Functional Gaussian Distribution Modelling of Mobility Prediction Accuracy for Wireless Users

Lu Liu, Wuyang Zhou, Sihai Zhang
Key Laboratory of Wireless-Optical Communications, University of Science and Technology of China
Hefei, Anhui, China 230027
Email: liuluzz@mail.ustc.edu.cn, {wyzhou, shzhang}@ustc.edu.cn

Wei Cai
School of Information and Science and Technology, University of Science and Technology of China
Hefei, Anhui, China 230027
Email: caiwei@ustc.edu.cn

Abstract—Mobility entropy is proposed to measure predictability of human movements, based on which, the upper and lower bound of prediction accuracy is deduced, but corresponding mathematical expressions of prediction accuracy keeps yet open. In this work, we try to analyze and model prediction accuracy in terms of entropy based on the 2-order Markov chain model empirical results on a large scale CDR data set, which demonstrates the observation that users with the same level of entropy achieve different levels of accuracy[1]. After dividing entropy into intervals, we fit the probability density distributions of accuracy in each entropy interval with Gaussian distribution and then we estimate the corresponding mean and standard deviation of these distributions. After observing that the parameters vary with increasing entropy, we then model the relationship between parameters and entropy using least squares method. The mean can be modelled as a linear function, while the standard deviation can be modelled as a Gaussian distribution. Thus based on the above analysis, the probability density function of accuracy given entropy can be expressed by functional Gaussian distribution. The insights from our work is the first step to model the correlation prediction accuracy and predictability entropy, thus shed light on the further work in this direction.

Index Terms—Human Mobility; Mobility Prediction; Real Entropy; CDR; Markov Chain;

I. INTRODUCTION

Human mobility prediction plays a critical role in many fields, such as urban planning[2], emergency management[3], transportation engineering[4], public health[5] and location-based services[6]. Various methods have been proposed to predict human mobility, such as Markov chain models[7], artificial neural networks[8][9], Bayesian networks[10][11], and so on. It has been now recognized that prediction accuracy relies on prediction models and data sets. [7] implements mobility prediction based on n-order Markov-based model and figures out that the 2-order Markov-based model has the highest accuracy, while [12] claims that 1-order Markov performs better.

After González et al. [13] demonstrate that human movements show a high degree of spatial and temporal regularity, Song et al. [14] propose an approach to measure the ability of predicting the whereabouts and mobility of individuals. They adopt mobility entropy to characterize the predictability of human mobility, where higher entropy indicates less predictable. Following this work, several research works have used this metric to explore the limits of predictability in human dynamics[12][15][16].

According to the theoretical entropy, though, there are one emerging important question: Can prediction accuracy be described using the mobility entropy? In other words, if someone’s mobility entropy is known, can we tell its prediction accuracy using mathematical formula without after actual prediction, even by just a specific prediction model? This paper is trying to answer this question. To be specific, [14] give a upper bound of prediction accuracy, and this work tries to build a mathematical model to describe the relationship between the achieved accuracy and predictability based on a real data set, even using one certain prediction model. Note that our previous work[11] has already provided the users’ predictability and their accuracy after predictability calculation and 2-order Markov prediction.

In order to analyse accuracies of users with the same level of entropy, entropy value are divided into multiple intervals. We analyze the probability density distributions of prediction accuracy corresponding to the intervals and find that the distributions approximate Gaussian distributions. Then, we fit the distribution of accuracy with Gaussian distribution and obtain the corresponding parameters, mean, µ, and standard deviation, σ, which vary with the increasing entropy. The relationship between the mean and entropy can be modelled as a linear function. As for the standard deviation, the relationship is not so simple to describe. We try three distributions and choose a Gaussian distribution based on MSE(mean square error) comparison at last. Finally, we draw a functional Gaussian distribution to describe the distributions of accuracy given entropy.

The contributions of this paper are as following:

• We model the prediction accuracy based on predictability entropy. Different from [14], which deduces the upper and lower bound of prediction accuracy based on entropy, this work is pioneering to describe the correlation between accuracy and entropy in the form of probability density distribution. Although this work is just based on one real world CDR data set and 2-order Markov model, it sheds light on future work in this direction.

• We provide an inference approach to model the correlations and estimate the parameters. We also conclude
that correlations can be modelled as a linear function and Gaussian distribution with \( \mu = -0.1726s + 0.9845 \) and \( \sigma = 0.09415e^{-(s-2.545)^2} \), where \( s \) is discrete entropy value. Following work may take this way to investigate more data sets and prediction models to obtain more general conclusions.

II. Data Sets

In this section, the details of the real mobile CDR data set used in this work will be introduced, which is provided by a Chinese telecom operator. This data set starts from Tuesday, July 1, 2014, and ends on Wednesday, Dec. 31, 2014, which possesses 194,336 anonymous mobile phone subscribers registering in one Chinese city. Totally 270,932,374 records provide the spatial and temporal information of telecom users. The record format is shown as below:

(SERVICE_NBR, CALL_TYPE, OPPOSITE_NO, TOLLTYPE_ID, ROAM_TYPE, START_TIME, END_TIME, DURATION, CITY_ID, ROAM_CITY_ID, OPPCITY_ID and LAC_ID)

LAC_ID(Location Area Code ID) is adopted as the user’s location, and 453,752 locations are included in the data set. In order to extract the trajectories of users, we take the spatial and temporal features and remove the others. Then, the four features are left: SERVICE_NBR, START_TIME, ROAM_CITY_ID, LAC_ID, for predictability calculation and next place prediction.

In order to make sure that every prediction object has enough spatial and temporal information, we define the active day concept as a user has at least one location update on that day and remove the users whose active days are less than 150. After the processing, 142,288 users with total 247,780,761 records are left.

III. Predictability Entropy and Prediction Accuracy

In this section, we describe the basic data of the work. First, we introduce the predictability measurement, entropy, and show the predictability measurement results of the users in the CDR data. Then, we show the prediction model and correlation between accuracy and entropy.

A. Entropy Calculation

Entropy is a basic quantity for measuring the degree of predictability characterizing a time series in spatial-temporal space[14]. The definitions of the three entropy measures for each individual’s mobility pattern are briefly described below and more details are shown in the original literatures mentioned above.

i) Random entropy: This entropy focuses on the number of unique locations that a person visits, assuming that each location is visited with equal probability by one user. Thus, it is defined as \( S_{\text{rand}} = \log_2 N \), where \( N \) is the number of unique locations to which a user has traveled.

ii) Temporal-uncorrelated entropy: This entropy takes the visiting frequency of locations into consideration. It is defined as \( S_{\text{unc}} = -\sum_i p(l_i) \log_2 p(l_i) \), where \( p(l_i) \) is the probability that location \( l_i \) was visited by the user.

iii) Real entropy: This entropy captures the full spatial-temporal order presented in a user’s mobility pattern. Suppose the location sequence of a person’s trajectory is \( \text{Traj} = l_1 \rightarrow l_2 \rightarrow l_3 \rightarrow \cdots \). The real entropy can be calculated as

\[
S_{\text{real}} = -\sum_{\text{Traj}'} p(\text{Traj}') \log_2 p(\text{Traj}'),
\]

where \( \text{Traj}' \) is a subsequence of \( \text{Traj} \) and \( p(\text{Traj}') \) is the probability of \( \text{Traj}' \) appearing in \( \text{Traj} \).

As real entropy incorporates a person’s spatial and temporal features, we only consider real entropy of users in this paper. Based on the previous work, we find that users with the same entropy have different prediction accuracies. In order to figure out the characteristic of the accuracy of users with the same level of entropy, we divide the range of entropy into intervals, length of 0.05. There are 84 entropy intervals in total. The number of users in each interval is shown in the Fig. [1].

B. Next Place Prediction

2-order Markov chain model is an effective prediction approach to make next place prediction. Thus we set it as prediction model. It assumes that an individual’s next place is related to the previous 2 visited locations and the trajectory of an individual can be modeled as a 2-order Markov chain. Based on the CDR data, transition probability can be calculated.

\[
P(X_{t+1} = l_{t+1}|X_t = l_t, X_{t-1} = l_{t-1}) = P(X_{t+1} = l_{t+1}|X_t = l_t, X_{t-1} = l_{t-1})
\]

where \( X_t \) is a variable representing the location for an individual at step \( t \).

According to the transition matrix \( P \), we can make next place prediction. The location corresponding to the maximum transition probability is chosen as the next place.

Based on the entropy and prediction accuracy of every user, we can get the relationship between predictability entropy and prediction accuracy as shown in Fig. [2]. Obviously, users with
the same entropy achieve different levels of accuracy. In the next section, we try to model the accuracy corresponding to different levels of entropy.

IV. PREDICTION ACCURACY MODELLING

In this section, we model the prediction accuracy based on the predictability entropy and the prediction result. The probability density distribution of accuracy can be fitted with Gaussian distribution, and we get the parameters, mean and standard deviation, respectively. The parameters vary with increasing entropy and they can be estimated by least squares method. After estimation, we draw a formula to describe the distribution of accuracy given entropy.

A. Accuracy Modelling Based on Entropy

Observing the correlation between accuracy and entropy, we are surprised that users with the same entropy achieve various prediction accuracies. Then, we decide to research the distribution of achieved accuracy of the users with the same entropy. In Section III, the range of entropy are divided into 84 intervals, and we analyze the results of accuracy in the same entropy intervals in this section.

In Fig. 2 we observe that the distribution of accuracy in every entropy interval is similar to unimodel distribution and the distribution shows a degree of symmetry. Accordingly, it is reasonable to fit the distribution of accuracy given entropy with Gaussian distribution.

A general formal of Gaussian distribution is shown below:

\[ P(x|s) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), -\infty < x < \infty. \]  

where \( x \) can be used to represent accuracy, \( s \) is denoted as entropy, \( \mu \) can represent the mean of accuracy, and \( \sigma \) can equal the standard deviation of accuracy.

In Fig. 2 we can see that users with higher entropy achieve lower accuracy. Thus, \( \mu \) decreases with increasing entropy. The size of accuracy range in the different entropy intervals also varies, and then \( \sigma \) is also a variable changing with entropy. We fit the probability density distribution of accuracy and make parameter estimation in the following subsections.

B. Fitting Accuracy of Entropy Interval

We analyze the probability density distribution of accuracy in each entropy interval which is estimated through kernel density estimation and use Gaussian distribution to fit them and obtain the parameters. The parameters for the 84 intervals are calculated, including mean \( \mu \) and standard deviation \( \sigma \).

In order to showing the performance, we take results of 9 intervals as examples shown in Fig. 3. For interval [0, 0.05], \( \mu = 0.999 \) and \( \sigma = 0.002 \). For interval [0.5, 0.55], \( \mu = 0.928 \) and \( \sigma = 0.019 \). For interval [1, 1.05], \( \mu = 0.814 \) and \( \sigma = 0.070 \). For interval [1.5, 1.55], \( \mu = 0.689 \) and \( \sigma = 0.084 \). For interval [2, 2.05], \( \mu = 0.580 \) and \( \sigma = 0.0865 \). For interval [2.5, 2.55], \( \mu = 0.500 \) and \( \sigma = 0.081 \). For interval [3, 3.05], \( \mu = 0.449 \) and \( \sigma = 0.079 \). For interval [3.5, 3.55], \( \mu = 0.419 \) and \( \sigma = 0.102 \). For interval [4, 4.05], \( \mu = 0.279 \) and \( \sigma = 0.057 \). Complete results of parameters are shown as the "\( \mu \) of Gaussian Fitting" and "\( \sigma \) of Gaussian Fitting" curves in Fig. 4 and Fig. 5. We can observe that \( \mu \) decreases with the increasing entropy. However, \( \sigma \) shows non-linear relationship with entropy.

C. Least Squares Parameter Estimation

When we get the parameters of distribution in each entropy interval, mean \( \mu \) and standard deviation \( \sigma \), which both vary with increasing entropy, it needs to find the correlation between parameters and entropy. For mean value, it shows linear relationship with entropy. However, it is not so simple for standard deviation. So, we try different distributions for standard deviation and make selection according to MSE.

The mean, which can seem as the average of prediction accuracy, decreases with the increasing entropy. The relationship between mean and entropy is similar to linear relationship as shown in Fig. 4. We set the linear function as the basis function. Through least squares method, the mean corresponding to entropy intervals can be described as \( \mu = -0.1726s + 0.9845 \), where \( s \) is the upper bound of entropy intervals. The result is shown in Fig. 4. The "\( \mu \) of Gaussian Fitting" is the mean of Gaussian fitting result in different entropy intervals. The "Linear" is the estimation result given entropy.

Contrary to the mean value, standard deviation has up and down with the increasing entropy. Consequently, we have tried different basis functions, including polynomial, Gaussian distribution and double Gaussian distribution. Using least squares method, the standard deviation of entropy intervals can be described as \( \sigma = -0.01689s^2 + 0.08373s - 0.01124, \sigma = 0.09415s \exp\left(-\frac{(s-2.548)^2}{0.552}\right), \sigma = 0.08199s \exp\left(-\frac{(s-3.534)^2}{0.552}\right) + 0.09348s \exp\left(-\frac{(s-1.92)^2}{0.552}\right) \), corresponding to "Polynomial", "Gaussian", and "Double Gaussian" in Fig. 5. The "\( \sigma \) of Gaussian Fitting" is the standard deviation of Gaussian fitting result in different entropy intervals.

Based on parameter estimation results, we can obtain the estimation of probability density function of accuracy corresponding to each entropy interval. We also take the same 9 intervals as examples to show the estimation results in Fig. 6.
In order to compare the performance of different estimating results for the standard deviation, we calculate the MSE respectively. The three estimations perform well in the middle of the range of entropy but perform a little poor at the both ends of the range. The MSE of the "Polynomial" is the largest, and the MSE of the "Gaussian" is smallest, which says that the "Gaussian" have the best performance which can be used to describe the relationship between standard deviation and entropy.

We use Kolmogorov-Smirnov test to check whether the Gaussian distribution based on the estimated $\mu$ and $\sigma$ follows the same distribution with the original distribution of accuracy. The testing results demonstrate the distributions based the estimated parameters follows the same distribution with the original distributions of accuracy on majority of the intervals. However, on 30% entropy intervals the results can’t pass the test, which means that the distribution on these intervals can’t be represented by the estimated result. It’s due to that the distribution of accuracy on these intervals are not completely asymmetrical. The Gaussian distribution based on the estimated results is just an approximation of the original distributions.

**D. Drawing A Functional Gaussian Distribution**

According to the parameters estimation, we obtain the parameters given entropy. Based on the expression form of Gaussian distribution, we can draw a functional Gaussian distribution to describe the probability density function of accuracy given the entropy interval in Eq. 4.
Standard deviation, $\sigma$, is a continuous variable, while $x$ is a discrete variable as $x$ is the discretized entropy. Based on the correlation, we draw a functional Gaussian distribution to describe the probability density function of accuracy given entropy.

Although we propose a formula to describe the correlation based on a real data set and a certain prediction model, it is not accurate and general enough. In the future work, we will try to find more accurate expression of the accuracy distribution and focus on more general conditions and deduce more general results.

**VI. Conclusions**

In this work, we model the accuracy based on entropy. As the users with the same level of entropy achieve different accuracies, we divide the entropy into intervals and analyze the probability density distribution of accuracy in each interval. All of the distributions follow unimodel distributions. We fit them with Gaussian distribution and obtain the mean and standard deviation of each interval. Then, we find the correlation between the parameters and entropy intervals. The mean can be modelled as $\mu = -0.1726s + 0.9845$, and standard deviation can be modelled as $\sigma = 0.09415 \exp\left(-\frac{(s - 2.548)^2}{1.96}\right)$, where $s$ is the discretized entropy. Based on the correlation, we draw a functional Gaussian distribution to describe the probability density function of accuracy given entropy.

ACKNOWLEDGMENT

This work was partially supported by Key Program of Natural Science Foundation of China under Grant(61631018), the Fundamental Research Funds for the Central Universities, Huawei Technology Innovative Research.

REFERENCES

[1] Lu Liu, Sihai Zhang, Wuyang Zhou, "Empirical Verification on Human Mobility Predictability using Call Detail Records," submitted to 2019 IEEE International Conference on Communications (ICC).
[2] Roberto Camagni, Maria Cristina Gibelli, Paolo Rigamonti, "Urban mobility and urban form: the social and environmental costs of different patterns of urban expansion," Ecological economics, vol. 40, no. 2, pp. 199-216, 2002.
[3] James P. Bagrow, Dashun Wang, Albert-László Barabási, "Collective response of human populations to large-scale emergencies," PloS one, vol. 6, no. 3, pp. e17680, 2011.
[4] Bin Jiang, Junjun Yin, and Sijian Zhao, "Characterizing the human mobility pattern in a large street network," Physical Review E, vol. 80, no. 2, pp. 021136, 2009.
[5] Amy Wesołowski, Nathan Eagle, Andrew J. Tatem, David L. Smith, Abdusalan M. Noor, Robert W. Snow, Caroline O. Buckee, "Quantifying the impact of human mobility on malaria," Science, vol. 338, no. 6104, pp. 267-270, 2012.
[6] Dimitrios Buhalis, Aditya Amaranggana, Smart tourism destinations, Information and communication technologies in tourism 2014. Springer, Cham, 2013, pp. 553-564.
[7] Sébastien Gambis, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez, "Next place prediction using mobility markov chains," Proceedings of the First Workshop on Measurement, Privacy, and Mobility, vol. 3, 2012.
[8] J. Tkačik and P. Kordík, "Neural Turing Machine for sequential learning of human mobility patterns," 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, 2016, pp. 2790-2797.
Fig. 6. Based on the parameter estimation, the estimation of probability density distribution of accuracy in entropy intervals. (a) Entropy interval [0, 0.05), (b) Entropy interval [0.5, 0.55), (c) Entropy interval [1, 1.05), (d) Entropy interval [1.5, 1.55), (e) Entropy interval [2, 2.05), and (f) Entropy interval [2.5, 2.55), (g) Entropy interval [3, 3.05), (h) Entropy interval [3.5, 3.55), and (i) Entropy interval [4, 4.05).

Fig. 7. MSE(mean square error) of the three estimation methods. [9] R. Jiang, X. Song, Z. Fan, T. Xia, Q. Chen, Q. Chen, and R. Shibasaki, "Deep ROI-Based Modeling for Urban Human Mobility Prediction;" [10] F. Alhasoun, M. Alhazzani and M. C. González, "City Scale Next Place Prediction from Sparse Data through Similar Strangers," Proceedings of ACM KDD Workshop, pp. 191-196, 2017. [11] Q. Lv, Y. Qiao, N. Ansari, J. Liu and J. Yang, "Big Data Driven Hidden Markov Model Based Individual Mobility Prediction at Points of Interest," in IEEE Transactions on Vehicular Technology, vol. 66, no. 6, pp. 5204-5216, June 2017. [12] Xin Lu, Erik Wetter, Nita Bhatt, Andrew J. Tatem and Linus Bengtsson, "Approaching the limit of predictability in human mobility," Scientific Reports, vol. 3, no. 2923, Oct 2013. [13] Marta C. González, César A. Hidalgo and Albert-László Barabási, "Understanding individual human mobility patterns," Nature, vol. 453, no. 7196, pp. 779-782, June 2008. [14] Chaoming Song, Zehui Qu, Nicholas Blumm, Albert-László Barabási, "Limits of predictability in human mobility," Science, vol. 327, no. 5968, pp.1018-1021, 2010. [15] Qin S-M, Verkasalo H, Mohtaschemi M, Hartonen T, Alava M, "Patterns, Entropy, and Predictability of Human Mobility and Life," PLoS ONE, vol. 7, no. 12, Dec 2012.
[16] Eunjoon Cho, Seth A. Myers, and Jure Leskovec, "Friendship and mobility: user movement in location-based social networks," in KDD, pp. 1082-1090, Aug 2011.