# 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.

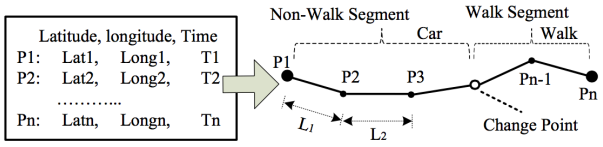


Figure 1: GPS logs and transportation mode prediction (From [5])

### 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

The paper dismisses many of the common assumptions associated with human mobility prediction by experimental evaluation of several mobility datasets. Here lies the major contribution of the paper as it established bounds for various aspects of prediction.

- The article presents a technique to measure the entropy associated with user movement. The entropies can be classified in to random entropy (number of distinct locations visited by the user), temporal-uncorrelated entropy (characterises the heterogeneity of visitation patterns) and the actual entropy (accounts the order in which the nodes are visited). This distinction aids to arrive at the conclusion that user movement can be predicted irrespective of the entropy thus dismissing the general assumption that only lower entropy implies higher predictability.
- The authors evaluate the Fano's inequality bound on predictability when a user with a given entropy moves between N locations. It was discovered that, despite of the apparent randomness of the individual trajectories there exists a high degree of potential predictability in user movement.
- The analysis also led to a conclusion that, the predictability across a large user base is insignificant and varies from person to person. Further, it was also found that, users covering larger distances on regular basis are just as predictable as users commuting in a small area.
- Similar results were obtained when experiments were performed on diverse demographics of varying ages and genders, i.e. only insignificant variations were found in regularity. This concludes that regularity and thus predictability is not imposed by demographic factors, but instead by intrinsic human activities.
- The combination of the empirically determined user entropy by the authors and Fano's inequality leads to a potential 93% average predictability in user mobility.

**Approach** A dataset representing the call patterns of 10 million mobile phone users was used containing the routing tower location. The data was filtered to have only the users with a sufficient calling frequency and high movement which was further characterised individual call/motion activity. This data was processed to construct a time series for each user to determine their entropy, movement regularity and dependence on the demographic and population density.

## 3.2 Discussion

- The used dataset was collected in a high income country where having a 93% potential predictability in user mobility is acceptable despite very large differences in travel distances due to the regularity and availability of transportation modes. However, it will be interesting to characterise similar parameters in the low income and densely populated countries which face much more extreme conditions so as to generalise the findings.
- Although the authors calculated the bounds on predictability by defining entropy and movement regularity, the paper did not show how close to the maximum potential predictability, the accuracy of actual algorithms can come in practice.
-

# 4 Privacy-Preserving Deep Learning [3]

## 4.1 Summary

### 4.1.1 Context

Today, commercial companies such as Facebook, Google and Apple collect a large amount of user data to learn about user preferences and suggest recommendations. This requires a large amount of data related to the users, a considerable part of which is highly sensitive personal information. This data is used to formulate user specific models, generation of which is aided by techniques such as deep learning. On the other hand, biomedical and clinical researchers cannot attain benefits from these techniques as not permitted to share their data. As a result privacy and confidentiality restrictions reduce the utility.

### 4.1.2 Problem

Deep learning presents a interesting avenue to extract highly accurate models by deriving complex features from high dimensional data. Although, such accurate models can give rise to high utility applications they present serious privacy issues due to user data stored in the servers, which is also used for monetary gains by these companies. As a result, there is a need to alter these model generation techniques to offer a satisfactory point in the utility/ privacy tradeoff space.

### 4.1.3 Contributions

The authors device a distributed deep learning technique that collaborates with multiple participants to learn a neural-network model on their own inputs, without explicitly sharing these inputs. The key contributions can be summarised as follows:

- A selective parameter update model: During training iterations, some attributes contribute largely towards a neural networks objective function as compared to others. This model selects parameters whose current value is far away from the local optima. Only these parameters are updated collaboratively to undergo bigger changes in subsequent iterations.
- Distributed collaborative learning: After every iteration of local training, participants asynchronously share the computed gradients with each other. Therefore benefiting from each others training data without actually sharing it. Thus preserving the privacy of the participants.

**Approach** The system architecture consists of a local training database for performing the training locally and parameter server which can be an actual server or a distributed system for running the parameter and gradient exchange protocol.

- Each participant maintains a local vector of neural network parameters. A training iteration is performed over the local training data. Next, the participant downloads the parameters uploaded at the server and computes the gradient of each parameter.
- The parameter server initialises the parameter vector and handles the participants upload and download requests.

- Distributed Stochastic Gradient Descent: While training alone, every participant is more likely to converge to a local optima. However using a distributed approach in learning, by using parameters trained on different datasets, can help to escape local optima resulting in more accurate models.

- Parameter exchange protocol: The authors assume round robin to run the gradient decent sequentially. Every participant downloads the most updated parameters from the server, runs local training and uploads selected gradients and the next participant follows in the fixed order.

## 4.2 Results and Discussion

The authors implemented distributed gradient decent with round robin, random order, and asynchronous parameter exchange protocols. The results were compared with two scenarios, running a gradient descent on the entire dataset and the other is standalone gradient descent where participants train only on their own training data without collaboration. The evaluation was performed on two major datasets used as benchmarks in deep-learning. The key results can be summarised as below:

- The authors achieved the same accuracy as simple stochastic gradient decent as compared to gradient descent ran by sharing only a small fraction of gradients at each iteration step.

- The results obtained show that even at sharing only 1 percent of parameters, results in higher accuracy than standalone or centralised learning.

- Distributed gradient descent using round robin parameter resulted in the highest accuracy and was discovered that round robin protocol is suitable for scenarios where all participants have similar computational capacity.

- It was also observed that number of participants has a lower impact on accuracy than the percentage of shared parameters. Assuming each participant shares his largest gradients with other participants.

- The system computes the differential privacy of each parameter and than decides which one to share with other participants resulting in lower privacy leakage. Further, the authors find the sensitivity index for each parameter, quantifying the amount of random noise which needs to be added to achieve a certain level of differential privacy.

The experiments conducted in the paper are based on supervised learning, it will be interesting to evaluate the accuracy of distributed stochastic gradient descent on unsupervised learning approaches.

# 5 Hiding Stars with Fireworks: Location Privacy through Camouflage [2]

## 5.1 Summary

### 5.1.1 Context

There has been a rapid proliferation of Location Based Services (LBS) in recent years due to ubiquitous wireless connectivity and GPS modules integrated with smart phones. These LBS rely on accurate, continuous and realtime streaming of location data. However, revealing this information to service providers poses a significant privacy risk. In this paper, the authors device a method to preserve user privacy without trading on the services offered by the LBS.

### 5.1.2 Problem

Existing research on user privacy protection in LBS takes the approach of obscuring user's path, compromising the accuracy of services offered by the LBS. Hiding parts of user's paths can lead to degrading the the spatial accuracy, increased delay in reporting user location or temporarily preventing the user from reporting locations completely. This leads to user data being less useful after enabling privacy protection. As a result a framework is needed which can protect the user against location tracking by the service providers at the same time offer high quality services.

### 5.1.3 Contributions

**Results**

**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] J. Petzold, F. Bagci, W. Trumler, and T. Ungerer. In *Euro-Par*, volume 4128, 2006.

[3] R. Shokri and V. Shmatikov. Privacy-preserving deep learning. CCS '15, 2015.

[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.