *hidden layer*. S'ha estudiat com varia l'accuracy i el temps d'execució en funció d'aquests dos paràmetres.

Els resultats de LSTM pel training es poden veure a la Taula 1 i els resultats pel testing a la Taula 2.

Com es pot observar, el millor valor d'*accuracy* pel training ha sigut 83%, i pel testing 81%. En ambdós casos, s'ha aconseguit aquest *accuracy* en el menor temps possible amb `n_layers=2` i `hidden_dim=128` (260s i 5.14s respectivament).

TAULA 1: RESULTATS DE LSTM (TRAINING). ACCURACY I TEMPS

LSTM Training		hidden_dim		
		64	128	256
n_layers	1	81% (150s)	82% (136s)	81% (151s)
	2	80% (246s)	83% (260s)	83% (337s)
	3	81% (391s)	82% (397s)	80% (552s)

TAULA 2: RESULTATS DE LSTM (TESTING). ACCURACY I TEMPS

LSTM Testing		hidden_dim		
		64	128	256
n_layers	1	79% (3.39s)	80% (3.41s)	80% (4.55s)
	2	79% (5.02s)	81% (5.14s)	81% (5.97s)
	3	81% (6.88s)	81% (7.04s)	81% (7.26s)

4.4.5 Gated Recurrent Unit (GRU)

També s'ha estudiat la variant GRU de la xarxa neuronal RNN, que té els mateixos paràmetres que la xarxa neuronal LSTM. Ambdues xarxes utilitzen un mecanisme de portes (*gates*) per implementar la memòria. Però, GRU és considerat menys complex que LSTM ja que manté menys portes (*reset, update*) mentre que LSTM utilitza tres portes (*input, output, forget*). En principi GRU és més ràpid pel training, com es pot apreciar també en les taules a continuació (Taules 3 i 4).

Els resultats pel training es poden veure a la Taula 3 i els resultats pel testing a la Taula 4.

TAULA 3: RESULTATS DE GRU (TRAINING). ACCURACY I TEMPS

GRU Training		hidden_dim		
		64	128	256
n_layers	1	81% (126s)	81% (137s)	82% (149s)
	2	83% (222s)	84% (256s)	84% (332s)
	3	84% (344s)	84% (399s)	84% (521s)

Com es pot observar, el millor valor d'*accuracy* pel training ha sigut 84%, i pel testing 82%. En ambdós casos,

TAULA 4: RESULTATS DE GRU (TESTING). ACCURACY I TEMPS

GRU Testing		hidden_dim		
		64	128	256
n_layers	1	78% (3.34s)	80% (3.42s)	80% (3.58s)
	2	81% (5.26s)	82% (5.73s)	82% (6.17s)
	3	82% (7.67s)	82% (7.11s)	82% (7.48s)

s'ha aconseguit aquest *accuracy* en el menor temps amb `n_layers=2` i `hidden_dim=128` (256s i 5.73s respectivament).

5 AVALUACIÓ DELS RESULTATS

En aquest darrer apartat es farà una breu avaluació i comparació dels mètodes estudiats.

S'han tingut en compte principalment l'*accuracy* i l'eficiència de processament.

5.1 Comparació dels *accuracies* obtinguts

A la Taula 5, es presenten de forma resumida els resultats obtinguts pel *training* i *testing* de tots els mètodes.

TAULA 5: COMPARACIÓ D'ACCURACIES OBTINGUTS

Model	Training Acc.	Testing Acc.
RF	99,59%	75,14%
LR	66,58%	63,83%
SVM	87,52%	69,61%
LSTM	83,00%	81,00%
GRU	84,00%	82,00%

Com es pot observar, GRU presenta el millor *testing accuracy*, seguit per LSTM, Random Forest, Support Vector Machine, i per últim Logistic-Regression.

5.2 Eficiència de processament

Pel que fa a l'eficiència, s'han considerat dos aspectes:

1. Processament del *raw data*. Això implica processar grans quantitats de dades (d'ordre de GB), filtrar-les i estructurar-les per tal de tenir-les a punt per la classificació.
   En aquest cas, s'han obtingut temps de processament grans (entre 30 minuts i 1 hora).
2. Processament dels mètodes de classificació. En tots el casos, el temps de testing ha estat per sota de 10 segons. Per tant, el temps més rellevant és el de training. Els temps obtinguts es poden observar a la Taula 6.

Com es pot observar, el model de classificació més ràpid és el de Random Forest, seguit de GRU, LSTM, Logistic-Regression, i per últim Support Vector Machine. Això és comprensible, ja que la funció per SVM de *sklearn* pot arribar a ser molt lenta per conjunts de dades grans.

TAULA 6: TEMPS DE TRAINING DELS MODELS DE CLASSIFICACIÓ

Model	Temps
RF	26 segons
LR	7 minuts i 48 segons
SVM	30 minuts i 58 segons
LSTM	4 minuts i 20 segons
GRU	4 minuts i 16 segons

5.3 Rendiment-Eficiència

Tal com s'ha pogut observar a les taules dels resultats, els dos mètodes de classificació clarament millors han sigut Random Forest en quant a temps, i GRU en quant a *accuracy*.

La decisió d'escollir un o l'altre dependrà primordialment de les necessitats que es tinguin de temps d'execució. Si és crític que el temps sigui el mínim possible, s'hauria d'escollir Random Forest, obtenint així un *accuracy* bastant bo. En canvi, si el temps no és crític i volem aconseguir el millor *accuracy* possible, s'hauria d'escollir GRU.

6 CONCLUSIONS

En aquest treball s'ha fet un estudi comparatiu de diversos mètodes de classificació sobre modes de transport (caminar, bus, cotxe, tren, etc.) en base de les traces de mobilitat humana. El problema de l'estudi de la mobilitat humana recentment ha rebut molt d'interès per la seva aplicació en models de ciutats, mobilitat urbana, impacte social en la vida de les persones, etc. Per fer l'estudi s'ha utilitzat un conjunt de dades (*raw data*) del projecte GeoLife de Microsoft realitzat a Pequín, Xina. Aquest conjunt (de l'ordre de GB) recull característiques interessant pel que fa l'estudi de mobilitat. Per poder fer l'estudi, s'ha cercat i analitzat la literatura i després s'han escollit 5 models d'aprenentatge computacional (Random Forest, Logistic Regression, Support Vector Machines, i les xarxes neuronals LSTM i GRU). Amb l'estudi s'ha pogut observar que Random Forest és el més ràpid i alhora un bon classificador dels modes de transport, encara que la xarxa GRU obté millors *accuracies* de testing, però necessita més temps per convergir.

Línies de continuació d'aquest treball: Pel que fa a les línies de continuació d'aquest treball considero que el problema de l'estudi de la mobilitat humana és força interessant i dóna moltes opcions de treball futur. Entre aquestes, es podrien mencionar:

- Estudiar la relació de la mobilitat humana amb les activitats realitzades per les persones en la seva vida quotidiana. Això permetria veure patrons de comportament de les activitats i la seva classificació, especialment, per veure activitats de caire socials En particular, aquest estudi es pot fer per dimensió de temps (dies laborals vs. cap de setmana, i/o per franges horàries durant el dia).
- Les àrees de stop locations i destinations (calculats en aquest projecte) podrien utilitzar-se per identificar les

àrees on fer publicitat i anuncis d'interès ja sigui per empreses o entitats públiques (ajuntaments, etc.).

- Estendre la llista de features amb d'altres més sofisticats per poder extreure més informació de les traces de la mobilitat.
- Per últim, el processament massiu en paral·lel de les traces de processament seria força interessant per poder analitzar grans volums de dades. Pel que s'ha pogut observar, per una banda els algorismes utilitzats a la literatura no estan prou optimitzats, i per l'altra, es podria aprofitar que plataformes com ara Spark pel *cluster computing*, ofereixen la possibilitat d'escriure programes en Python per ser executat en una plataforma Spark.

7 AUTO-AVALUACIÓ

Al final del treball amb aquest projecte, considero que ha sigut un projecte interessant. He après coses noves com ara el tema de la mobilitat humana i treballar amb un conjunt de dades reals (GeoLife) de referència internacional. Pel que fa els mètodes estudiats, he pogut aprofundir en els conceptes i els models de xarxes neuronals, i la LSTM i GRU en particular.

Voldria mencionar que també he hagut d'utilitzar noves llibreries de Python. També m'han sigut útils varis projectes en Github sobre el tema.

Per últim, de les assignatures cursades durant la carrera, m'han ajudat més les de Coneixement, Raonament i Incertesa, i Aprenentatge Computacional, si bé altres assignatures m'han ajudat d'una manera o altra.

Per tot, considero que s'han acomplert els objectius del treball establerts a l'inici del projecte.

AGRAÏMENTS

Voldria agrair al Prof. Ramón Baldrich per l'oportunitat de realitzar el TFG sota la seva direcció, pel seu suport i ajuda al llarg del TFG.

REFERÈNCIES

[1] H. Barbosa, F. Lima, A. Evsukoff, and R. Menezes. The effect of recency to human mobility. *EPJ Data Science*, 4, 2015.

[2] S. Bertoli. Where people stay - extracting destinations from GPS data (GitHub). `https://github.com/sebastianbertoli/Github-internship_human_mobility`.

[3] F. Calabrese, M. Diao, G. Di Lorenzo, J. Ferreira, and C. Ratti. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation Research Part C: Emerging Technologies*, 26:301–313, 2013.

[4] Ch. Yeung Ch. Zhao, A. Zeng. Characteristics of human mobility patterns revealed by high-frequency cell-phone position data. *EPJ Data Science*, 10, 2021.

[5] A. Cuttone, S. Lehmann, and M. González. Understanding predictability and exploration in human mobility. *EPJ Data Sci.*, 7(1):2, 2018.

[6] A. Farrokhtala, Y. Chen, T. Hu, and S. Ye. Toward understanding hidden patterns in human mobility using wi-fi. In *CCECE 2018*, pages 1–4, 2018.

[7] Github. Classificadors. `https://github.com/taspinar/GPSMachineLearning`.

[8] Github. LSTM. `https://senzhangwang.github.io/`.

[9] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997.

[10] Python. Llibreria geopy. `https://pypi.org/project/geopy/`.

[11] Jing Liang *i altres*. An enhanced transportation mode detection method based on GPS data. In *Data Science -3d Int'l Conf. of Pioneering Computer Scientists, Eng. and Educators, ICPCSEE 2017, China, Proc.*, volume 727 of *Communications in Computer and Information Science*, pages 605–620. Springer, 2017.

[12] M. Traunmueller, N. Johnson, A. Malik, and C. Kontokosta. Digital footprints: Using wifi probe and locational data to analyze human mobility trajectories in cities. *Comp., Env. and Urban Syst.*, 72:4–12, 2018.

[13] Sh. Zhang and X. Li. Mobility patterns of human population among university campuses. In *APCCAS 2016*, pages 50–53, 2016.

[14] Y. Zheng, X. Xie, and W. Ma. Geolife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.*, 33(2):32–39, 2010.

APÈNDIX

A.1 Resultats de l'anàlisi de l'estat de l'art

A.1.1 Font: Elsevier

Calabrese i altres [3]:

- Main issue: Understand mobility patterns in a metropolitan area
- Mobility type: All (walking, driving, public transportation), Vehicle Safety Inspection Records
- Data sensing technologies: Cellular network
- Data type: GPS location and timestamp
- Datasets: Not available.
- Data info: Boston Metropolitan Area (USA), 3 months, 1M people
- Patterns studied: Sequence of locations
- Methods used: Statistical analysis, distributions
- Software: –

Traunmueller i altres [12]:

- Main issue: Understand mobility patterns in a metropolitan area
- Mobility type: All
- Data sensing technologies: Wi-Fi
- Data type: Location and time
- Datasets: Not available
- Data info: New York City (USA), 1 week, 800k people
- Patterns studied: Sequence of locations
- Methods used: Network Analysis
- Software: –

A.1.2 Font: Springer

Cuttone i altres [5]:

- Main issue: Which modelling strategies produce better predictability accuracy.
- Mobility type: All
- Data sensing technologies: Mobile (GPS, WiFi), every 15 minutes
- Data type: Location and timestamp
- Datasets: Not available
- Data info: Copenhagen, 3 months - 1 year, 800 people
- Patterns studied: Sequence of cells, places, and time-bins.
- Methods used: Toploc, Markov
- Software: –

Zhao i altres [4]:

- Mobility type: Understand mobility patterns with high-frequency data
- Data sensing technologies: 4G, per sec.
- Data type: Location and timestamp
- Datasets: Not available
- Data info: Shijiazhuang (China), 14 days, 5.3M people
- Patterns studied: Sequence of locations
- Methods used: First-order Markov model
- Software: –

Barbosa i altres [1]:

- Mobility type: Understand the effect of recency on human mobility
- Data sensing technologies: Mobile
- Data type: Location and timestamp
- Datasets: Not available
- Data info: Metropolitan area in Brazil, 6 months, 30k people
- Patterns studied: Sequence of locations
- Methods used: Statistics and probabilities
- Software: –

A.1.3 Font: IEEE

Ghosh i altres [1]:

- Main issue: Predicting next location from sparse GPS data and contextual information
- Mobility type: All
- Data sensing technologies: GPS
- Data type: Location and timestamp
- Datasets: https://www.idiap.ch/dataset/mdc/download
- Data info: Lake Geneva region (Switzerland), 9 months, 200 people
- Patterns studied: Sequence of locations
- Methods used: Hierarchical and layered Hidden Markov Model
- Software: –

Farrokhtala i altres [6]:

- Main issue: Understand mobility patterns
- Mobility type: All

- Data sensing technologies: Wi-Fi
- Data type: Location and timestamp
- Datasets: https://crawdad.org/buffalo/phonelab-wifi/20160309/
- Data info: SUNY Buffalo University Campus (New York), 5 months, 284 people
- Patterns studied: Sequence of locations
- Methods used: Eigenvectors
- Software: –

Zhang i altres [13]:

- Main issue: Understand mobility patterns, differentiating spontaneous and non-spontaneus mobility
- Mobility type: Walking
- Data sensing technologies: Wi-Fi
- Data type: Location and timestamp
- Datasets: Not available.
- Data info: Fudan University Campus (Shanghai), 5 months, 7,7K people
- Patterns studied: Sequence of locations, time spent in locations
- Methods used: Statistics and probabilities
- Software: –