GPS Trace Modality Classification

Diana Juarez Madera[1], Matej Kosec[1], Yi Cao[2]
Department of Aeronautics and Astronautics[1], SCPD[2], Stanford University
{djuarezm, mkosec, ycao4}@stanford.edu

Motivation

Open Street Maps (OSM) is the most popular open-source world-wide digital map. Users can download map vector data for free, and contribute back by uploading their edits. However, in order to be able to also navigate the map effectively and give users accurate predictions of travel time, it is necessary to collect real data of the road network. This is done by associating GPS traces of real trips with physical entities (roads, bridges, etc.) in the OSM database.

This project investigates the feasibility of using unsupervised learning (k-means) to characterize the transportation modes in particular GPS traces. Supervised learning (SVM) is used to validate the clustering performance.

Fig. 1 Satellite image near Palo Alto.

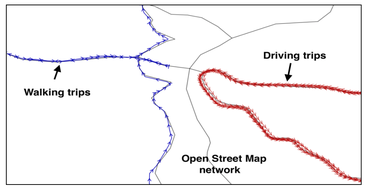

Fig. 2 Walking and driving trips comparison.

Data and Feature Engineering

The project integrates OSM map data [1] consisting of road network geometry, and building footprints with GPS trace data which contains spatial and temporal information. All data used is freely available through the OSM GPX portal and OSM website.

Additionally, each GPS trip was segmented such that it included a single mode of travel. Then 21 features were generated. Features such as maximum speed and acceleration were found to be too noisy and were later disregarded. Ultimately, only 16 were used for the clustering.

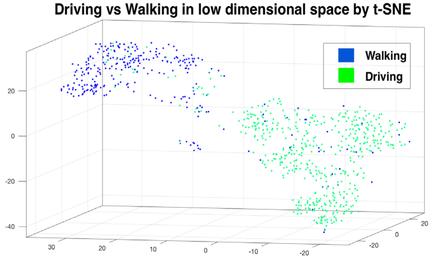

Fig. 3 Visualization of clusters using t-SNE.

The t-Distributed Stochastic Neighbor Embedding (t-SNE) tool projects high dimensional similarity into lower dimension space. It was used to evaluate whether a given set of features can be used to cluster the data-set effectively. The two major data clusters can be identified as walking and driving (see Fig. 3).

Learning Model

The methodology followed incorporates both supervised and unsupervised learning elements across OSM and GPS data as outlined in Fig. 4.

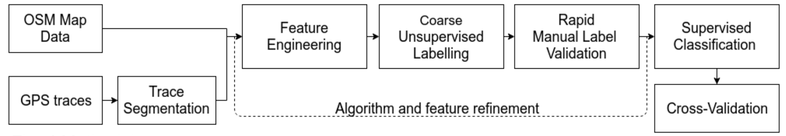

Fig. 4 Methodology.

Unsupervised Labeling: k-means

In order to achieve unsupervised labeling of the data, k-means was utilized to generate two clusters: driving and walking. Trips with a mixture of different transportation modes were discarded. Table 1 gives the mode centroids in terms of the 16 features used.

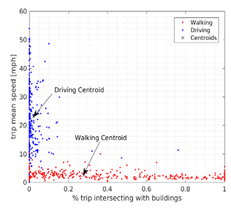

Table 1: Table summary of k-means centroids when using OSM and GPS data

	Speed [mph]		Acceleration [mph^2]		Distance bw. waypts [m]			Building intersec.
Feature	*mean*	*std.*	*mean*	*std.*	*mean*	*max*	*std.*	*percentage of trip*
Walking	2.45	1.63	0.561	0.724	4.88	16.6	2.75	27.8
Driving	23.0	11.9	1.28	1.31	15.26	38.9	7.88	2.13
	Δ Heading [deg/m]		Δ Heading [deg/pt.]		Distance to road [m]			
Feature	*mean*	*std.*	*mean*	*std.*	*min*	*mean*	*max*	*std.*
Walking	67.2	150.6	25.8	31.1	133.4	192.7	361.4	71.8
Driving	32.4	152.4	8.70	18.7	29	171.4	1072	222

Fig. 5 Combination of GPS and OSM data in k-means.

It is observed that the pairing of OSM data (e.g. percentage of trip intersecting a building) with GPS features (such as mean speed), allows for highly effective clustering (see Fig. 5). This is also demonstrated by the k-means silhouette values being close to '1' (Fig. 6).

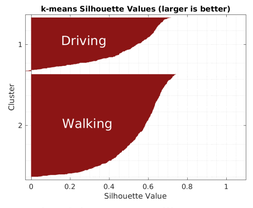

Fig. 6 k-means silhouette.

Supervised Learning: The Benchmark

The manually obtained labels were used in a supervised learning setting to establish a comparison benchmark for the performance of the unsupervised learning (k-means). Past work on GPS trace modality classification indicates that for data collected on smart-phones, SVM provides a high prediction accuracy as compared to other classifiers [2].

To evaluate the SVM performance in this particular application, 10-fold cross validation was implemented in MATLAB. Overfitting was mitigated by applying a feature selection algorithm to reduce the number of features and find the most relevant ones.

Results and Discussion

Table 2 shows that k-means accurately clustered 95.4% of walking trips, and 88.2% of driving trips. Cumulatively this results in an error of just 7.2% in the classification of trips as either walking or driving. Note that the accuracy of the k-means decreases significantly if mixed modes of transport are present. It turns out these are quite common in reality. For instance it is common to walk to the train station, and then take the train to the city.

Table 2: Accuracy of k-means data clustering and labelling

k-means accuracy	**Percentage**	**Trips**
Correctly classified walking trips	95.4	496 of 520
Correctly classified driving trips	88.2	253 of 287
Total of correctly classified trips	**92.8**	**749 of 807**

Fig. 7 shows the performance of the supervised learning benchmark. The results of cross validation and feature selection show that the smallest generalization error achievable was 5%. This corresponds to 4.9% training error with a subset of 8 GPS-derived and OSM-based features. The error performance demonstrates the typical bias-variance trade-off.

Overall these results show that k-means is nearly as effective at labelling data, as SVM is in training and generalization. As such, k-means can be valuable in speeding-up the labelling of GPS traces of trips.

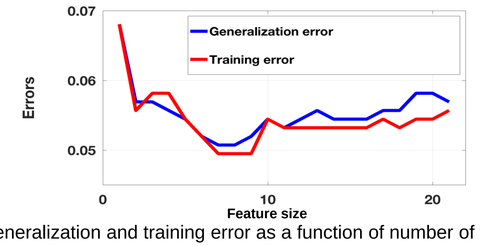

Fig 7. Generalization and training error as a function of number of features (SVM).

Future work

While this project serves as a good proof of concept, future work should extend the method to a larger set of transport modes including: rail, cycling, flying, and ferries. This will require the acquisition of a more extensive data-set, perhaps of the entire United States.

References

[1] Download OpenStreetMap data for this region: California. (2016). GeoFabrik GmbH. Retrieved 20 November 2016, from http://download.geofabrik.de/north-america/us/california.html
[2] M. A. Shafique, E. Hato, A Comparison among various Classification Algorithms for Travel Mode Detection using Sensors' data collected by Smartphones. CUPUM, 2015.