Understanding the Impact of Data Sparsity and Duration for Location Prediction Applications

Alasdair Thomason, Matthew Leeke, and Nathan Griffiths

University of Warwick, Coventry,
United Kingdom, CV4 7AL
{ali,matt,nathan}@dcs.warwick.ac.uk

Abstract. As mobile devices capable of sensing location have become pervasive, the collection and transmission of location data has become commonplace, enabling the creation of models of behaviour that support location prediction. With such devices often heavily resource-constrained, the nature of data used in location prediction must be understood in order to optimise storage and processing requirements. This paper specifically explores data sparsity and collection duration. The results presented provide insight which suggest: (i) a relationship of diminishing returns in predictive accuracy when collecting user location data at increased rates over a fixed period, and (ii) the duration over which a fixed size sample of location data is collected has a greater impact on predicative accuracy than data sparsity.

Key words: Collection; Data; Duration; Location Prediction; Sparsity

1 Introduction

Location-aware devices are routinely used by a significant proportion of the global population [8]. Data pertaining to a user's location can now be sensed, stored and shared in real-time through devices such as smartphones and tablet computers. Many applications might benefit from location prediction, including city planning, law enforcement, marketing, etc., however relatively little is understood regarding the necessary quality of data for forecasting. Much existing work assumes that location data can be stored indefinitely and at the highest rate afforded by a collection method [3, 16]. These assumptions are inconsistent with the devices typically used to perform location analysis, which are generally battery-powered portable devices carried by an individual with limited storage and memory capability. As a result, data collectors must be able to justify the resolution and duration of collection mechanisms to users.

In this paper we consider two dimensions of data quality for location prediction. Specifically, we investigate data sparsity and collection duration, with a view to informing the design of data collection mechanisms and addressing user privacy concerns. Through the application of three established techniques in location prediction to data varying in sparsity and collection duration, it is shown that: (i) there is a relationship of diminishing returns in predictive accuracy when collecting user location data at increased rates over a fixed period, and (ii) the duration over which a fixed size sample of location data is collected has a greater impact on predicative accuracy than data sparsity.

2 Related Work

Location prediction is widely recognised as being beneficial to providing location-aware services. Early work in this area considered the problem of predicting future locations within small, enclosed environments with a fixed number of discrete user locations, typically employing neural networks, Markov models or dynamic Bayesian networks [2, 10, 13]. Solutions to this problem have applications within offices, homes and public buildings but do not lend themselves to location prediction in large uncontrolled environments.

Motivated by applications such as cell tower handover — seamlessly passing a connection from one cell tower to another when a device is moving — research has considered location prediction in more open environments [7, 17]. Ashbrook and Starner investigated the use of algorithms to extract a user's 'significant locations' from GPS data, and using these locations as the basis for the development of Markov models for location prediction [3]. Similar investigations have been conducted on GPS traces [14, 15], online check-in data [9], and discrete real-world locations such as cell towers [4, 6, 16]. In contrast to the variety of location prediction approaches, existing work has generally considered near-continuous data collected over long time periods, an assumption explored in this paper.

3 Modelling the Location Prediction Problem

We characterise location data as a set of n-tuples, called *points*, containing location and time values, where a *location* is an identifier given to a distinct geographic area on the surface of the earth. The dataset, D_u, of a user, u, is therefore the set of all points associated with u, having inherent sparsity and duration.

$$D_u = \{x_{1,u}, x_{2,u}, ..., x_{n,u}\}$$

The mapping between a time range and set of visited locations for a user can be represented by unknown function, f_u, such that $f_u([start : end]) = S$, where S is a non-empty, potentially large, set of locations visited by the user during that period. It is the aim of location prediction to construct an approximation of the unknown function, $\hat{f}_u$, given a training set $TR_u \subseteq D_u$, such that $\forall y \in TR_u : time(y) \prec start$, which ensures that predictions are in the future.

3.1 Evaluation Model

The function $\hat{f}_u$ can then be used to produce a set of estimated locations, $\hat{S}$, for a specified time range, known as an evaluation window. This set can then be compared against the known set of visited locations for the same window, S. True Positives (TP), False Positives (FP) and False Negatives (FN) are intuitively defined as $S \cap \hat{S}$, $\hat{S} - S$ and $S - \hat{S}$ respectively. We define the set of True Negatives (TN) as $loc(TR_u) - (S \cup \hat{S})$, where $loc(TR_u)$ is the set of locations that exist within a user's training data. We can now define *accuracy* as:

$$ACC = \frac{|TP| + |TN|}{|TP| + |FP| + |TN| + |FN|}$$

4 Experimental Setup

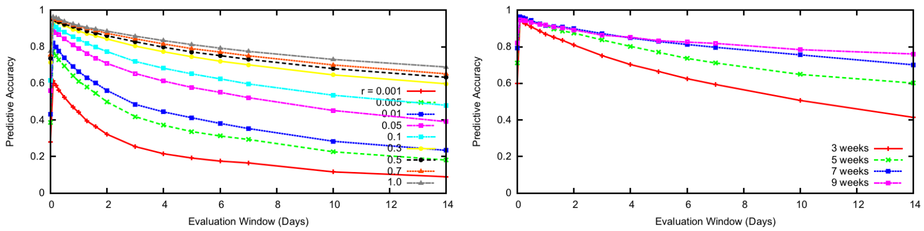
Data was collected by installing a bespoke smartphone application on mobile phones belonging to 5 members of the Department of Computer Science at the University of Warwick. Users ran the application for a period of several months, which recorded the time, latitude and longitude of the device every minute. We generated datasets of different sparsities and durations by selecting a random continuous subset of length n weeks from each collected dataset, and then sampling the truncated data according to a retention probability, r, where each point within the dataset had probability r of being included. Although our approach is limited to using data from only 5 users, this represents an improvement over existing work. The collection of such data is challenging, and existing approaches typically rely on data collected from a single individual [3, 5], or on artificial simulated data [4, 7, 16, 17].

Since location clustering remains an open problem, cell tower regions were used to discretise locations for prediction. There is no loss of generality with regard to the defined data model, since cell tower regions can be considered arbitrary geographical regions designed to maximise coverage.

4.1 Location Prediction Techniques

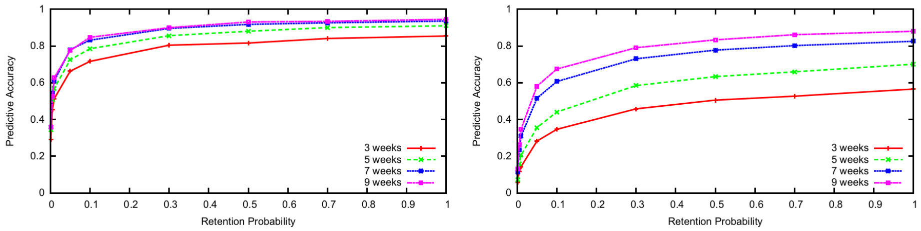
Formalising location prediction as a classification problem allows machine learning techniques to produce predictions for locations given a specific time. A training set of instances, in this case a set of points, is used to represent attributes of the user's current location. Each instance in the space, x_i, has a single classification, $f(x_i)$, where this classification is the location visited. A classifier is able to generate a prediction for a single instance of time, rather than for a time range. To obtain results for a time range, classifiers can be provided with test instances for every time step, in this case each minute, throughout the test range and the results merged to form a set spanning the evaluation window.

Several classification techniques have been shown to be effective for the problem of location prediction, including neural networks [10, 13], decision trees [1, 12] and support vector machines [11, 12]. To ensure the results presented are representative, each of these techniques is used in this paper. Experiments were performed for each user using 4 different durations and 9 sparsities, with each repeated 50 times.



(a) Split on retention probability (b) Split on training length

Fig. 1: Predictive accuracy against evaluation window averaged across classifiers



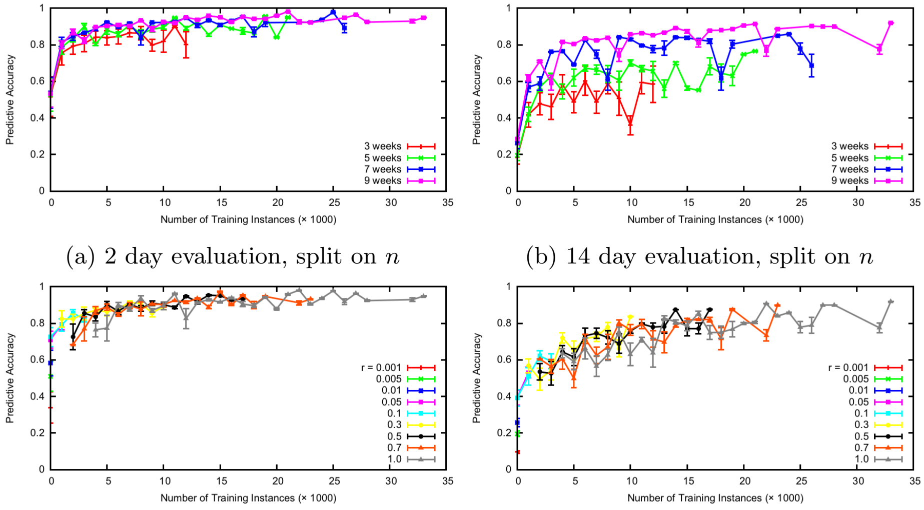
(a) 2 day evaluation window (b) 14 day evaluation window

Fig. 2: Retention probability against predictive accuracy for different durations

5 Results

Each classifier performed similarly across all experiments, and so for brevity we present results averaged across classifiers. Figure 1 shows how predicative accuracy varies with evaluation window length. With an evaluation window of 0 days, the resulting predictions are for a single instance, meaning that $|\hat{S}| = |S| = 1$. As the evaluation window is increased to approximately 1 hour, prediction accuracy increases. This is because any error in a set of predictions made on an individual visiting a small number of locations, especially a single location, is likely to negatively skew predictive accuracy. As the set of visited locations increases, any single error has a reduced impact. Despite this, predictive accuracy declines as the evaluation window is increased further, likely due to the inherent complexity of human mobility. This finding can be used to inform the design of location-aware services, not least because the selection of an appropriate evaluation window can impact the utility of the service.

We now consider how sparsity and duration impact the performance of prediction techniques. Figure 2 shows how predicative accuracy changes with levels of sparsity for the different training durations. In particular, Figures 2a and 2b show these results for evaluation windows of 2 and 14 days respectively. It can be seen that an increase in the proportion of location data, i.e., a reduction in sparsity, consistently yields increased predictive accuracy, although the increase is non-linear. This change in growth rate is significant, since it demonstrates



(a) 2 day evaluation, split on n

(b) 14 day evaluation, split on n

(c) 2 day evaluation, split on r

(d) 14 day evaluation, split on r

Fig. 3: Training instance count against predictive accuracy

that the increase in quality of service afforded is not necessarily linear with the amount of location data collected for prediction.

Figure 3 shows the relationship between number of training instances and predictive accuracy. The number of training instances is a function of both sparsity and duration — the same number of instances can be generated from a short duration with low sparsity or a longer duration at higher sparsity. In order to investigate the interplay between sparsity and duration, each graph shows the result of dividing the range of instance values $(0 - 35000)$ uniformly into groups of 1000, with data points falling into each grouping being averaged.

It can be observed from Figures 3a and 3b that there is a marked difference in terms of predictive accuracy when drawing on training instances of longer duration (and therefore higher sparsity). With a fixed number of training instances, those drawn from a longer duration perform nearly uniformally better than those from a shorter duration. This finding is reinforced by Figures 3c and 3d which show a less pronounced relationship between predictive accuracy and number of training instances when split on different retention probabilities. This substantiates the finding that the duration over which location data is collected is at least as, if not more, important to predictive accuracy than sparsity.

6 Conclusion

This paper has explored the impact of sparsity and duration on the accuracy of location prediction, with a view to informing the design of location data collection mechanisms. Our analysis is based on data collected from 5 individuals, which, although limited, improves on previous approaches that use a single individual's data [3, 5], or on artificial simulated data [4, 7, 16, 17].

In particular, we have demonstrated the performance of established location prediction techniques under general purpose models of data, prediction and evaluation. These results provide insight which suggests: (i) a relationship of diminishing returns in predictive accuracy when collecting user location data at increased rates over a fixed period, and (ii) the duration over which a fixed size sample of location data is collected has a greater impact on predicative accuracy than data sparsity.

References

[1] A. Noulas et al. Mining User Mobility Features for Next Place Prediction in Location-Based Services. *ICDM*, pages 1038–1043, 2012.

[2] A. Roy et al. Location Aware Resource Management in Smart Homes. In *PerCom*, pages 481–488, 2003.

[3] D. Ashbrook and T. Starner. Learning Significant Locations and Predicting User Movement with GPS. In *ISWC*, pages 101–108, 2002.

[4] E. Lu et al. Mining Cluster-Based Temporal Mobile Sequential Patterns in Location-Based Service Environments. *IEEE Transactions on Knowledge and Data Engineering*, 23(6):914–927, 2011.

[5] J. Fukano et al. A next location prediction method for smartphones using block-models. *IEEE Virtual Reality*, pages 1–4, 2013.

[6] S. Akoush et al. Bayesian Learning of Neural Networks for Mobile User Position Prediction. *Computer Communications and Networks*, pages 1234–1239, 2007.

[7] G. Yava et al. A Data Mining Approach for Location Prediction in Mobile Environments. *Data & Knowledge Engineering*, 54(2):121–146, 2005.

[8] Google. Our Mobile Planet. `http://think.withgoogle.com/mobileplanet/en/`. Accessed: July 2014.

[9] H. Cao et al. Mining Frequent Spatio-temporal Sequential Patterns. In *ICDM*, pages 82–89, 2005.

[10] J. Petzold et al. Comparison of Different Methods for Next Location Prediction. In *Euro-Par Parallel Processing*, pages 909–918, 2006.

[11] J.B. Gomes et al. Where Will You Go? Mobile Data Mining for Next Place Prediction. In *DaWaK*, pages 146–158, 2013.

[12] L. Nguyen et al. PnLUM : System for Prediction of Next Location for Users. In *Mobile Data Challenge by Nokia Workshop at Pervasive*, 2012.

[13] L. Vintan et al. Person Movement Prediction Using Neural Networks. In *1st Workshop on Modeling and Retrieval of Context*, 2004.

[14] S. Gambs et al. Next Place Prediction Using Mobility Markov Chains. In *1st Workshop on Measurement, Privacy, and Mobility*, pages 1–6, 2012.

[15] S. Scellato et al. NextPlace : A Spatio-Temporal Prediction Framework for Pervasive Systems. In *Pervasive*, pages 152–169, 2011.

[16] M. Vukovic. Adaptive User Movement Prediction for Advanced Location-aware Services. In *SoftCOM*, pages 343–347, 2009.

[17] Y. Zhang et al. Location Prediction Model Based on Bayesian Network Theory. In *GLOBECOM*, pages 1–6, 2009.