International Journal of Advanced Smart Convergence Vol.7 No.4 27-39 (2018)
http://dx.doi.org/10.7236/IJASC.2018.7.4.27

IJASC 18-4-4

Design of a machine learning based mobile application with GPS, mobile sensors, public GIS: real time prediction on personal daily routes

Hyunkyung Shin

Department of Mathematical Finance, Gachon University, Seongnam-si, Gyeonggi-do, Korea
e-mail: hyunkyung@gachon.ac.kr

Abstract

Since the global positioning system (GPS) has been included in mobile devices (e.g., for car navigation, in smartphones, and in smart watches), the impact of personal GPS log data on daily life has been unprecedented. For example, such log data have been used to solve public problems, such as mass transit traffic patterns, finding optimum travelers' routes, and determining prospective business zones. However, a real-time analysis technique for GPS log data has been unattainable due to theoretical limitations. We introduced a machine learning model in order to resolve the limitation. In this paper presents a new, three-stage real-time prediction model for a person's daily route activity. In the first stage, a machine learning–based clustering algorithm is adopted for place detection. The training data set was a personal GPS tracking history. In the second stage, prediction of a new person's transient mode is studied. In the third stage, to represent the person's activity on those daily routes, inference rules are applied.

Keywords: *global position system(GPS), public geographic information system(GIS) data, mobile sensors, machine learning, inference rules, daily route prediction.*

1. Introduction

GPS log data has simple format with latitude, longitude, and altitude (optional), and time stamp. From this raw GPS data, transition mode (staying, walking, and riding) detection has long been studied by analysis on speed of person's move as described at next section. Recently, it was studied again [1], however, it was on general purpose smartphone device rather than on special purpose positioning sensor. Motion sensor data also has simple format with x-, y-, and z- directional quantity(dx, dy, dz). From this raw data, human behavioral activity analysis has been studied [2]. Identification of human poses such as lay down, sit, stand, walk, run, cycle ride, and car ride and detection of transition from one pose to another has been main topics of the studies. Recently, motion and environmental sensors have been equipped into mobile devices including

Manuscript Received: Sep. 6, 2018 / Revised: Sep. 13, 2018 / Accepted: Sep. 17, 2018
Corresponding Author: hyunkyung@gachon.ac.kr
Tel: +82-31-750-8674, Fax: +82-31-750-5152
Department of Mathematical Finance, Gachon University, Seongnam-si, Gyeonggi-do, Korea

wearable devices as well as smart phone. Accelerator and pressure sensors are the examples.

As can be seen in

Figure **1** where GPS log data put on map layer, it represents daily route. The single colored lines show part of daily route of person with GPS embedded device. The figure shows public GPS data from OpenStreetMap[3] near Washington Square Park at lower Manhattan in New York City. For presentation purpose, we added some of place markers (pin shaped marker with 'x') on the map. The place markers can be obtained from GIS (OpenStreetMap GIS in our case). From the aspect of computational model, it is natural idea that person's daily route can be interpreted as a graph of nodes and links, where a node represents a place where the person visited (e.g., not passing through) while a link does a path on which the person pass through between two places.

Figure 1. Example GPS log data near Washington Square Park, New York City. Courtesy of Open Street map [3].

Conversion from GPS log data into daily route is a mapping from sequence of points (curved lines on the map above) to graph structure, which is not a dimension preserving. It is map from one dimensional line to 1 + α dimensional tree structure. Mathematical stability of the mapping is relied on invariance of count of nodes. In other words, well-defined clustering method on the sequence of positions is required. For construction of cluster nodes from sequence of points, various methods are available including KNN (k-nearest neighbor), SOM (self-organization map), and graph theoretic clustering etc. In this case of person's daily route problem, GIS is useful information and can provide good candidates for cluster node. For example in the map above, the place mark in the center of the map represents park office with public bathrooms. When we find GPS signals stay around the region, we can define the place with marker as a

cluster of the signals (route). Stability of clustering algorithm can be improved.

We just claimed that stability for cluster construction from GPS log data can be enhanced with help of public GIS. In this paper we also claimed that accuracy of cluster construction could be improved by using human activity recognition from motion sensor.

Figure **2** illustrates graph of motion signals. For this study we created a mobile application to capture signals from motion sensors equipped to smart-phone. The figure demonstrates characteristic dependency of magnitude and frequency of graph on type of motions: sit, walk, and run. At a glance, the difference between graphs is quite obvious. Our idea was that combined recognition result of person's motion behavior with candidate place of cluster node would improve accuracy of finding cluster nodes.

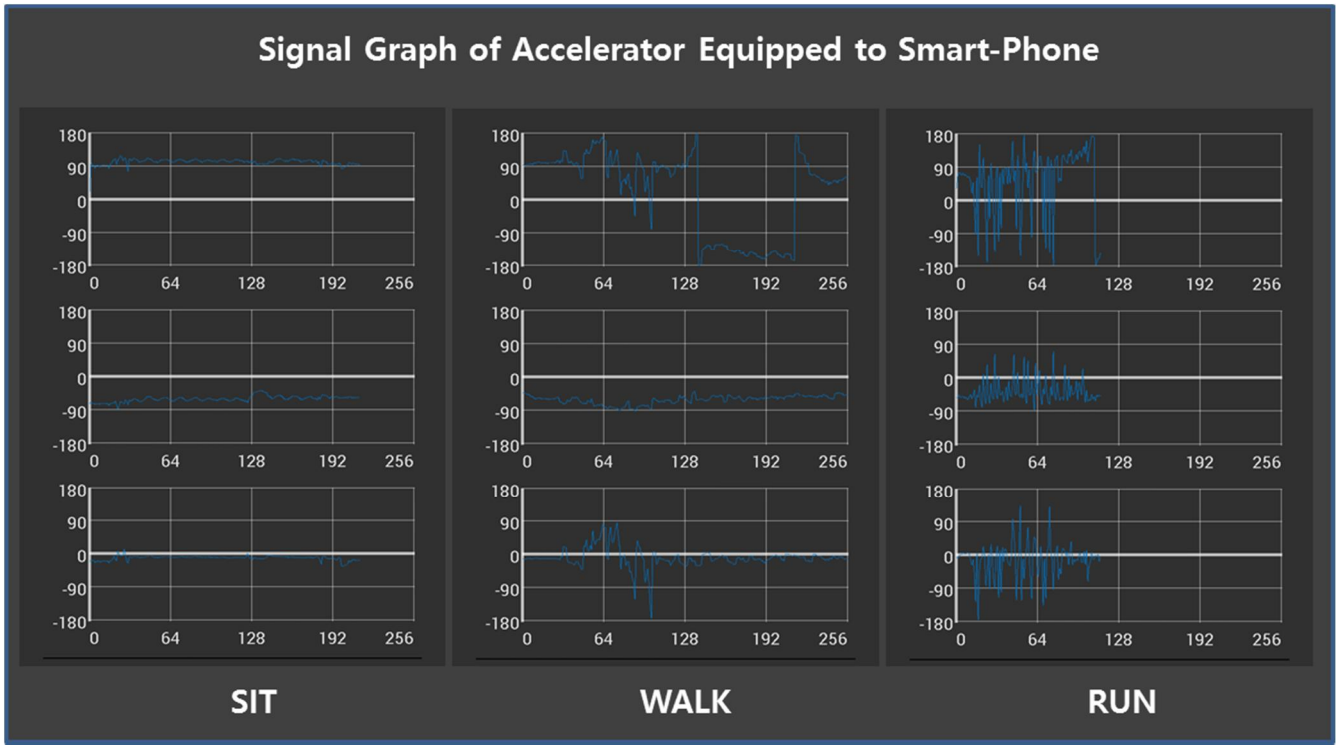


Figure 2. Example of motion sensor signals by behavioral pose. In this demonstrational graph, we captured signal from accelerator. In the graph, x- axis represents progressing time and y- axis does degree of angle.

As we demonstrated applicability of GIS data and motion sensor data for the problem of cluster node construction of daily route graph from GPS log, our new approach is use of both of the data in addition to GPS log data. For clarity of presentation we formulated our idea as follows:

$$M|c : Q \underset{s,x}{\rightarrow} G \tag{1}$$

From the expression (1), Q denotes GPS log data (for one day), i.e., $Q = \{q_i\}$. Each $q_i \in Q$ is a vector of [latitude, longitude, altitude, Time]. G denotes graph $G(N, E)$ of daily route, where N and E denote cluster nodes (visiting places) and edges (paths passed through). The "s" and "x" under the arrow indicate motion sensor and GIS information, respectively, as the confounding variables. The 'c' denotes clustering

method applied and acts as an implicit argument of the map, M, on constructing cluster nodes from GPS log.

Development of a mobile application for understanding a person's daily activity in real time, for example, staying at one of person's places or being in transition between two places in one's daily route, by analyzing signals from person's mobile device is a challenging problem. Our objective in this paper is construction of a real time prediction (machine learning based) model on person's daily route using the sensor data available from smartphone (GPS and motion sensor) and public GIS database. Our question to be addressed here is as the following: will the extra information of GIS and motion sensor data be necessary for the model parameters? Our claim is that GIS can be useful to improve stability of clustering and motion sensor analysis can be helpful to improve accuracy of clustering.

This paper is organized as follows: in section 2, relevant previous research works are briefly introduced, in section 3, our three-stage model is explained in detail, in section 4, simulation results are presented, in section 5, conclusion remarks are added.

2. Related Research

Our study is about a fused technique between GPS and human activity recognition. Since the last decade, geographic information system (GIS) has been accepted as a part of geography and has create a new area named spatial science. GIS related researches has created new level of data in public domain. Place detection and transient mode prediction have been two major contributions in the area of personal daily route problem.

For place detection, Liao et al. proposed machine learning based Relational Markov Networks (RMN) [5]. and Zhou et al. proposed DJ-Cluster method[6]. Khalaf-Allah et al., described a approach based on Bayesian filtering formulation[7]. Our approach is reinforce learning considering that each individual users have different pattern in signal. For unsupervised stage, we applied graph-theoretic clustering, and for supervised stage, MLP.

Regarding transient mode detection, the details of remote motion sensor on Opportunity over the surface of Mars were described in the article [2]. Theodoridis at el. studied pattern recognition techniques on human activity [8]. Zheng et al. proposed machine learning based transportation mode prediction [9][10][11][12]. Ellis et al. presented their long term field works on transportation mode (travel behavior) [13].

On public domain datasets, Reiss et al. published PAMAP2 for physical activity monitoring [14, 15] and Anguity et al. published HAPT for human activity recognition [16]. our approach to transient mode detection is inspired by HAPT and improve it by adopting gait analysis technique.

3. Model Description

Since dynamic and environmental sensors have been equipped to the mobile devices, various results of researches on understanding person's activity by analyzing sensor signals from mobile device have been published. Among the researches, we have focused on two topics: 1) prediction on transient mode and 2) detection on the person's places (location where a person is visiting not just passing by) from GPS track data. If we consider input data adopted by those researches, GPS track is commonly used, accelerometer is also used most of recent researches, GIS data is used most recently in conjunction with person's daily route in order to compare to public transportation route.

Prior to addressing the objective of this study, three kinds of notions need to be defined: place, path, and person's daily route. In this paper, type of person's place (as mentioned above) was adopted from paper [4]: "essential and frequently visit place" "non-essential and frequently visiting place", "essential and infrequently

visiting place", and "non-essential and infrequent visiting place". We summarized at Table 1.

Table 1. Cateogrization of type of places with examples

	Frequent	Infrequent
Essential	**Home, Office**	**Hospital, Mom's house**
Not essential	**Convenient store**	**Diner**

As can be seen at Table 1, each of the four categories from type of places (TYPE_OF_PLACE) was further divided into the sub-categories and labeled with user customized place such as "home", "office", "grocery place", and so on.

Place is an abstract object representing location where a person visits. It will be classified by TYPE_OF_PLACE introduced at the table above. As notable properties, arrival and leaving times, east-north and west-south bound, Google place type (which is about whether the location is exact or approximate), and name (user specified or from GIS, sub-category of TYPE_OF_PLACE) were considered. Table 2 shows definition of place in pseudo language.

Table 2. Object design: place definition

```
@Entity
class Place
    @id Long id
    @kind TYPE_OF_PLACE type
    @prop TIME arrival, leaving
    @prop BOUND EastNorth, WestSouth
    @prop GP_TYPE cover_type
    @prop String name
```

Path is an abstract object representing a person's route between the two places. Table 3 shows definition. It will be classified by TRANSITION_MODE as to be seen in section 4.2. As notable properties, it contains start/end time, start/end places, and name (user specified or from GIS).

Table 3. Object design: path definition

```
@Entity
class Path
    @id Long id
    @kind TRANSITION_MODE transition
    @prop TIME start, end
    @prop Place start, end
    @prop list<GPS> listGps
    @prop String name
```

Route is an abstract object representing a person's daily route. It is a graph of places and paths as described above. Table 4 shows definition. It will be classified by ROUTE_MODE (explained below). As notable properties, date, list of places and paths were used.

Table 4. Object design: route definition

```
@Entity
class Route
    @id Long id
    @kind ROUTE_MODE type
    @prop DATE date
    // graph representation
    @prop list<Place> listNode
    @prop list<Path> listPath
```

On the table above, for the clarity of presentation, we adopted JPA-like notation. @Entity is for scalability guaranteed object, @id for key of the object, @kind for category, @prop for properties to be used for query. For the term TRANSITION_MODE, we adopted HAPT2 project for this study and its related outputs are shown at section 4.2. The term ROUTE_MODE (mentioned in the table) categorizes user's activities in daily basis. In this paper, for the purpose of presentation, we used a simple predefined set such as normal week day, abnormal week day, normal weekend, and abnormal weekend. In addition to the data structures for graph representation of personal daily route, we used knowledge database for knowledge/semantic representation of personal daily route to support inference rules depending on business logic of the application. Table 5 shows an example of schema for our database model.

Table 5. schema of personal knowledge database

Example of predicates	
Predicate	leftHome(), cameHome(), wentWork()
Unary	visitedPlace(A), passedLocation(A)
Binary	visitedPlaceAt(A, T), passedLocationAt(A, T), cameAcrossPerson(A, B)

We present a machine learning based prediction model on person's activity along with one's daily route. The model consists of three-staged mutually independent models: at the stage 1, person's places detection, at the stage 2, prediction on person's transient mode, the stage 3, inference rules with person's place and transient mode. In this paper, 'place' is a specific term for one of GPS location where a person is visiting not passing by. The three staged model is explained throughout the following three subsections as below. The topics of subsections are detection of place from GPS log, classification of transient mode with the signals from mobile sensors, and problem specific inference rules, respectively. As a summary is illustrated.

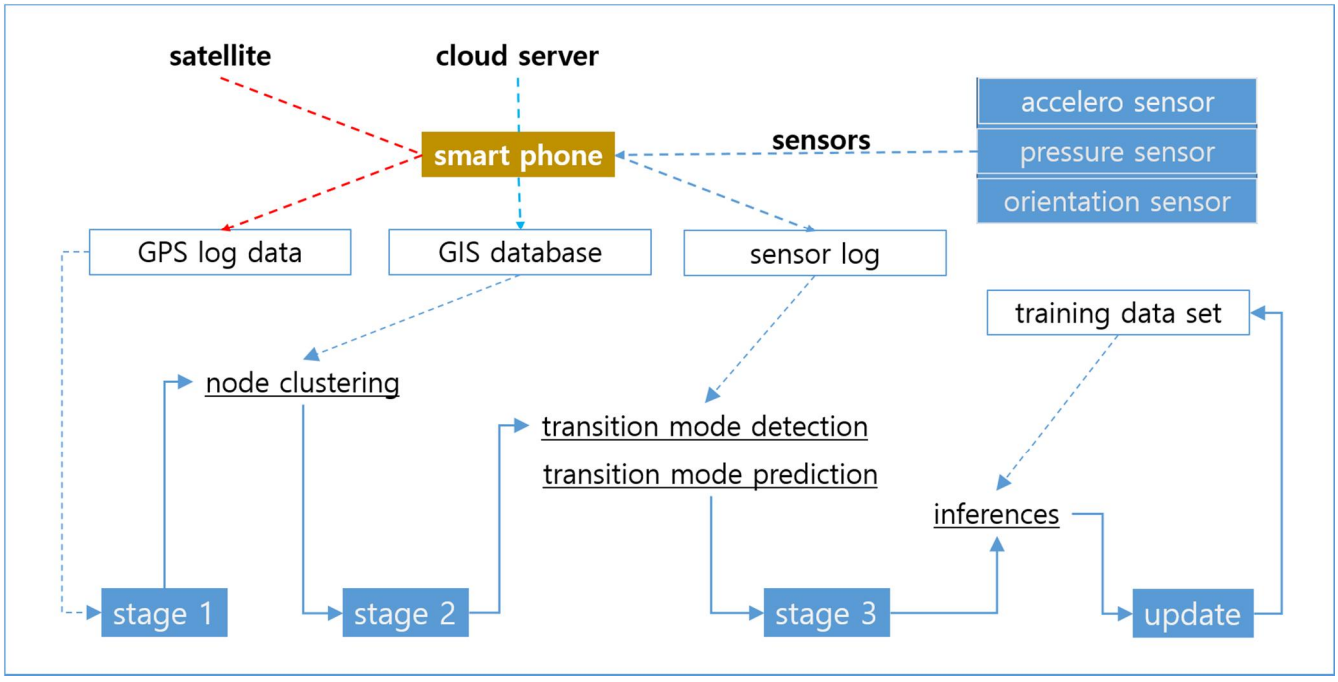


Figure 3. Flow chart of daily route prediction model

3.1 stage 1: 'place' detection from GPS tracks with history of personal daily route and GIS data

Supervised learning is not appropriate for detection of places from GPS track. The problem is that person's visiting place is user dependent. In this paper, we adopted a kind of reinforcement learning. First of all, we build a universal classifier for detecting place on GPS track data using unsupervised clustering algorithms. Use the classifier to identify candidate places from GPS log. Secondly, validate the candidates with GIS data and personal history of places. Thirdly, convert GPS track into graph with node of places.

As mentioned above, the term 'place', in this paper, has special meaning: it is a cluster of locations where a person spends some amount of time, not just passing by. Various methods, from KNN clustering to rule based thresholding with respect to staying time, have been studied. In this paper we present a new approach, two-step method. At the first step, candidates of place are detected using thresholding on staying time. The threshold value was obtained by using graph theoretic clustering. At the second step, filter the candidates out using GIS information and personal history. Details are presented at the next section.

Figure **4** below illustrates clustering GPS track points in terms of time stamp. The red place markers are obtained from public GIS database. Our idea is that we merge the cluster points near a GIS marker into the marker.

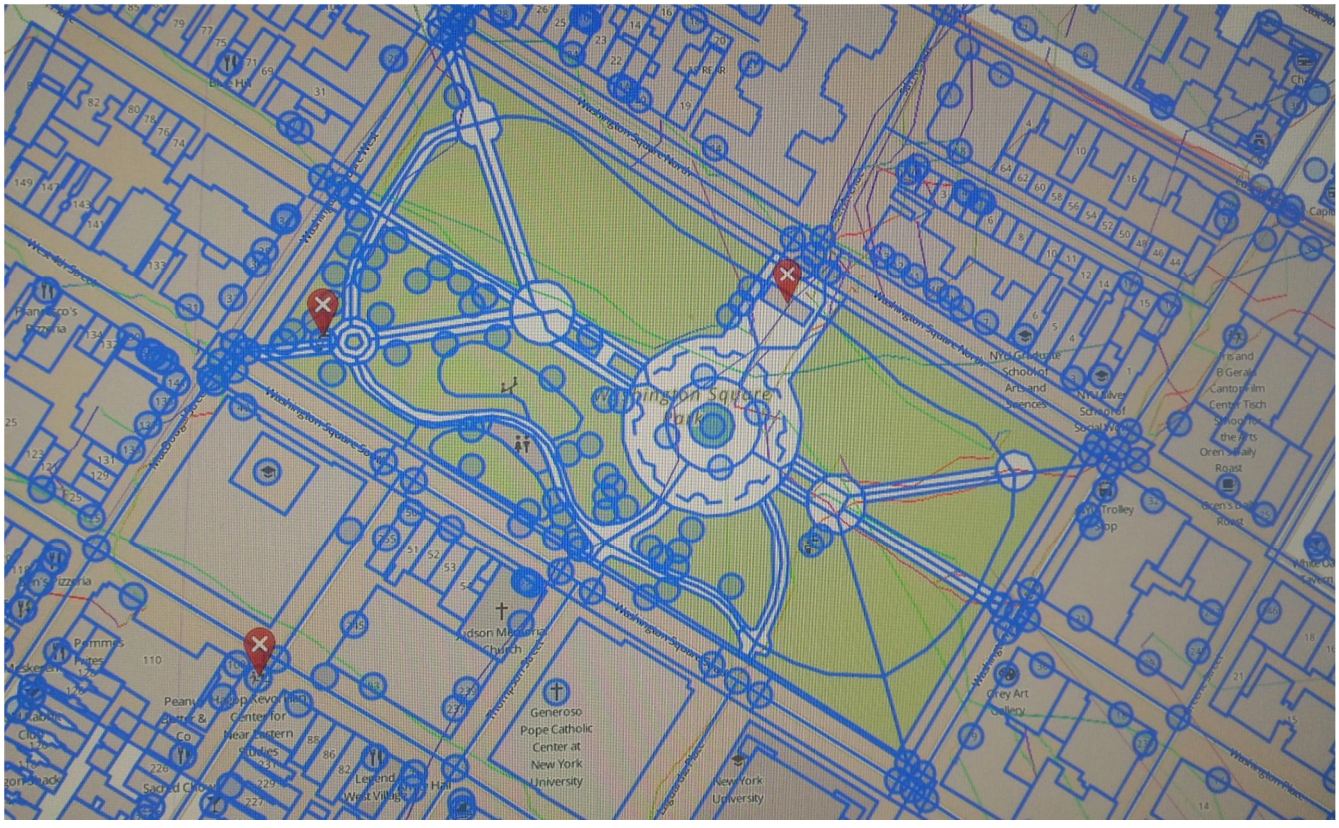


Figure 4. Example of node clustering output. The blue circles are output of clustering algorithm on GPS points. Confer to Figure 1 for original GPS track

3.2 stage 2: machine learning based transient mode prediction from GPS track and sensor database

Problem of determination on person's transient modes is typically approached by supervised machine learning on GPS track and accelerometer signals. In this study, we did not try to improve transient mode detection methods per se, on the other hand, using the places (cluster nodes) constructed at the previous stage, we tried to determine the transient mode on a path between two distinctive places (edges of graph). Similarly to the other studies introduced at the previous section, in this paper, mobile sensor signals (including accelerometer) were mainly used to build the training data set. Timestamped GPS track log was also used to correct the speed estimated by accelerometer. It is clear that a personal GPS log cannot be used as training data due to user dependency of the data, therefore we used it only for measuring relative movement. One of the major differences of our approach with the previous studies is in that, using the graph of places and paths; our method takes advantage of more contextual information on the path. Transient mode would be unchanged on a path between the two places.

Hardware and software imbedded in smartphone device should follow international standards such as ISO and ITU. As of writing this paper (fall, 2016), however, dynamic- and environmental- sensors embedded in smart phone do not have a standard to follow. This can cause a problem when we use raw signal data as is. Scale difference of the raw signals between mobile devices should be normalized.

3.3 stage 3: inference rules on predicting a person's activity

For daily route mode prediction, we applied a first order inference engine with input data of graph, where the graph is consisted of place node and transient mode. Considering a first order logic

$$\text{FOL} := \langle S|V|C|Q \rangle,$$

where S is represented as non-logical symbol, V as variable, C as connectives including boolean connectors, disjunction, conjunction, and implication, Q as quantifiers. Based on the FOL, definition of syntax of FOL using our data as follows: the place identified at the stage above played role of term, and the transition mode identified played role of binary relation. Using the two (place and transition mode), we could construct atomic formula. The collection of the formula build our knowledge database on a person's daily route.

For presentation purpose, suppose that a user's location and transition mode should have been identified through the previous stages. Then we can construct set of logical statements as the following way:

- suppose the current location is X, the transition mode is R (one from the list in Table 7),
- set of two possible places connecting X is S = {(P1, P2), (P3, P4), …} which can be represented as edges, i.e. L = {E1, E2, E3…}.
- Evaluation of logical statement R(X, L) = {R(X, E1), R(X, E2), …} will give us inference rules based on consistency checker model. In other word, R(X, E) will be passed as long as it exists in current database, failed otherwise.
- Failure in inference leads to predict abnormal activity.

In this stage, knowledge database described at the previous section is continuously updated. Depending on the application, this stage plays role of business logic layer. For example, if the application is child protection from abduction, then this stage will check the current location whether it is matched with predetermined personal daily routes. If this application is public traffic flow prediction, then this stage will estimate probability of passing predetermined traffic grid points at a given time.

3.4 update data

Here, today's data record is continuously updated into the existing history data. In the morning every day, with setting a property for date to today, allocation and initialization of a graph object (introduced at the previous section) is performed prior to the beginning of the first stage. Let us denote the graph (route) object by G. At the end of the first stage, if a new place P has been detected then register the place P into the place list in G (G.placeList), set the P into the path E (E.end = P), and create a new path ENew with ENew.start = P. Otherwise (no P detected) no update. At the end of the second stage, the location data used and the transient mode predicted are registered into E (E.listGps) which is an element in the list in G. At the end of the third stage, the decisions induced by the inference rules are recorded depending on business logic.

4. Simulations

4.1 Analysis on PLACE clustering

As briefly mentioned in the introduction, we designed four stages for constructing machine learning based clustering classifier on status of user's daily route with the sensor signals as input data. Detailed description is as follows.

1. perform training data acquisition;
2. perform training error analysis (and cross validation if necessary);
3. build classifier using machine learning package in R;
4. perform clustering with the classifier obtained at the previous stage.

4.2 Analysis on transient mode

For transportation prediction, we selected features from signals of accelerometer (m/s^2), magnetometer, and gyroscope data (rad/s) for each three dimensional component (x-, y-, z-) values. For the simulation, we made use of R with 'rpart' package.

Table 6. R statements

fit <- rpart(V1 ~ ., data = PAMAP, method = 'class', control=rpart.control(minsplit=1000, cp=0.0001), perms=list(split='gini', loss=lmat, shrink=0))

The Figure 5 at below Illustrates measured cross validation error for machine learning on transient mode, for the purpose of compatibility, using a public dataset for physical activity monitoring (PAMAP2) [13, 14].

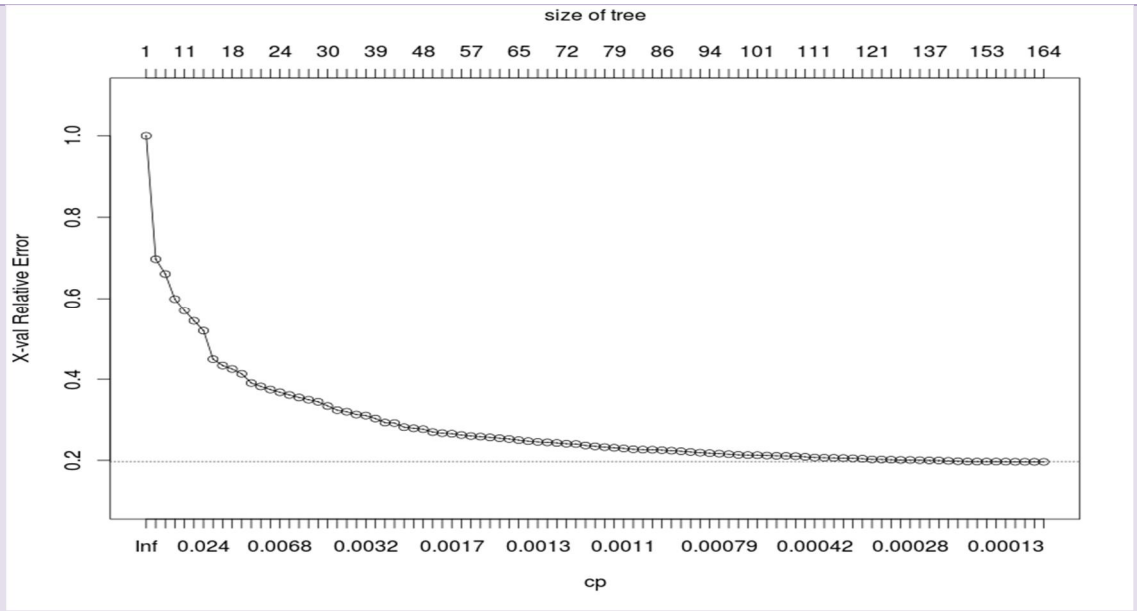


Figure 5 Cross validation for training transportation mode

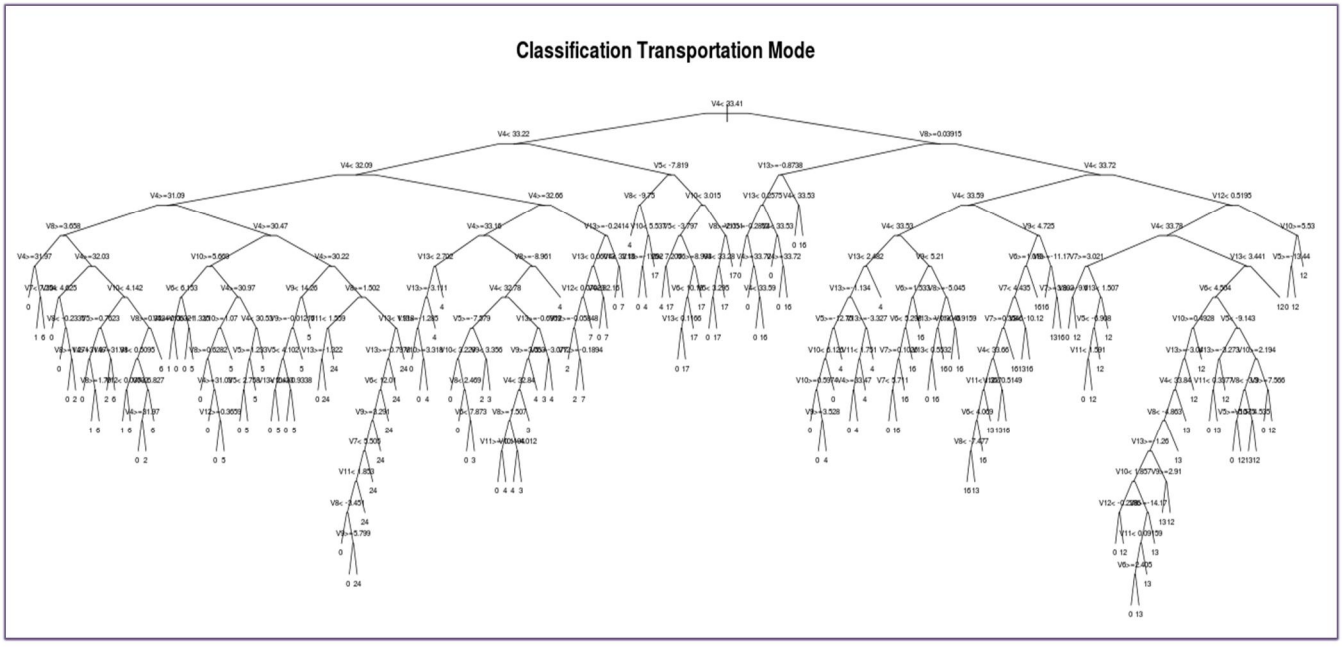
n=374963 (1454 observations deleted due to missingness)

	CP	nsplit	rel error	xerror	xstd
1	0.0389	0	1.0000	1.0000	0.0012

2	0.0365	6	0.6974	0.6974	0.0012
3	0.0310	7	0.6609	0.6609	0.0012
4	0.0292	9	0.5989	0.5990	0.0012
5	0.0255	10	0.5697	0.5698	0.0012
6	0.0245	11	0.5441	0.5443	0.0012
7	0.0235	12	0.5197	0.5198	0.0012
8	0.0156	15	0.4491	0.4493	0.0011
9	0.0111	16	0.4335	0.4338	0.0011
10	0.0109	17	0.4224	0.4252	0.0011
	⋮				
88	0.0002	149	0.1919	0.1971	0.0008
89	0.0002	152	0.1914	0.1970	0.0008
90	0.0001	153	0.1912	0.1970	0.0008
91	0.0001	155	0.1910	0.1969	0.0008
92	0.0001	156	0.1908	0.1966	0.0008
93	0.0001	158	0.1906	0.1966	0.0008
94	0.0001	160	0.1904	0.1965	0.0008
95	0.0001	163	0.1900	0.1964	0.0008

As can be seen in the right panel of the above figure, among 0.37 million counts of data were used and 19% of cross validation error level was observed.

For the purpose of visualization, decision tree for classification obtained from the training is displayed at Figure **6** below. With total 164 counts of leaf nodes, it is too big to draw (top) due to lack of resolution. At the bottom of figure, zoomed-in shot is ready for readability.



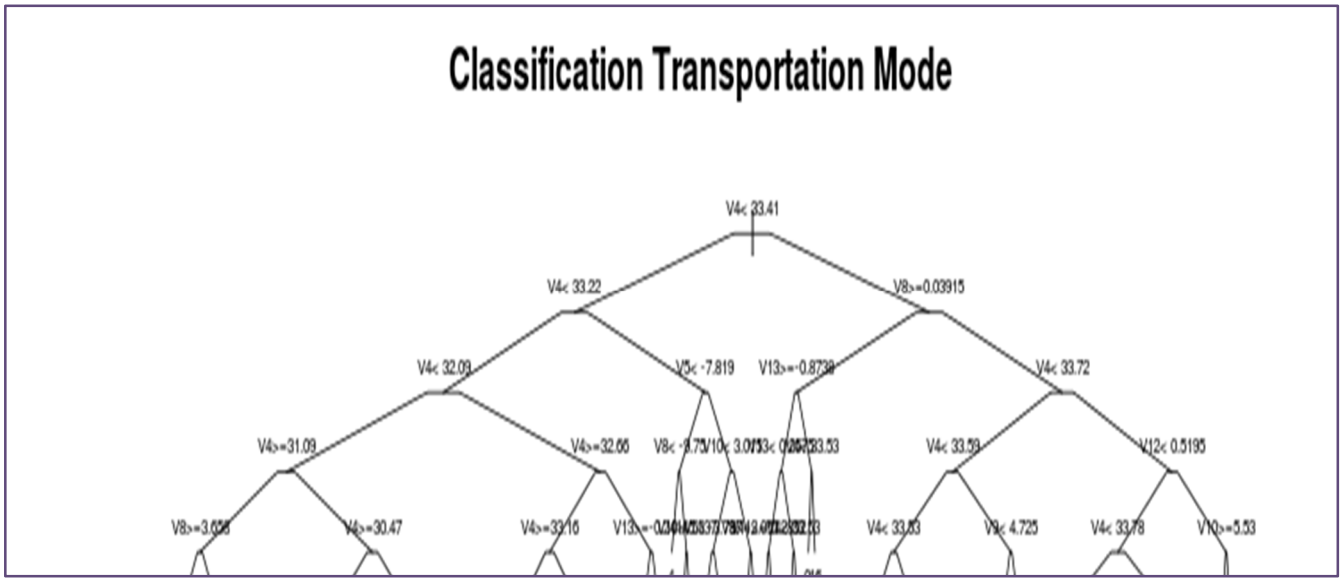


Figure 6. Classification tree for transient mode. The original output is too big to draw (top), part of region was zoomed-in for readability.

PAMAP2 dataset (a public dataset used in this study) has activity ID for label of training data [2]. The list is summarized in Table 7.

Table 7. list of activity ID for PAMAP2 data set

ID	activity	ID	activity	ID	activity
0	other (transient activities)	1	lying	2	sitting
3	standing	4	walking	5	running
6	cycling	7	Nordic walking	9	watching TV
10	computer work	11	car driving	12	ascending stairs
13	descending stairs	14	vacuum cleaning	15	ironing
16	folding laundry	17	house cleaning	18	playing soccer
19	rope jumping				

5. Conclusion

In this paper we presented a three stage model for prediction on person's daily route. As input data, GPS track, signals of mobile sensors are used including accelerometer, magnetometer, and gyroscope. As contextual data, GIS information and incremental personal history were used. The test results performed on each of the modules using public data showed that our design worked well. Real time understanding of user's daily routine has great impact on intelligent system. As our future work, this technique can be used as main engine of a mobile application for child protection from abduction: child's mobile device will detect abnormal behavior in daily routine and will invoke GPS2SMS to alert to one's guardians.

References

[1] Jorge-L. Reyes-Ortiz, Luca Oneto, Albert SamÃ , Xavier Parra, Davide Anguita. “Transition-Aware Human Activity Recognition Using Smartphones”, Neurocomputing. Springer, pp. 754, 2015.

[2] Ricardo Chavarriaga, Hesam Sagha, Alberto Calatroni, Sundaratejaswi Digumarti, Gerhard Tröster, José del R. Millán, Daniel Roggen. "The Opportunity challenge: A benchmark database for on-body sensor-based activity recognition", Pattern Recognition Letters, Vol 34, pp.1780-1788, 2013.

[3] OpenStreetMap. www.openstreetmap.org

[4] Manuel J. A. Eugster and Thomas Schlesinger. “osmar: OpenStrretMap and R”, R Journal Vol 5. pp.1, 2013.

[5] Lin Lao, Dieter Fox, and H. Kauts. “Location based activity recognition”, In Proc. NIPS, 2005.

[6] C. Zhou, N. Bhantnagar, S. Shekhar, and L. Terveen. “Mining Personally Important Places from GPS Tracks”, ICDEW ’07 Proc. 2007 IEEE 23rd ICDEW.

[7] M. Khalaf-Allah. “A Novel GPS-free Method for Mobile Unit Global Positioning in Outdoor Wireless Environments”, Wireless Personal Communications, Vol. 44, No. 3, pp, 311–322, 2008.

[8] T. Theodoridis, A. Agapitos, H. Hu, and S. M. Lucas. “A QA-TSK Fuzzy Model versus Evolutionary Decision Trees Towards Nonlinear Action Pattern Recognition”, IEEE International Conference in Information and Automation (ICIA-2010), pp. 1813-1818, 2010.

[9] Y. Zheng, L. Lie, L. Wang, X. Xie. “Learning Transportation Mode from Raw GPS data for Geographic Applications on the Web”, Proc. WWW .08 of the 17 ICWWW, 2008.

[10] H. Azami, M. Mosavi, S. Sanei. “Classification of GPS Satellites Using Improved Back Propagation Training Algorithms”, Wireless Personal Communications, Vol. 71, No. 2, pp. 789-803, 2013.

[11]Z. Salcic, E. Chan. “Mobile Station Positioning Using GSM Cellular Phone and Artificial Neural Networks” , Wireless Personal Communications Vol. 14, No.3, pp. 235–254, 2000.

[12] D. Wang, M. Fattouche, F. Ghannouchi, D. Wang. “Geometry-Based Doppler Analysis for GPS Receivers”, Wireless Personal Communications , Vol. 68, No. 1, pp. 1–13, 2013.

[13] K. Ellis, S. Godbole, S. Marshall, G. Lanckriet, J. Staudenmayer, and J. Kerr. “Identifying Active Travel Behaviors in Challenging Environments Using GPS, Accelerometers, and Machine Learning Algorithms”, frontiers in Public Health, vol. 2., pp. 1-8, 2014.

[14] A. Reiss and D. Stricker. “Introducing a New Benchmarked Dataset for Activity Monitoring”, The 16th IEEE International Symposium on Wearable Computers (ISWC), 2012.

[15] A. Reiss and D. Stricker. “Creating and Benchmarking a New Dataset for Activity Monitoring”, The 5th Workshop on Affect and Behavior Related Assistance (ABRA), 2012.

[16] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra and Jorge L. Reyes-Ortiz. “A Public Domain Dataset for Human Activity Recognition Using Smartphones”. 21th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2013. Bruges, Belgium, 2013.