Exploring Cell Tower Data Dumps for Supervised Learning-based Point-of-Interest Prediction

Ran Wang[1] Chi-Yin Chow[1] Sarana Nutanong[1] Yan Lyu[1] Yanhua Li[2]
Mingxuan Yuan[2] Victor C. S. Lee[1]
[1]Department of Computer Science, City University of Hong Kong, Hong Kong
[2]Huawei Noah's Ark Lab, Hong Kong
{ranwang3-c@my., chiychow@, snutanon@, yanlv2-c@my.}cityu.edu.hk, {li.yanhua1, yuan.mingxuan}@huawei.com, csvlee@cityu.edu.hk

ABSTRACT

Exploring massive mobile data for location-based services (LBS) becomes one of the key challenges in mobile data mining. In this paper, we propose a framework that uses large-scale cell tower data dumps and extracts points-of-interest (POIs) from a social network web site called Weibo, and provides new LBS based on these two data sets, i.e., predicting the existence of POIs and the number of POIs in a certain area. We use Voronoi diagram to divide a city area into non-overlapping regions, and a k-means clustering algorithm to aggregate neighboring cell towers into region groups. A supervised learning algorithm is adopted to build up a model between the number of connections of cell towers and the POIs in different region groups, where a classification or regression model is used to predict the POI existence or the number of POIs, respectively. We studied 12 state-of-the-art classification and regression algorithms, and the experimental results demonstrate the feasibility and effectiveness of the proposed framework.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications-Spatial databases and GIS; H.2.8 [**Database Management**]: Database Applications-Data mining

General Terms

Algorithms, Experimentation

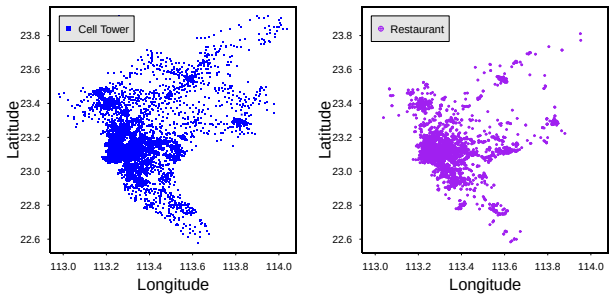
Keywords

Spatio-temporal data analysis, classification, regression, cell tower data dumps, point-of-interest prediction

1. INTRODUCTION

SIGSPATIAL'14, November 04-07 2014, Dallas/Fort Worth, TX, USA

http://dx.doi.org/10.1145/2666310.2666478.



(a) Cell Towers (b) Restaurants

Figure 1: Geographical distribution of cell towers and restaurants in the Guangzhou city of China.

The ubiquity of mobile devices such as smartphones and tablet computers enables us to collect useful spatial and temporal data in a large scale and also opens up the possibility of extracting useful information from the data.

In this paper, we focus on a specific type of mobile-user data known as *cell tower data dumps*, which contain connection records collected by 9,563 cell towers operated by the China Mobile Limited[1] in Guangzhou, China, as illustrated in Figure 1(a). This data set was collected within a time period of six days (from 4 September 2013 to 9 September 2013). For the purpose of this investigation, we focus on records produced by phone calls and SMSs. For each record, we use the connection time, identifier and location of each cell tower. We extracted 18,290 restaurants in Guangzhou from Weibo[2], a popular Chinese social network web site, as our point-of-interest (POI) data set, as depicted in Figure 1(b).

The main objective of this research is to make use of the cell phone and POI data sets to help predict the existence or number of POIs in the vicinity of a cell tower. Our investigation is driven by a hypothesis that *there is a correlation between the collective behavior of mobile users and the existence of a certain type of POIs in a certain area.* In general, the contributions of our work can be summarized as follows.

- We formulate spatial and temporal representation methods of cell tower data dumps for mobile user behaviors.

[1]http://www.chinamobileltd.com
[2]http://weibo.com

- We design a framework with classification and regression algorithms to build up a model between mobile users' behaviors and LBS.
- We conduct extensive evaluation of our framework on real cell tower data dumps and POI data set.

The remainder of this paper is organized as follows. Section 2 highlights related work. In Section 3, we describe the proposed framework for POI predictions. In Section 4, we present implementation details and analyze extensive experimental results. Finally, Section 5 concludes this paper.

2. RELATED WORK

Most existing work on mobile and spatio-temporal data focuses on recommender systems [1, 10, 12], urban planning [2], discovering [11], social networking services [13], etc. The most commonly used techniques include collaborative filtering, density estimation, image and signal processing, etc. However, none of them put their focus on machine learning, especially supervised learning, which is also a potential tool to mine useful information and make accurate prediction on mobile phone call data or cellular network data for valuable location-based applications.

Supervised learning [4] refers to the problem of inferring a model from a set of labeled training samples, in order to achieve accurate predictions on unseen data. Given a training set $\mathbb{X}$ with N labeled samples, i.e., $\mathbb{X} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, each sample is associated with a set of conditional attributes $\mathbf{x}_i = \{x_{i1}, x_{i2}, \ldots, x_{iL}\}$ and a decision attribute y_i. The goal is to learn a function $f : \mathbf{x} \rightarrow y$, such that given a new unlabeled sample $\hat{\mathbf{x}} = \{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_L\}$, its desired output value could be predicted by $\hat{y} = f(\hat{\mathbf{x}})$. Besides, the learning task is classification or regression if the decision attribute is discrete or continuous, respectively.

Supervised learning covers a wide range of application domains such as image processing, text classification, face recognition, video indexing, etc. Besides, several learning techniques have been applied on mobile and spatio-temporal data in recent literature. In [6], kernel-based SVM is used as a classifier in the detection of harmful algal blooms in the Gulf of Mexico based on mobile data. In [9], the random forest approach is used to classify the land usage in a city based on mobile phone activities. In [3], a density-based clustering algorithm is proposed for a wide range of spatio-temporal data. To the best of our knowledge, no one has applied supervised learning models to predict the POI existence or the number of POIs in a certain region of a city using cell tower data dumps.

3. USING SUPERVISED LEARNING FOR POI PREDICTIONS

In this section, we formulate spatial and temporal representation methods of cell tower data dumps for mobile user behaviors in Sections 3.1 and 3.2, and then present our proposed framework for POI prediction in Section 3.3.

3.1 Pre-clustering of Cell Towers

As demonstrated in Figure 1, the geographical distributions of the cell towers and POIs in Guangzhou city are roughly consistent with each other. That is to say, if a given region has a larger number of cell towers, it also has a high chance to cover a larger number of POIs, and vice versa. Besides, the density of cell towers is also related to the user visiting rate. For example, the downtown is usually the most popular and busiest area in a city, so it records the highest user visiting rate, and thus needs more cell towers. In comparison, relatively fewer people visit the suburb in a day, thus the density of cell towers is lower in such area. Having these basic observations, it is possible to predict the POI existence or the number of POIs in a region based on the user visiting rate, which is reflected by the number of connections established by cell towers in that region.

Given N cell towers $\mathbb{T} = \{T_1, \ldots, T_N\}$ with geographical location information, we denote $T_i = (t_{i1}, t_{i2})$, where t_{i1} and t_{i2} represent the longitude and latitude of T_i, $i = 1, \ldots, N$, respectively. The intuitive scheme is to divide the city into N regions $\mathbb{R} = \{R_1, \ldots, R_N\}$, such that each region contains one cell tower. These regions could be defined by the Voronoi diagram [5], which treats each cell tower as a seed. Given a point in a region, the point is closer to the seed of the region than the seeds in other regions, i.e., $\forall \mathbf{x} \in R_i, d(\mathbf{x}, T_i) \leq d(\mathbf{x}, T_j)$, where $i \in \{1, \ldots, N\}$, and $j = 1, \ldots, i-1, i+1, \ldots, N$.

Suppose there is a set of M POIs $\mathbb{P} = \{P_1, \ldots, P_M\}$ with geographical location information, we denote $P_i = (p_{i1}, p_{i2})$, where p_{i1} and p_{i2} represent the longitude and latitude of P_i, $i = 1, \ldots, M$, respectively. For a given POI P_i, the region that covers P_i could be discovered by a nearest neighbor (NN) search process among $\mathbb{T}$. Finally, the number of POIs in each region is computed as the target that we aim to predict. However, when it comes to a real application, we have to consider the following two issues:

- The signal intensity of a cell tower is not stable, which leads to an unreliable relation between the POI density and the number of connections.
- Due to an unbalanced distribution of cell towers, the separated regions may be too small in downtown and too large in suburb. As a result, the number of POIs covered by them may be balanced out and have no obvious difference.

In order to overcome the above-mentioned problems, we conduct a pre-clustering process on the cell towers, such that the cell towers with similar geographical information are grouped into one cluster. Accordingly, their regions defined by the Voronoi diagram are merged, and the numbers of their covered POIs are summed up as the target that we aim to predict. As the most widely used one, k-means clustering technique [7] is adopted, which aims to partition N observations (i.e., $T_1, \ldots, T_N$) into k sets (i.e., $\mathbb{S} = \{S_1, \ldots, S_k\}$), so as to minimize the within-cluster sum of square, i.e., $\text{argmin}_{\mathbb{S}} \sum_{i=1}^{k} \sum_{T_j \in S_i} ||T_j - \mu_i||$, where $\mu_i = \frac{1}{k_i} \sum_{T_j \in S_i} T_j$, and k_i is the number of cell towers in the i-th cluster.

In a real application, it is always difficult to get optimal k. Thus, we test the values of $\{250, 500, 1000, 2500, 5000\}$. Due to space limitation, we only plot the clustering result when $k = 250$, as shown in Figure 2.

3.2 Refine Time Resolution

We aim to use spatio-temporal data to perform POI predictions. In Section 3.1, we have introduced how to make use of the spatial data. In this section, we further discuss how to make use of the temporal data.

Basically, the time in a day can be divided into 24 slots in the unit of an hour. Each slot defines a feature for the cell tower T. Each connection record indicates that a user has visited the region covered by T, thus the connection

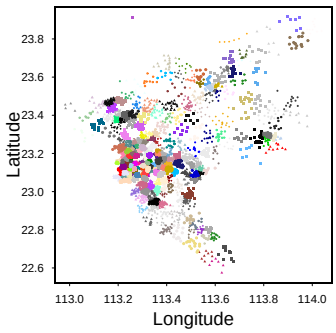
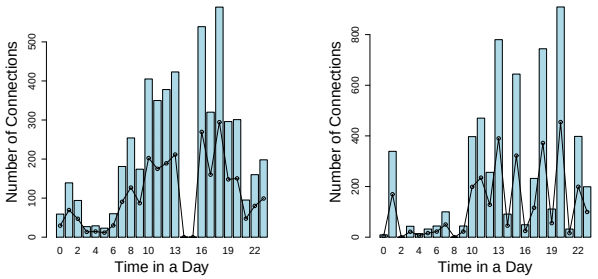


Figure 2: Pre-clustering result of cell towers when $k = 250$.



(a) Wednesday (b) Saturday

Figure 3: Connection frequency of the cell towers in a region in two different days.

frequency distribution of a region could possibly reflect the characteristics of its user visiting rate. Given a region, the distributions in different days are supposed to be similar. However, this statement does not hold in the reality. Figures 3(a) and 3(b) demonstrate the connection frequency distributions of a region on Wednesday and Saturday, respectively. We pay attention to two observations: 1) the distribution in a weekday is more uniform than that of weekend; and 2) there may be some missing values, which give zero connection in a time slot.

The first observation is easy to explain, since people always have quite different living habits in weekdays and weekends. In weekdays, they have a regular time schedule for working and rest. While in weekends, there is no regular pattern, even the same person can take part in quite different activities. As for the second observation, it is possibly caused by some facility problems, such as a poor signal intensity or periodic maintenance of the cell towers. Thus, we refine the time resolution of a weekday into seven new time slots as listed in Table 1, and compute the new features as the average number of connections during the slots. As for the weekend, the 24 time slots are retained. Finally, the feature vector of a given region R_i is denoted as $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iL})$, where $L = 7 * 4 + 24 * 2 = 76$ (i.e., four weekdays and two weekend days), with each dimension reflecting its user visiting rate in a specific time slot.

3.3 The Proposed Framework

Given a certain area in a city, a company wants to know whether there should exist any POI (i.e., restaurants) or how many POIs should be there for business planning. Thus, it is useful to resolve these problems from the viewpoints of both classification and regression. Finally, the POI prediction framework is sketched in the following three steps.

Table 1: Time slots in a weekday

Time slot	Duration	Activity
0:00 to 7:00	7 hours	Sleeping hours
7:00 to 9:00	2 hours	Morning rush hours
9:00 to 12:00	3 hours	Morning working hours
12:00 to 14:00	2 hours	Lunch hours
14:00 to 18:00	4 hours	Afternoon working hours
18:00 to 21:00	3 hours	Evening rush & dinner hours
21:00 to 24:00	3 hours	Home hours

Step 1: The Voronoi diagram step. In this step, the city is divided into a number of consecutive regions based on the Voronoi diagram by taking the cell towers as the seeds. Then, the number of POIs located in each region is found.
Step 2: The clustering step. This step performs the k-means clustering algorithm on the cell towers, and the cell towers with similar geographical locations are grouped into the same cluster. Each cell tower cluster defines a region of the city, with a feature vector (i.e., including the cell tower identifier and time of each connection) extracted from the cell tower data dumps.
Step 3: The supervised learning step. Finally, the POI existence (treated as positive if there exists any POI and negative if no POI exists) or the number of POIs is taken as the output target of the region. Based on these labeled regions, a learner f is built up based on these labeled regions for a classification or regression model that will be used to predict POI existence or the number of POIs in new regions.

Once the learner f is trained based on a set of given regions, it can be used in two directions. (1) Prediction of unknown regions: when there comes a new region without any POI information, we can extract its feature vector from the user connection records of the cell tower data dumps and predict the POI existence or the number of POIs by f for business decision making. (2) Evaluation of existing regions: given a region, if the regression number of POIs is larger than or equal to the real one, there may be adequate number of POIs; however, if the regression number is smaller than the actual one, it indicates a possibility to set up more POIs in the future.

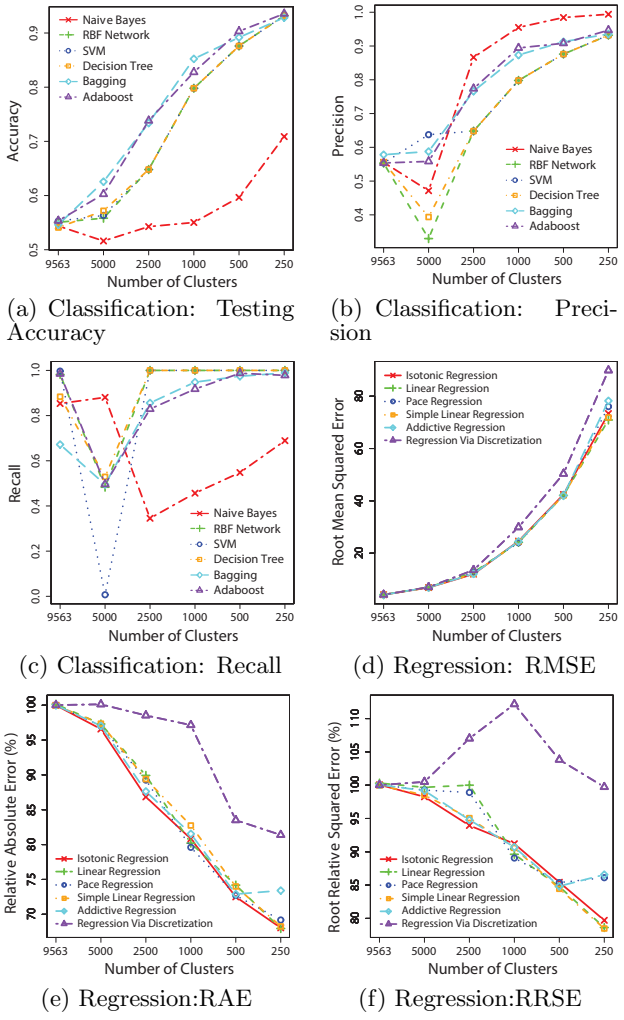
4. IMPLEMENTATION AND ANALYSIS

We here present the implementation details of the proposed framework. Note that the model selection is not the main concern in this work, thus we just adopt several widely used parameter settings for the learning algorithms.

Classification mode. The purpose is to correctly identify whether there exists any POI in a given region of a city. We study six state-of-the-art algorithms, which are naive Bayes classifier (NBC), radial basis function (RBF) network, SVM, decision tree, bagging, and adaboost.

Regression mode. The purpose is to predict the number of POIs located in a given region of a city. We also study six state-of-the-art algorithms, which are isotonic regression, linear regression, pace regression, simple linear regression, addictive regression, and regression via discretization.

We first conduct the experiments on the original data without a pre-clustering process, then set the number of clusters as $\{250, 500, 1000, 2500, 5000\}$ and observe the trend of the result. In order to avoid random effects, we conduct 10-fold cross validation 10 times, and observe the average values. The experiments are conducted with the standard machine learning toolbox WEKA [8], which are performed



(a) Classification: Testing Accuracy

(b) Classification: Precision

(c) Classification: Recall

(d) Regression: RMSE

(e) Regression:RAE

(f) Regression:RRSE

Figure 4: Comparative results of different learning models.

on a computer with an Intel Core 2 Duo CPU with 4GB memory, it runs on 32-bit Windows 7.

We adopt three metrics to evaluate the classification performance, i.e., testing accuracy, precision, and recall, which respectively give the rate of correctly classified testing samples, the correct rate in the set that has been classified as positive, and the correct rate in the real positive set.

As for the regression results, the most commonly used evaluation metric is the root mean squared error (RMSE). However, the ranges of the prediction targets differ a lot with different numbers of clusters, which lead to some incomparable results. In this case, we adopt another two metrics, i.e., relative absolute error (RAE) and root relative squared error (RRSE). Basically, the RAE takes the total absolute error and normalizes it by dividing the total absolute error of the simple predictor, and the RRSE takes the total squared error and normalizes it by dividing the total squared error of the simple predictor. For both the RAE and RRSE, smaller values are better, and 100% represents the baseline of just predicting the mean. Thus, values less than 100% are considered as effective for predicting the number of POIs.

The average values of the 10×10 results (10-fold cross validation 10 times) with different numbers of clusters are shown in Figure 4. Basically, we have two observations: (1) for both classification and regression modes, the prediction is more accurate when the number of clusters is smaller; (2) different learning algorithms have different advantages regarding different evaluation metrics, thus it is important to select an appropriate model based on the problem requirement.

5. CONCLUSION

In this paper, we proposed a supervised learning-based framework for predicting the existence of POIs and the number of POIs in a given region using the spatio-temporal features extracted from cell tower data dumps in Guangzhou, China and the information of a set of restaurants collected from the Chinese social network Weibo. We have studied 12 state-of-the-art classification and regression algorithms. Experimental results show the feasibility and effectiveness of the proposed framework.

6. ACKNOWLEDGMENTS

R. Wang and C.-Y. Chow were partially supported by a research grant (CityU Project No. 9231131). S. Nutanong was partially supported by a CityU research grant (CityU Project No. 7200387). Y. Lyu and Victor C. S. Lee were partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU Project No. 115312).

7. REFERENCES

[1] J. Bao, Y. Zheng, and M. F. Mokbel. Location-based and preference-aware recommendation using sparse geo-social networking data. In *ACM SIGSPATIAL*, 2012.

[2] R. A. Becker, R. Caceres, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky. A tale of one city: Using cellular network data for urban planning. *IEEE Pervasive Computing*, 10(4):18–26, 2011.

[3] D. Birant and A. Kut. St-dbscan: An algorithm for clustering spatial–temporal data. *DKE*, 60(1):208–221, 2007.

[4] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., 2006.

[5] S. Ghosh, K. Lee, and S. Moorthy. Multiple scale analysis of heterogeneous elastic structures using homogenization theory and voronoi cell finite element method. *IJSS*, 32(1):27–62, 1995.

[6] B. Gokaraju, S. S. Durbha, R. L. King, and N. H. Younan. A machine learning based spatio-temporal data mining approach for detection of harmful algal blooms in the Gulf of Mexico. *IEEE J-STARS*, 4(3):710–720, 2011.

[7] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Applied Statistics*, pages 100–108, 1979.

[8] G. Holmes, A. Donkin, and I. H. Witten. Weka: A machine learning workbench. In *ANZIIS*, 1994.

[9] J. L. Toole, M. Ulm, M. C. González, and D. Bauer. Inferring land use from mobile phone activity. In *ACM UrbComp*, 2012.

[10] M. Ye, P. Yin, W.-C. Lee, and D.-L. Lee. Exploiting geographical influence for collaborative point-of-interest recommendation. In *ACM SIGSPATIAL*, 2011.

[11] J. Yuan, Y. Zheng, and X. Xie. Discovering regions of different functions in a city using human mobility and pois. In *ACM SIGKDD*, 2012.

[12] J.-D. Zhang and C.-Y. Chow. iGSLR: Personalized geo-social location recommendation: A kernel density estimation approach. In *ACM SIGSPATIAL*, 2013.

[13] Y. Zheng, Y. Chen, X. Xie, and W. Y. Ma. Geolife2.0: A location-based social networking service. In *IEEE MDM*, 2009.