Reading Exam

Vaibhav Kulkarni

June 6th 2016

Contents

1 Motivation

Mobility prediction is at the heart of my research subject. More specifically, the main goal concerns how to predict future locations of humans from their mobility traces. A wide number of techniques may be used to achieve it. Consequently, to gain a better overview of them, a better understanding of their accuracy and their complexity, I would like to focus the reading exam on these techniques. Since mobility prediction uses geolocated data and that mobility prediction must include security, I have also chosen a paper about security related to geolocated data. This work will be very helpful for my research.

2 Understanding Mobility Based on GPS Data [5]

2.1 Summary

2.1.1 Context

A considerable amount of research studies have focussed on detecting significant locations in the user trajectory data, predicting future locations of a user or recognising user specific activity at a particular location. On the contrary, classifying user GPS trajectories based on transportation modes has not received substantial attention. Knowledge regarding the transportation modes is significant in order to provide pervasive computing systems with meaningful context information.

2.1.2 Problem

To date, the transportation mode classification, relies on manual labelling by the users or utilised GSM radio signals, which could only discriminate between simple motions such as moving and being stationary. On the other hand, identification methods based on trivial approaches such as velocity based classification leads to inherent errors due to the frequency at which users switch the travel modes. The velocity of the travel modes is also influenced by traffic conditions and weather. In this paper the user GPS trajectories are used to understand and distinguish between users transportation mode, such as walking, driving or taking a bus based on supervised learning.

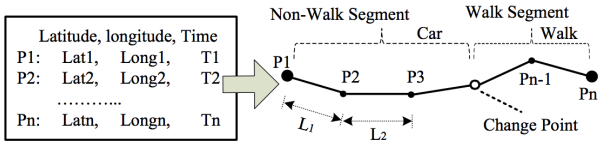

Figure 1: GPS logs and transportation mode prediction (From [5])

2.1.3 Contributions

Results

In order to infer the transportation mode, the authors identify a set of novel features beyond simple velocity and acceleration. These features along with a post processing algorithm are robust to the traffic condition and contain significant information of users motion. The technique was evaluated using GPS logs collected by 65 users over a time period of 10 months. The authors evaluated the proposed approach by considering various combinations of the suggested features and achieved a prediction accuracy of 72.8 %.

Approach

The framework to infer the transportation mode consists of an offline learning stage and a online processing stage. In the offline learning part, the GPS trajectories are partitioned into segments depending on the direction change points. The individual segments are than utilised to extract several features which are used to train a classification model for the online inference stage.

When the GPS trajectory comes, it is first partitioned into segments from which several features such as direction change rate, velocity change rate, stop rate and heading direction are extracted. Using these features, the inference model is utilised to predict

the transportation mode for a given segment in a probabilistic manner. As a classification model, decision tree is used based on a change point based segmentation method. The change points are grouped into several clusters using a density-based clustering algorithm. The derived clusters are used to construct an undirected graph with nodes as individual cluster and edges being the transportation between nodes. A spatial index is built over the graph to improve the efficiency of accessing information of each node and edge. Finally, the probability distribution is calculated of different transportation modes on each edge.

change point based segmentation method, variable traffic conditions, graph based post processing algorithm, independent of any additional database of road networks or POI's. Independent of sensor data such as GSM signal, heart rate, and map information like road networks and bus stops. Therefore can be deployed in a range of web applications. Non user-specific model.

2.2 Discussion

The contribution of the above paper is a methodology to infer the mobility mode, given raw GPS data logs. The non-user specific model presented in the paper is independent of external sensors such as GSM signals, heart rate and map information like road networks and bus stops. The change point based segmentation method devised by the authors is proved to be robust to variable traffic conditions. Thus, it can be deployed in a range of web applications.

- The technique presented, involves an offline training phase which uses the extracted features to generate the online transportation mode inference model. It would have been interesting to quantify the data required and computational time/complexity involved at the offline stage to arrive at a sufficient online prediction accuracy.
- The paper states that the devised approach is a non-user specific model. However, sufficient evaluation of the experimental results regarding generalisation of user specific models is not presented in the paper.
- The authors derive several features from the raw GPS logs to train the inference model. Although, it is evident that a combination of certain features results in a sufficient prediction accuracy, a combination of a particular features reduces the accuracy. The reasoning behind such a behaviour is lacking in the evaluation section.
- Considering majority of the population owning a smart phone have Internet connection, it will be interesting to map the extracted segments to road maps and points of interests in the training phase to increase the prediction accuracy.
- The segmentation technique described, extracts segments from consecutive GPS points. However, there will exist some segments which do not offer any valuable information for feature extraction. It will also be interesting to extract the only segments providing valuable information based on other attributes.

3 Limits of Predictability in Human Mobility[4]

3.1 Summary

3.1.1 Context

Human behaviour is characterised by spontaneity, randomness and change. This paper puts forth an important contribution in stating to what extent human mobility is predictable. The authors identify a metric, "Entropy" as a means to measure the potential predictability in user mobility. Although, the paper might seem trivial, it can be seen as a base research in order to make explicit predictions on user whereabouts.

3.1.2 Problem

The authors take an interesting approach to model human activity which is not stochastic based. There exist several probabilistic based models, as a result it is necessary to know the degree of randomness in human behaviour and the extent to which human mobility is predictable. Furthermore, it is important to know the bounds of predictability characterised by the mobility entropy, travel distance, frequency of visits, time spent at a particular location and heterogeneity of the visitation patterns. This papers fills this void by experimentally analysing the above bounds and concluding that human mobility is characterised by high regularity and this is predictable.

3.1.3 Contributions

Results

- *Prediction accuracy* is the number of correct predictions divided by the number of deliverable predictions (deliverable predictions are those that provide a result, because some techniques cannot provide a result if current pattern occurs the first time);
- *Quantity* is the number of deliverable predictions divided by the number of requested predictions;
- *Stability* is the difference between the minimum and the maximum of the prediction accuracies reached with different parameters related to the prediction method such as time;
- *Learning* is the time taken between the beginning of the training phase of a network or a model and when it can be effectively used;
- *Relearning* is the time taken to learn a new habit;
- *Memory cost* is the minimal number of bits to store the current state of the technique;
- *Computing cost* can be a table look-up, a specific training or another cost;
- *Modelling effort* is the effort made to find variables, parameters and other elements in order to obtain a model or a network;
- *Expendability* indicates if it is possible to use the network or the model with more locations;

- *Time prediction* highlights if the time prediction is integrated or in parallel.

The results obtained show that the most accurate techniques (state predictor followed by Markov predictor) are the ones including the confidence counter. However, if the confidence counter is not taken into account, the most accurate technique is the Elman net. Furthermore, the state predictor is the fastest relearning method. Finally, the results also indicate that it is impossible to highlight a 'one size fits all' technique. Indeed, persons who want to predict future locations in a specific application must define the most important criteria of this particular application and find a technique that matches these criteria.

Approach In this paper, a location prediction model is described as a function with input data (history) and output data (prediction result). The input is composed of a sequence of the past visited locations with the entry time for each of these locations, while the output is the most likely future location and its forecast entry time. For the evaluation of these five techniques, a set of benchmarks called the 'Augsburg Indoor Location Tracking Benchmarks' are used. These benchmarks contain the movements of four persons in an office building divided in two different sets (fall and summer data). The summer data set is used for training while the fall data set for computing the prediction accuracy. Furthermore, this evaluation does not take into account contextual knowledge (the personal schedule of a person for example).

3.2 Discussion

This comparison is a helpful guide aiming at facilitating the choice of prediction methods. Indeed, we can easily see what technique is better according to which criterion, especially as all the criteria are well described with some examples in the paper. In addition, we can talk about several limits and/or improvements.

- The choice of the five techniques is not clearly explained and motivated. The authors only indicate that their comparison is focused on the evaluation of next location methods but they do not really explain why they have chosen these methods. Are they the only existing methods?
- Some of the chosen techniques are not sufficiently detailed. Indeed, some techniques are very complex and it would have been better to describe in detail how they work and how we can implement them.
- In this paper, it is written that the best settings for each technique have not always been used during the evaluation. Consequently, this choice can promote some techniques and disadvantage the others. In addition, the real accuracy might not be properly evaluated.
- As the confidence counter seems to improve the Markov predictor and the state predictor, it could also be included in the dynamic Bayesian network method, in the multi-layer perceptron method as well as in Elman net method in order to see if we observe the same accuracy improvements.

4 Show Me How You Move and I Will Tell You Who You Are [2]

4.1 Summary

4.1.1 Context

Due to the constant increase of applications on smartphones or computers that extract and manipulate geolocated data about users, inference attacks on these geolocated data constitute a serious risk. Indeed, with this type of data and inference techniques, discovering users' behaviours become a rather easy task. For instance, based on the movements of a user, it is possible to learn where she lives, where she works as well as her social network, typically by correlating her movements with those of other users.

4.1.2 Problem

In order to protect geolocated data, there exist sanitization mechanisms, which add uncertainty to the data and remove sensitive aspects. A sanitization process has an impact on the power of a potential adversary (an entity that tries to infer users' information from their geolocated data) and should make its work harder. However, there exist a lot of sanitization mechanisms and it may be hard to make a relevant choice among them.

4.1.3 Contributions

Results In order to help researchers to evaluate various sanitization mechanisms and inference attacks on geolocated data, this paper presents some results of experiments that show the impact of several inference attacks according some sanitization mechanisms. Furthermore, by using a flexible toolkit named GEPETO (GEoPrivacy Enhancing TOolkit), this work demonstrates how this tool is helpful to compare sanitization mechanisms according different inference attacks. The results of this paper highlight that the inference attacks can diverge significantly depending on the sanitization mechanisms. In other words, it shows that sanitization mechanisms do not offer equal protection of the data.

Approach In this paper, three experiments illustrate these results.

- For the first experiment, private data (coordinates) about taxi drivers in San Francisco was loaded in GEPETO in order to highlight critical Points Of Interests (POIs), such as home place or work place, of these drivers by applying a 'Begin and end location finder' heuristic inference attack on these data. For 20 to 90 taxi drivers, their home location was found precisely and checked on a map;
- In the second experiment, three clustering algorithms and the previous technique were compared according two sanitization mechanisms (sampling and perturbation). These four techniques also aim at the discovery of POIs of these taxi drivers from their geolocated data and play the role of a potential adversary. The results show that two clustering algorithms are quite resilient to sampling, while with the perturbation technique (distortion), none of the clustering algorithms performed with a precision greater than 50 % under a distortion of magnitude 400 meters;
- Finally, the last experiment shows that the mobility Markov chain (also used like an inference attack) is a compact and reliable representation of the mobility behaviour

of a user. Although this structure is relatively robust to sanitization mechanisms, some POIs with a lesser density may be loose when the sanitization mechanisms are applied on data. Indeed, the Markov chain algorithm enables to highlight transitions between POIs, found by a clustering algorithm. In addition, a probability is assigned to each transition and corresponds to the probability of moving from one state to another.

For these three experiments, only several sanitization mechanisms and inference attacks were implemented in GEPETO, but obviously it is possible to enrich/extend it with other techniques.

4.2 Discussion

This paper introduces a very helpful and flexible tool for the geolocated data privacy domain. Although GEPETO is an important tool in order to make a choice among several sanitization mechanisms, the overall results and GEPETO present some limits and/or might include the following improvements.

- Time dimension is not taken into account in the implementation of the mobility Markov chain. This addition might be a good improvement in order to increase the quality and the accuracy of the representation of the mobility behaviour of a user.
- Obviously, not all the sanitization mechanisms, known in research literature, are compared in this paper. Only sampling and perturbation mechanisms are presented, but other mechanisms exist such as aggregation, spatial cloaking, mix-zones, as well as swapping and should be implemented.
- GEPETO should allow us to combine several different sanitization mechanisms in order to see if the data is better protected.
- GEPETO could include the possibility to add additional knowledge about users (user's calendar for example) in order to create more sophisticated inference attacks and to test them with GEPETO.
- To finish, the comparison of several users models should be possible with GEPETO in the interest in discovering links between them in order to find their potential social network for example.

5 Context-prediction performance by a dynamic Bayesian network: Emphasis on location prediction in ubiquitous decision support environment [3]

5.1 Summary

5.1.1 Context

The interest in context-aware devices, such as smartphones, is increasing. Since these devices have more and more interactions with their users, they are ubiquitous in our lives. These interactions have been made possible because smartphones react according to the context and because they are able to adapt to changes in context. In addition, such devices can be good at helping people in making timely decisions. For example, when a user receives an alert on her smartphone indicating that an electrical problem just occurred on a particular route of the transportation network, she can decide to leave later.

5.1.2 Problem

Existing context-aware systems are limited by the fact that they cannot provide proactive decision support, but only reactive decision support as in the previous example. In order to obtain such proactive behaviours, context prediction becomes essential. To offer support for proactive decision, the system must be able to provide more useful and personalized information to the user according to her potential future location.

5.1.3 Contributions

Results This paper proposes an inductive approach to predict future user's locations by creating a dynamic Bayesian network model (DBN). This model is compared with three other selected probabilistic prediction methods: General Bayesian Network (GBN), Tree Augmented Naïve Bayesian Network (TAN) and Naïve Bayesian Network (NBN). The models induced by these methods are evaluated with a tenfold cross-validation (a data set is generally divided in 2 sub-sets: one to train a model and the other to validate it). The results demonstrate that the DBN model outperforms all other models in terms of average accuracy with a percentage of 72.67 %. The TAN model is the second best performing prediction model with an accuracy of 69.29 %. As for the GBN and the NBN, they obtain low accuracy compared to the two other models (45.88 % and 55.27 % respectively).

Approach In order to clearly understand the Bayesian models considered in this paper, a brief description of each of them is given below. Bayesian networks are Directed Acyclic Graphs (DAGs), which represent the dependencies between nodes and provide a compact representation of full joint probability distributions. Nodes represent variables, or in other words, occurrences of an event or features of an object. NBN is the simplest Bayesian network where there is only one parent node (root node) of all other nodes (child nodes of the root node). TAN is a NBN with also directional links between child nodes. In a GBN, the parent node can also be a child of some child nodes. Contrary to the three others, DBN takes the time into account, more specifically it contains a sequence of static Bayesian networks where each of them represents the state of a variable at different times. For the evaluation, a set of contextual data from 336 undergraduate students has been aggregated and used. Students had to record their daily routines over a period of two

days on campus. For the GBN, TAN and NBN model induction, Weka machine learning has been used to create three location prediction models. During this automatic learning step, similar to a process of discovery knowledge, different variables have been highlighted: the user's previous action, the user's current action, the user's location and route. Then, in order to induce the new Bayesian approach of the paper (the DBN model) four steps have been applied:

1. Identify domain variables;
2. Examine dependencies between domain variables and how they change over time;
3. Describe how the conditional probability distributions are constructed from the user's action and location data;
4. Develop procedurally the belief update in order to use it for propagating beliefs through the DBN.

5.2 Discussion

This paper provides a good overview of the different Bayesian network methods for context prediction: from the simplest to the more sophisticated. However this work has several limits and/or might add some potential improvements or extensions.

- Since students have different habits and perhaps follow different class schedules, the results of this paper may not be accurate because recorded data may not be homogeneous. The authors could use students data of a same class in order to see the differences with the current results.
- The paper uses a cross-validation. The other possible evaluation approach would have been to use an application for the students who had participated to the research in order to evaluate the location prediction models created. This application would have notified the students with a message containing their next possible location and they would have had to answer if the notice is correct or not.
- Finally, the last important limit is related to the previous and concerns how we can really integrate these location prediction models in a real application. In addition, with this implementation, we should be able to see how react the models if changes occur in the students' life.

6 Where to go from here? Mobility prediction from instantaneous information [1]

6.1 Summary

6.1.1 Context

The interest in studying human mobility is increasing. Currently, a lot of applications are able to collect human locations. These locations reflect people lifestyle, their tastes as well as their behaviour. Therefore, the value of these data collected is increasing. In addition, all these locations can be very useful in order to predict future locations of a human.

6.1.2 Problem

There exist an important number of location predictors. However, all of them have not the same prediction accuracy and do not necessarily take into account the frequent changes in human's life such as home or work changes. Consequently, it is obvious that an analysis must be done about these predictors in order to reveal the best performing ones and to find a way to increase this performance.

6.1.3 Contributions

Results This paper examines a wide set of predictors and highlights the most accurate among them (Gradient Boosted Decision Trees with a percentage of 52.55 %). In addition it reveals a complex blending strategy that enables to improve prediction accuracy of 4 % compared to the most accurate predictor.

Approach The work focuses on comparing three families of predictors in order to predict the next place of a human with instantaneous information only. These three families are based on graphical models, neural networks and decision trees. It is also important to note that this work is the result of the participation to the Nokia mobile data challenge, which consisted in responding to the following challenge: *'predict the next destination of a user given the current context, by building user-specific models that learn from their mobility history, and then applying these models to the current context to predict where the users go next'*. The mobility traces have been collected by the organizers and extracted from the smartphones of 80 users over periods of time ranging from a few weeks to almost two years. The research presented in this paper won the challenge. The first result consists in the comparison of several predictors including one tailored model named Dynamical Bayesian Network (DBN) and two generic algorithms called Artificial Neural Network (ANN) and Gradient Boosted Decision Trees (GBDTs). These predictors are enhanced with an aging algorithm in order to adapt them quickly if changes appear in a user's life. In addition, these methods are compared with two baseline predictors: most visited and first order Markov chain. The predictors have been trained with the two first sets of data and evaluate on the third. The results show that the most accurate predictor is GBDT with a percentage of 52.55 %. But the two others are very close (52.12 % for DBN and 51.43 % for ANN). The two baseline predictors offer low accuracies (35.21 % for the most visited and 44.37 % for the first order Markov chain). The accuracy of each predictor varies a lot according to user data. This variation depends on the fact that the quality of each user data set is not equal nor homogeneous. Due to this high variability and to take advantage of it, a combination of predictors has also been created and tested

using different blending strategies. Blending consists in the creation of a new predictor by combining others. More specifically, the new one should be more accurate than any of the individual ones. The second result demonstrates that an accuracy gain of 4 % has been achieved compared to the 52.55 % of the GBDT. These accuracies have been measured on the third set of data and validated with the fourth set that was undisclosed and revealed by the organizers of the challenge at the end in order to choose the winner. Thus, the best strategy found is the following: the ten best predictors of each family have been selected in order to create a subset. Then, the final or new predictor is a mixture of this subset weighted by their performance on the second data set (computed during the training phase).

6.2 Discussion

The work of this paper is valuable because it shows the creation of a new predictor taking into account the performance of other predictors. However, this paper presents some limits and/or might take into account the following improvements:

- Although the results underscore that a blending strategy seems a good option if we want to increase the prediction accuracy, the paper do not describe in detail how we can concretely implement a blending predictor. It would have been very useful to see how to achieve this goal with the description of the algorithm (for the best strategy obviously).
- The authors chose to compare the selected predictors with the first order Markov chain predictor and obtained a low result for this predictor. However, there are other implementations or extensions of this Markov chain and it has clearly been revealed, in other research papers, that some extensions have better accuracy results (between 70 and 95 % at most) than the simple first order Markov chain used in this paper. It would have been interesting to implement some better extensions of this Markov chain predictor in order to see which results would be obtained.
- Finally, the authors do not give any explanation concerning the choice of the 10 best predictors for the best strategy. Why have they chosen this number? Can we obtain the same accuracy result with 15 or 5 best predictors?

References

[1] V. Etter, M. Kafsi, E. Kazemi, M. Grossglauser, and P. Thiran. Where to go from here? Mobility prediction from instantaneous information. *Pervasive and Mobile Computing*, July 2013.

[2] S. Gambs, M.-O. Killijian, and M. Núòez del Prado Cortez. Show me how you move and i will tell you who you are. *Trans. Data Privacy*, 4(2):103–126, Aug. 2011.

[3] S. Lee and K. C. Lee. Context-prediction performance by a dynamic bayesian network: Emphasis on location prediction in ubiquitous decision support environment. *Expert Syst. Appl.*, 39(5):4908–4914, 2012.

[4] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 2010.

[5] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma. Understanding mobility based on gps data. UbiComp '08. ACM, 2008.