Identifying events in mobility data

Branko Kavšek[1,2]
[1]Artificial Intelligence Laboratory
Jožef Stefan Institute
Ljubljana, Slovenia
branko.kavsek@ijs.si
[2]Department of Information Sciences and Technologies
University of Primorska
Koper, Slovenia
branko.kavsek@upr.si

Dunja Mladenić
Artificial Intelligence Laboratory
Jožef Stefan Institute and
Jozef Stefan International Postgraduate School
Ljubljana, Slovenia
dunja.mladenic@ijs.si

Omar Malik
Network Science and Technology Center
Rensselaer Polytechnic Institute
Troy, USA
maliko@rpi.edu

Boleslaw K. Szymanski
Network Science and Technology Center
Rensselaer Polytechnic Institute
Troy, USA
szymab@rpi.edu

ABSTRACT

Today we are used to being interconnected via our smartphones and having our phone location tracked by different apps. ICT technology enables real-time monitoring and processing the user location data from GPS coordinates of a phone. Based on observing the user mobility, Artificial Intelligence methods can be used to improve transportation, proactively provide mobility recommendations and acquire knowledge using the user context. This paper describes the application of machine learning algorithms on user mobility data to identify and understand potentially interesting events. The data for this research was collected from a sample of users consenting to be monitored through our in-house developed smart phone app. A pilot study that includes 227 users that were tracked over a period of 7 years yields fairly positive evaluation results in terms of predictive accuracy of identified events but succeeds in identifying exclusively "well-known" events related to users going to or coming from the office and/or lunch. This shows that machine learning methods can be a suitable choice for identifying events in mobility data but there is still room for improvement.

CCS CONCEPTS

• CCS Information systems Information systems applications Data mining

KEYWORDS

Users mobility, network analysis, event detection, machine learning, clustering.

1 Introduction

Given the data of user mobility, we were looking into using social network analysis and machine learning methods to understand causal templates and identify and predict events in the user mobility data. To this end we have defined an event as an action that is a consequence of some user and/or environment property. For instance, such event is the user driving in the morning if the weather is cold, otherwise the user would be using some other means of transportation. The weather being cold is a cause for the event of driving.

The idea for identifying events is to build a social network of locations that the users are frequently visiting and compare traces of different users to identify typical behaviors. Once we have the traces of typical behaviors, we look for significant diversions in traces and hypothesize that they are consequences of some specific user or environmental context, for instance, from work the user is usually going home but every Tuesday afternoon we observe that the user is going to gym instead not to home. We use machine learning methods to categorize the events based on identified properties of the users/environment correlated with diversions of traces (these properties are seen as potential cause of an event). For instance, on Tuesday afternoons, when the previous location is work, the user frequently uses a bicycle. Then we find regularities in the properties to group the events (and causes). For instance, under specific circumstances some users go from work to gym instead of going home (relevant circumstances here could be that a user likes exercising and the period is Tuesday afternoon).

This paper is organized as follows. Section 2 shortly lists all related research that was done on the same or very similar data to the data used in this research. In Section 3 the data is presented together with the performed pre-processing. Section 4 describes the experimental evaluation with descriptions of the methodology and results. Section 5 provides interpretation and discussion of the experimental results. Section 6 concludes the paper and gives directions for future work.

2 Related research

When referring to user mobility data nowadays, we mostly refer to GPS data provided by the user's smart phone or other wearable device. Sometimes, this data also includes the readings of other sensors, if present and functional (e.g. accelerometer). Tracking a user thus means collecting a series of GPS coordinates readings in a time sequence.

In [6] the authors argue that the raw GPS data is noisy and messy. Thus, when analyzing user paths, they group the GPS coordinates based on time and distance resulting in the detection of the so-called stay-points or locations where a user spent more time. Figure 1 depicts the idea of stay-point detection by clustering in

space and time. The blue points in Figure 1 represent a spatio-temporal cluster of GPS coordinates called a stay-point.

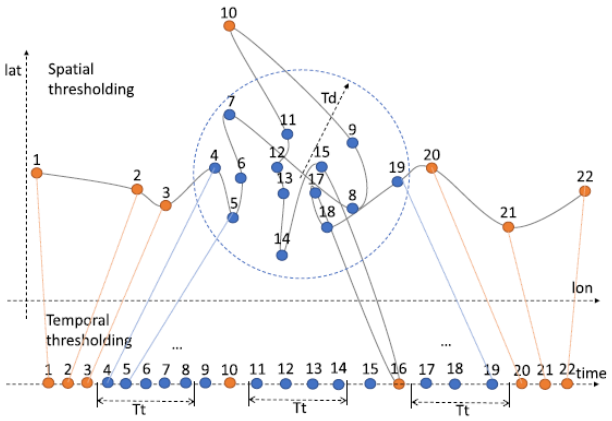

Figure 1: Outlier removal and spatio-temporal clustering [6]

The authors of [7] went a step further and used the data produced by the previous spatio-temporal clustering [6] to try to identify and give a rating to points of interest (PoI) based on users' behavior.

In [2] natural language processing (NLP) methods are used by the authors in combination with previously mentioned methods and crowdsourcing to provide additional user context.

Predicting users' mobility is what the authors of [5] tried to achieve by predicting the next location and mobility pattern of a user using probabilities and the Markov state-space model.

In [9] the authors aimed at detecting the most likely transportation mode of a user and in [1] they tried to motivate the users to make a more ecologically-friendly transportation choices.

Finally, the authors of [8] devised a methodology for visualizing qualitative patterns in multivariate time series that was tested also on user mobility data.

3 Experimental data

Raw data was collected from 432 users in a span of nearly 7 years (from 5.7.2012 until 7.4.2019). The users installed a mobile app that tracked their whereabouts by sending a reading to the database every 30 seconds. Every reading sent to the database included: the "activity ID", "the user ID", "timestamp", "GPS coordinates" (LAT and LON), additional data (accelerometer readings, GPS accuracy readings, …). Not all 432 users were sending data continuously for all 7 years (some users came later or left earlier, some smart phones switched off because of power source issues, sometimes GPS signal was out of range, …).

Because GPS data from users' phones was noisy and messy as argued by the authors of [6], we used their method to preprocess our raw data. The pre-processing steps taken to "clean" our data are described in Section 3.1.

3.1 Pre-processing the data

Data pre-processing was performed in two steps. First, clustering in space and time was applied to a set of uninterrupted 30 seconds GPS readings. Second, any remaining outliers were removed.

3.1.1 Clustering in space and time

This type of clustering is best understood by looking at Figure 1 (taken from [6]). Points, marked with numbers from 1 to 22 and connected with a line in this figure, represent a series of 22 uninterrupted 30 seconds GPS readings from one user. Every point has an associated timestamp and the values for LAT and LON. The clustering is performed using one time and one space threshold. A time threshold of 5 minutes and a space threshold of 120 meters (the threshold values that were actually used throughout our experiments) mean that all GPS readings that fall within a radius of 120 meters for more than 5 minutes will be clustered together to form one stay-point. Start and end times in this stay-point correspond to the first and last GPS readings in the cluster, respectively. A GPS coordinate for this stay-point is the average of LAT and LON values of all GPS readings in the cluster. The non-clustered GPS coordinates represent the so-called paths. In Figure 1 we can notice two paths (1-3 and 20-22) and 1 stay-point (all blue points 4-19).

3.1.2 Outlier removal

When performing the spatio-temporal clustering described in Section 3.1.1, we requested that all the remaining paths must contain at least two GPS coordinates. The GPS coordinates that do not belong neither to a stay-point, nor to a path after clustering, are considered outliers and thus removed.

3.1.3 The pre-processed data

After clustering and outlier removal described in Sections 3.1.1 and 3.1.2, the data contains 235,683 records, of which 114,923 are stay-points and 120,760 are paths. Every stay-point is described by a start time, an end time and a GPS location of its center. The paths, on the other hand, are ordered sets of readings, where each reading has a timestamp and a GPS location. Some paths can contain hundreds of readings, some of them can even be circular (starting and ending in the same GPS location).

Since some of the users that were tracked traveled a lot to all parts on the globe, we decided to simplify things by considering just those stay-points and paths for which all GPS coordinates were inside a rectangle (N 45° - 47° LAT, E 13° - 17° LON) that is limited to Slovenia in the Ljubljana nearby area. This also simplified our dealing with time, as all the data is in the same time zone. We also did not consider daylight-saving times. This reduction leaves our data with 110,072 records from 227 users, of which 58,188 are stay-points and 51,884 are paths.

Since our goal is to identify events in user mobility data, we need additional features describing the data that may later serve as event descriptors. The only two features we have at the moment are "time" and "position" (in space). From "time" we created six new features as follows:

- Time of day,
- Hour,
- Weekday,
- Weekend,

- Season, and
- Holiday.

"Time of day" is a discrete feature with 10 values (see Table 1), "Hour" is just the hour part of the timestamp, "Weekday" is a discrete feature with values MON – SUN, "Weekend" is a binary feature (T if SAT or SUN, F otherwise), "Season" is one of the 4 seasons (Winter, Spring, Summer or Autumn), Holiday is a discrete feature denoting all known Slovenian holidays.

Table 1: Values for the discrete feature "Time of day"

Timestamp	*Value*
6 AM – 8 AM	Early morning
8 AM – 11 AM	Morning
11 AM – 1 PM	Mid-day
1 PM – 3 PM	Early afternoon
3 PM – 5 PM	Afternoon
5 PM – 7 PM	Late afternoon
7 PM – 10 PM	Evening
10 PM – 12 PM	Late evening
12 PM – 4 AM	Night
4 AM – 6 AM	Dawn

From "position" we created just one additional feature, namely "Region" that maps a GPS coordinate to one of the 5 geographic regions in Slovenia (Štajerska-Prekmurje, Dolenjska-Hrvaška, Primorska-Istra, Gorenjsa-Avstrija, Gorica-Italija); a sixth "region" was added for the capital (Ljubljana).

4 Experimental evaluation

In this experiment we decided to additionally simplify things by "reducing" all the 51,884 paths to just the initial and final positions, disregarding all the 30-seconds position readings in-between. By doing so the notions of "path" and "stay-point" lose their meaning, since now we can consider a stay-point as a path whose initial and final positions are the same. Thus, we can drop the "type" (stay-point/path) feature and consider all 110,072 activities from 227 users in the same way.

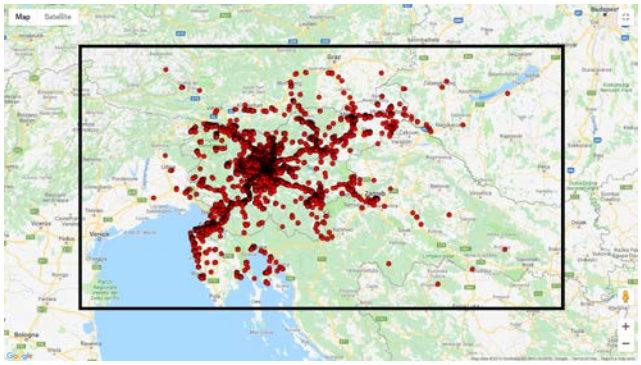

Figure 2: Visualization of experimental data on the map of Slovenia (Google Maps API)

We also decided to round all LAT and LON values to 2 decimal places. This was done since minor fluctuations in the GPS signal were being treated as different locations. By reducing our precision, we smoothed out this noise. The visualization of this data on a map of Slovenia is shown in Figure 2 – the black rectangle represents the observed region.

We now observe the 20 most visited GPS locations. 18 of the 20 most visited locations are all located around or near one of the most popular locations – they are depicted in Figure 3, with the black circle representing the most popular one. For each location we sample all paths that contain this location either at the beginning, the end or on both sides. This generates 20 new datasets. Just the results for the dataset associated with the most frequent location is presented in this paper, since for the other 19 datasets the results are very similar, and this is just the first experiment intended to be more of a proof of concept than a thorough result.

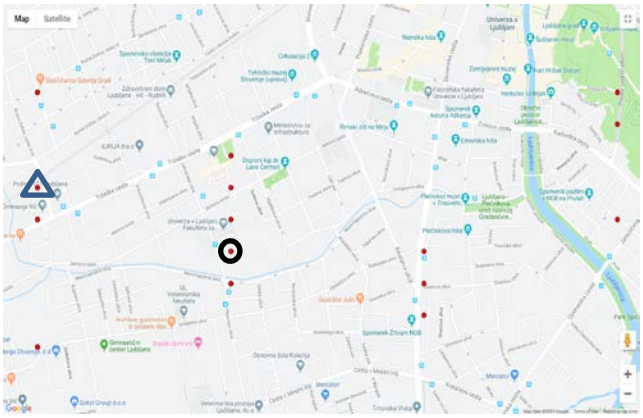

Figure 3: The 18 most popular locations (Google Maps API)

The most frequent location's dataset now contains 17,582 activities (2-point paths). At this point we decide to observe the difference between users that come to the most frequent location, those that leave the most frequent location and those that stay at the location. We create a new *Class* attribute that will serve as our dependent variable for the predictions and assign the values "In" (5,356 examples), "Out" (6,947 examples) and "Stay" (5,225 examples) to it, reflecting the users coming, leaving or staying. So, we end up with a dataset with 17,582 examples, 14 independent attributes – (6 for "time", 1 for "space") x 2 (for start and end point) and a quite balanced class attribute.

4.1 Methodology

For machine learning we used the WEKA workbench [3,4]. The algorithms used were PART (rule learning), J4.8 (decision trees), SMO (SVM), Random Forrest and Naïve Bayes. All the algorithms were ran with the default parameters; the evaluation was performed using 10-fold cross-validation observing classification accuracy as the performance measure. The task we are addressing is supervised learning to build a model for distinguishing between the three types of users that are visiting the most frequent location. In our data the most frequent location turned out to be the Jožef Stefan Institute, which is the working place for most of the users.

4.2 Results

The results of this experiment are presented in Table 2. Classification accuracies are presented as percentages together with standard deviations.

Table 2: Results of selected algorithms on most frequent location's dataset

ML algorithm	*Average accuracy (%) / STD*
Majority class	*39.5 ("Out")*
Naïve Bayes	65.5 / 2.45
J4.8	68.1 / 2.76
PART	68.5 / 2.19
SMO	**71.4 / 1.99**
Random Forrest	70.2 / 2.01

The results in Table 2 show that all five algorithms perform within the 65% to 70% classification accuracy, with SMO having slightly higher accuracy. The majority class value in this case is "Out" appearing in just 39.5% of all the examples.

Not shown in Table 2 is the co-occurrence of certain attribute values with specific class values: "Time of the day = Morning" frequently co-occurs with class value "In" in the generated models; "Time of the day = Mid-day" frequently co-occurs with both class values "In" and "Out"; "Time of the day = Late afternoon" frequently co-occurs with class value "Out". There is a lot of migration between the most frequent location (black circle on Figure 3) and one of the other top 20 frequent locations (blue triangle on Figure 3).

5 Discussion

As the results in Table 2 clearly show, the Support Vector Machine classifier (SMO) has the highest accuracy, but the difference compared to the second best, Random Forrest, is not big. All selected machine learning algorithms clearly outperform the majority classifier, but still with around 70% accuracy, these classifiers cannot be considered good predictors.

The frequent co-occurrence of attribute values with specific classes show the following:

- in the morning people tend to come "In" to the frequent location (they come to work),
- in the late afternoon people tend to go "Out" from the frequent location (they leave the office),
- at mid-day (around noon), both "In" and "Out" links suggest people go for lunch or a snack,
- a lot of migration between the most frequent location and one of the other frequent locations suggests people have some sort of engagement on this other frequent location – indeed it turned out that the other frequent location is in fact the building where a lot of mobility users work in their spin-off companies.

In Figure 3 the "grid effect" of rounding up the GPS coordinates is clearly visible and sometimes the rounded coordinates do not correspond exactly to the physical locations of the points-of-interest.

6 Conclusions and future work

Our pilot study on identifying events in mobility data provided fairly positive experimental evaluation results in terms of predictive accuracy of identified events. However, the events identified are "well-known" events related to the users going to or coming from the office and/or lunch.

On the other hand, over-simplification of the mobility data did not "pay off" in our case, which is clearly visible in the form of the "grid effect" of rounded GPS positions and lack of interesting/surprising relationships in the constructed models.

One possible direction that we are looking at for the future research is to re-run the experiments on the original pre-processed data (described in Section 3) and focus our attention on the changes in user paths.

ACKNOWLEDGMENTS

This work was partially supported by the Slovenian Research Agency and the European Commission RENOIR project H2020-MSCA-RISE-691152. Authors are grateful to Jasna Urbančič for providing help with preparing and interpreting the raw data.

REFERENCES

[1] E. Anagnostopoulou, J. Urbančič, E. Bothos, B. Magoutas, L. Bradeško, J. Schrammel, G. Mentzas. *From mobility patterns to behavioural change: leveraging travel behaviour and personality profiles to nudge for sustainable transportation*. Journal of Intelligent Information Systems, pp. 1-22, 2018.

[2] L. Bradeško, M. Witbrock, J. Starc, Z. Herga, M. Grobelnik, D. Mladenić. *Curious Cat-Mobile, Context-Aware Conversational Crowdsourcing Knowledge Acquisition*. ACM Transactions on Information Systems (TOIS) 35 (4), 33, 2017.

[3] E. Frank, M. A. Hall, I. H. Witten. *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*. Morgan Kaufmann, Fourth Edition, 2016.

[4] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten. *The WEKA Data Mining Software: An Update*. SIGKDD Explorations, Volume 11, Issue 1, 2009.

[5] B. Kažič, J. Rupnik, P. Škraba, L. Bradeško, D. Mladenić. *Predicting Users' Mobility Using Monte Carlo Simulations*. IEEE Access, 2017.

[6] M. Senožetnik, L. Bradeško, B. Kažič, D. Mladenić, T. Šubic. *Spatio-temporal clustering methods*. Conference on Data Mining and Data Warehouses (SiKDD 2016).

[7] M. Senožetnik, L. Bradeško, T. Šubic, Z. Herga, J. Urbančič, P. Škraba, D. Mladenić. *Estimating point-of-interest rating based on visitors geospatial behaviour*. Computer Science and Information Systems 16 (1):131-154, 2019.

[8] L. Stopar, P. Škraba, M. Grobelnik. Streamstory: exploring multivariate time series on multiple scales. IEEE Transaction on Visualization and Computer Graphics, 12(8):1-10, 2018.

[9] J. Urbančič, L. Bradeško, M. Senožetnik. *Near real-time transportation mode detection based on accelerometer readings*. Proceedings of the 19th international multiconference information society (IS 2016).