Exploiting Foursquare and Cellular Data to Infer User Activity in Urban Environments

Anastasios Noulas, Cecilia Mascolo
Computer Laboratory
University of Cambridge
name.surname@cl.cam.ac.uk

Enrique Frias-Martinez
Telefonica Research
Madrid, Spain
efm@tid.es

***Abstract*—Inferring the type of activities in neighborhoods of urban centers may be helpful in a number of contexts including urban planning, content delivery and activity recommendations for mobile web users or may even yield to a deeper understanding of the geographical evolution of social life in the city . During the past few years, the analysis of mobile phone usage patterns, or of social media with longitudinal attributes, have aided the automatic characterization of the dynamics of the urban environment.**

In this work, we combine a dataset sourced from a telecommunication provider in Spain with a database of millions of geotagged venues from Foursquare and we formulate the problem of urban activity inference in a supervised learning framework. In particular, we exploit user communication patterns observed at the base station level in order to predict the activity of Foursquare users who checkin-in at nearby venues. First, we mine a set of machine learning features that allow us to encode the input telecommunication signal of a tower. Subsequently, we evaluate a diverse set of supervised learning algorithms using labels extracted from Foursquare place categories and we consider two application scenarios. Initially, we assess how hard it is to predict specific urban activity of an area, showing that Nightlife and Entertainment spots are those easier to infer, whereas College and Shopping areas are those featuring the lowest accuracy rates. Then, considering a candidate set of activity types in a geographic area, we aim to elect the most prominent one. We demonstrate how the difficulty of the problem increases with the number of classes incorporated in the prediction task, yet the classifiers achieve a considerably better performance compared to a random guess even when the set of candidate classes increases.

I. Introduction

The long-standing vision of smart cities, where intelligent infrastructure and technologies are employed to improve services for its citizens is being progressively materialized. A number of urban authorities around the world have launched projects that range from optimizing transport and communications, to minimizing the impact of urban activity on the environment [2], [8]. Novel services are being designed by telecommunication and IT companies[5], whereas academic and industrial research units have been organizing conferences and inter-disciplinary meetings [1], [7]. Already, this activity has resulted in a diverse research output in the area [19], [20], [26].

The movement towards the *city-operating system* [6], has been largely driven by the ever increasing usage of smartphones and mobile web services. These computationally efficient mobile devices present an excellent platform for the development of novel applications, many of which exploit GPS sensors to *anchor* photos, users or real places to a digital map. Facebook, for instance, has already introduced a feature that allows users to geo-tag every type of post in the service [12], whereas the number of tweets with geographic information is progressively increasing compared to microblogging content that lacks any geographic reference. The rapid entrance to the mobile web era has also been marked by the rise of online social networks that explicitly use *location* as an essential element of their services. Examples are Gowalla, recently acquired by Facebook [29], and Foursquare [3] which currently numbers more than twenty million users [30] and which is the dominant force in the area of mobile social networks. Nevertheless, the datasets emerging from on-line and mobile web services constitute only one of the many digital information layers covering the geography of the city. Since, the introduction of mobile phones, almost twenty years ago, there has also been a generation of massive datasets describing human telecommunication activity in the city as captured by BTS (base transceiver station) cells.

In this work we are combining a dataset of millions of Foursquare venues with cellular data to infer *types* of urban activity in the neighbourhoods of a city. More specifically, we have collected a large set comprised of millions Call Detail Records (CDR) recorded by the BTS towers of a large Telecommunication provider in Spain. By processing these records that correspond to the telecommunication patterns of approximately twelve million users, we mine a set of machine learning *features* in order to represent the *local* telecommunication *signal* of a BTS tower. Further, we analyze the check-in patterns of Foursquare users at places that are geographically close to BTS towers and exploit their semantic annotations in order to characterize nearby areas in terms of well known urban activities such as *Food, Nightlife, Work* and *Travel* among others. We formulate a supervised learning task that aims to build a learning function to associate the input telecommunication signal of a BTS tower to the Foursquare categories of nearby places. This would potentially lead to the characterization of geographic areas where only CDR data is available.

More specifically our work answers two questions:

- **Is a given urban activity (Food, Work etc.) the most prominent activity in a geographic neighbourhood or, instead, the area is dominated by some other activity?** We design this problem as a *binary classification* task

and we assess the difficulty of detecting distinct urban activities. We discover that it is easier to predict the *Nightlife* and *Entertainment* areas in a city, whereas *Academic* and *Shopping* spots are those in which the discriminative power of the classifiers presents the worse performance.

- **Assuming the existence of candidate set of urban activities that may take place at an area, which is the most popular one?** We test the prediction accuracy of supervised learning algorithms by exposing them to a multi-class classification scenario, where a single urban activity has to be elected amongst a set of those. We show how the prediction problem becomes increasingly difficult as we increase the number of classes. Further, we note that as we move beyond four or five candidate classes the prediction accuracy of all classifiers drops below $50(\%)$, but the gains over a random guess persist at all instances of the multi-class classification problem.

We expect the findings of the present work to provide insights that could benefit applications such as urban planning through the automatic characterization of urban activities in cities. Moreover, to our knowledge this is one of the first attempts to bring together, in the context of a well defined machine learning task, cellular data and a dataset collected from a mobile social network such as Foursquare. In the next section we perform an analysis on the usage patterns of these two datasets. Next, we define the urban activity prediction task and we present a list of features to encode the telecommunication signal processed by BTS towers. Finally, we proceed with the evaluation of a diverse set of machine learning algorithms and we close with references to related work and conclusions.

II. Supervised Urban Activity Inference

In this section we provide a formulation for the urban activity inference task. Given a geographic area in a city our goal is to infer the type of activity carried out by people nearby. We formulate this problem as a supervised learning prediction task where the input vectors are designed according to the telecommunication signal at the BTS tower level, whereas we consider the popularity of nearby Foursquare places as a proxy to characterize human activity.

A. Problem Formulation

We now define the urban activity inference task given a set of input towers T and places P in a city. Formally, for an area a_i associated to tower i, we represent the communication patterns occurring within the range of the tower using a multi-dimensional input vector $\mathbf{x_i}$. Our goal then is to infer the most prominent urban activity y_i associated to the area. We assume that y_i is the label of the Foursquare venue category z that features the highest number of check-ins in geographic area a_i. For instance if three place categories are observed in the area, namely *Work, Food and Shops*, with check-ins 380, 200 and 120 respectively, then y_i is set to be equal to *Work*. Considering the eight general types of places in Foursquare that we introduce and analyze in the next section, we can define the task of inferring the activity of an area as a supervised classification problem where we seek to learn a function f such as

$$y_i = f(\mathbf{x_i})$$

whose input is the tower's multi-dimensional communication signal and output a class label that corresponds to the inferred most prominent nearby activity. In other words, we exploit quantitative telecommunication patterns observed through BTS towers to identify qualitative aspects of activity in the area. In Figure 1, we depict the spatial distribution of BTS towers and Foursquare venues observed in one of the two cities we focus on in the present work, Barcelona. As formulated above, each tower will be associated to nearby places. Due to the irregular geographic distribution of the place and tower entities, imbalances may exist. For example, if two BTS towers are very close to each other, then the Foursquare place labels associated to them will be different, while the towers could cover effectively the same urban area in terms of the activity to be inferred. We will discuss in detail and tackle this issue in Section IV, where we describe a technique that we name *tower aggregation*.

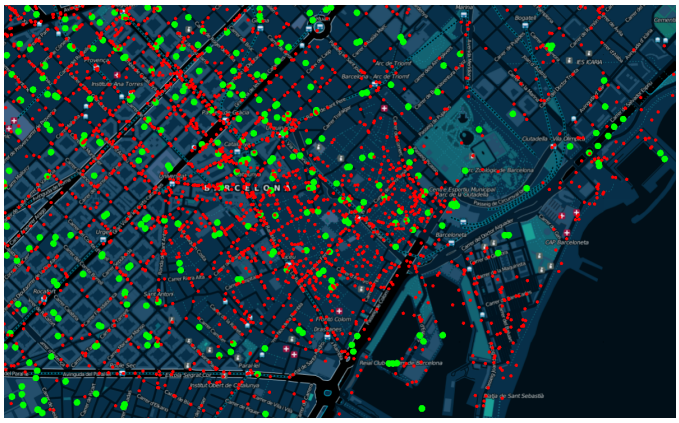

Fig. 1. Geographic distribution of BTS Towers (large green circles) and Foursquare venues (small red circles) in the center of Barcelona. The present work aims to mine the telecommunication signal received by BTS towers and associated with the semantic annotations of nearby Foursquare places.

Finally, during evaluation we will discuss two variations of the supervised learning problem depending on the type of application question we are interested in answering. Next we provide an analysis driven discussion of the two mobile datasets employed in this work, whereas the features we have mined in order to build a tower's input signal $\mathbf{x_i}$ are being formulated in Section IV.

III. Mobile Dataset Analysis

In this section we analyze the two datasets we employ in the present work. Initially an overview of our Call Detail Record from Spain is presented, followed by the Foursquare venue dataset. Further, we discuss the most important facets of the data in light of the urban activity chacterization problem.

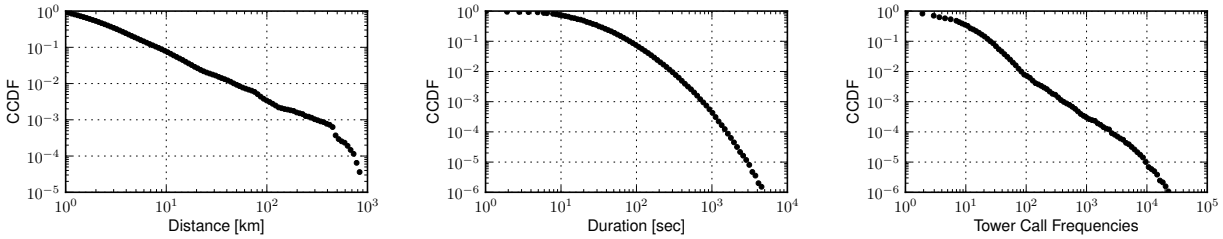

Fig. 2. Complementary Cumulative Distribution Functions of the Call Detail Record dataset: In (a) and (b) the distance and duration distributions of calls are shown respectively whereas in (c) we plot the distribution of the number of sightings (total call frequency) for each tower.

A. Call Detail Records

Call Detail Records (CDRs) are data records collected by telecommunication providers when cell phone users use one of their services, most commonly when they initiate or receive a *voice* call. Each time a user participates in a telecommunication interaction, his or her position can be approximately inferred by knowing the geographic coordinates of the nearby BTS tower that has processed the call. Since every interaction in these systems involves a pair of users, each record can be processed to retrieve the position of two users. There are two possibilities in this case: users could be interacting under the same BTS tower (users are at same area) or their call may be handled by two different towers (users are at different areas). We have processed a dataset of Call Detail Records spanning a month's duration in September 2009 collected by a large telecommunication provider in Spain, through approximately 100 thousand BTS Towers deployed on the country's territory. Twenty million timestamped calls (together with text messages) were recorded by twelve million users with per second temporal granularity. Further, the duration of each voice call was recorded.

In Figure 2(a) we plot the Complementary Cumulative Distribution Function (CCDF) of call distances in the dataset. The log-log plot shows that the distribution follows a power-law functional form with the probability of encountering a certain value to decrease as distance increases. Interestingly, more than 90% of calls occurs within a distance of $10km$ suggesting that a large amount of activity in those systems is concentrated within the urban boundary of a city. This is particularly important for this work as we concentrate our experiments on the two largest cities of Spain, Madrid and Barcelona, where call records account for approximately 70% of the country. In Figure 2(b) we present the CCDF of call durations. About 90% of durations are less than 100 seconds and 1% of calls last more than 5 minutes. We will show in the next section how we exploit information both on call distances and durations to build machine learning *features* for supervised learning classifiers. Next, in Figure 2(c) the CCDF of call frequencies per BTS tower during a period of 24-hours is shown. We can observe large heterogeneities in the volume of activity at each tower, as the frequency values span large orders of magnitude. A small number of towers, potentially those positioned at the centers of large cities, deal with more than 10 thousand calls daily. Finally, with respect to the CDR data, in Figure 3 we show the temporal evolution of call frequencies during the course of a week. The call frequency signal extracted from cellular usage patterns presents strong periodicities on a daily basis. Moreover, there are two peaks in the frequency of calls during lunchtime and dinner times. Telecommunication activity drops during the weekends, but differently to the rest of the days, dinner time activity is a peak, especially on Sundays. We think that those temporal heterogeneities may signal different types of urban activities carried out by mobile users as they shift their behavioral patterns over time. We aim to exploit this information in the present work by mining features that characterize BTS towers according to the times of the week they are mostly active.

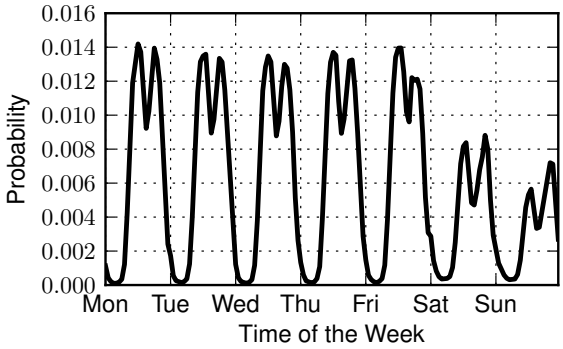

Fig. 3. Temporal evolution of call frequencies aggregated over the period of a week. Strong periodicities together with variations between weekdays and weekends are observed.

B. Foursquare Venue Dataset

So far we have analyzed only one of the datasets we employ in the present work. While we will be using Call Detail Records to encode the telecommunication *signal* of a BTS tower, the information we use as a proxy of urban activity for a given geographic area is sourced from Foursquare. The service was created in 2008 and since then has attracted more than 20 million users [30] who *check-in* at venues around the world informing their social network on their whereabouts. Despite that the application was launched in the form of a game where users could become mayors of a place if they had the highest number of *check-ins* there, it has now evolved to a large social network and a mobile recommendation engine focused around physical places.

During September 2010, we collected information for about 2.5 million Foursquare venues spread around the globe. For each venue, or place, we are aware of its geographic coordinates. Further, category information about each place has been crowd sourced by Foursquare users. Thus a venue has been associated with a semantic label that signifies its type. There are two hierarchical level of places categories in Foursquare [4], a more general one that describes the place in an abstract way (for instance Food) and a more specific one (for instance Paella Restaurant). In the context of the present work we will use the more general category hierarchy for which we have eight possible variations: *Arts and Entertainment, College And Education, Food, Work, Nightlife, Parks and Outdoors, Shops* and *Travel Spots*. Moreover, each venue in the Foursquare dataset contains information on the total number of check-ins that its users have carried out since the inception of the service. Here, we will exploit information on the popularity of Foursquare venues and their categories in order to characterize geographic areas in Madrid and Barcelona, covered by BTS towers. Our assumption is that Foursquare check-in activity can be used to identify the type of urban activity occurring in a city's neighbourhood.

In Figure 4 we plot the Complementary Cumulative Distribution Function of the *check-in* number observed at each Foursquare venue. We observe large variations in the popularity of venues with most of them concentrating a small number of *check-ins* and a small minority to concentrate a large number of user visits in the order of thousands. The bias in popularity of Foursquare places is relevant to this work as this is already an indication that the intensity of user activity may differ significantly across space. To shed further light in this direction, we analyze the way place and *check-in* popularity varies when considering different Foursquare categories. First, in Figure 5(a) we show the number of places per category in Foursquare. Most places, approximately 630 thousand, are associated with *Food*, a category which includes restaurants and coffee shops. Next, *Work* and *Shops* categories also feature a large set of places, while the rest of venue types (*Travel, College, Arts and Nightlife*) have a significantly smaller number of settlements. The striking observation however, comes when we shift our view to Figure 5(b), where the number of *check-ins* per category is shown. The most popular Foursquare category is *Travel* with more than 80 million *check-ins*, at least two times higher than any other category. The ratio between places and *check-in* number for this particular category signifies that a large number of *check-ins* is focused on very popular Travel spots, such as airports and train stations as also has been demonstrated previously [22]. As mentioned in Section II, for a particular area a_i, the corresponding urban activity label y_i is extracted by considering the *check-in popularity* of a Foursquare venue category in the area (Figure 5(b)), as opposed to the *venue popularity* which would enumerate simply the number of venues for a category in a_i.

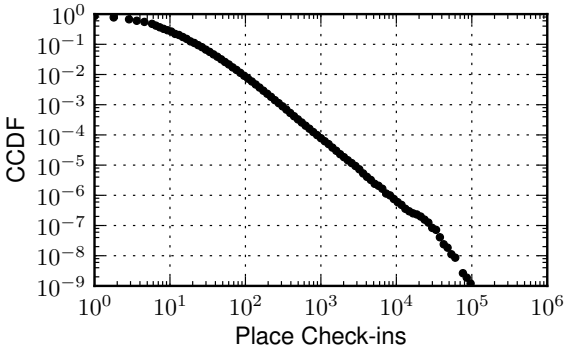

Fig. 4. Complementary Cumulative Distribution Function of place check-ins in Foursquare.

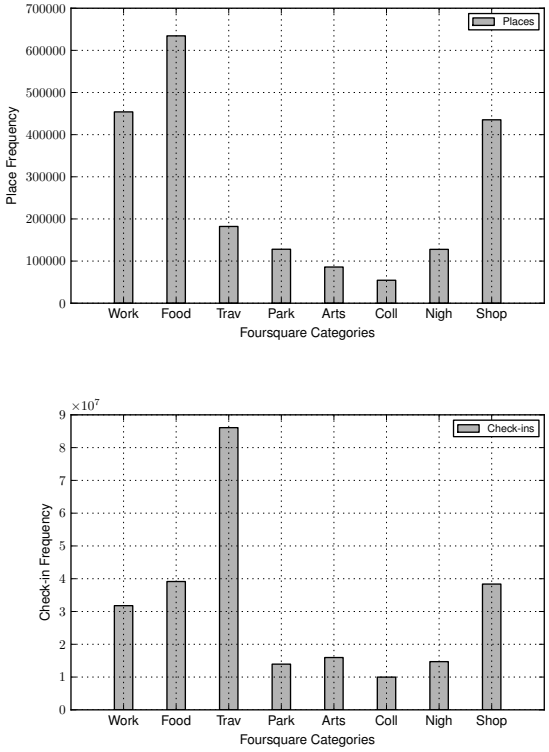

Fig. 5. Frequency of places and check-ins for each category in the Foursquare dataset.

IV. Mining User Telecommunication Activity

In this section we describe and formulate the machine learning features mined by analyzing the user telecommunication activity around BTS towers. All features for a given tower i are combined in a vector $\mathbf{x_i}$ that represents the input telecommunication signal of the area a_i covered by tower i. Supervised learning algorithms will then be applied on the signal vectors to infer the activity y_i in the area. Finally, in this section we introduce a BTS tower aggregation technique that aims to combine the signal of neighboring towers in order to improve the quality of the learning task.

A. Learning Features

User Communication Entropy: We represent with c_{ik} the number of calls by user k at tower i during the measurement period. Then, by considering the total number of calls at tower i, $|C_i|$, we define the user's proportion of calls at the tower as $p_{ik} = \frac{c_{ik}}{|C_i|}$. Our goal is here to capture how call records at a tower are distributed across users. Thus, we define E_{ik} as the entropy of the distribution of p_{ik} values, taking into account all n users that are *seen* at a tower.

$$E_{ik}^{users} = -\sum_{k=1}^{n} p_{ik} \log p_{ik} \tag{1}$$

A high entropy value implies that communication activity at the tower is distributed across many users who have very few calls, whereas smaller values can be interpreted as signs of more stable and periodic user behavior. For instance, *Work* environments may involve a lot of users returning to those regularly, whereas *Arts and Entertainment* areas may attract many opportunistic visitors.

Outgoing Tower Entropy: In a similar spirit, we define the entropy of a BTS tower, by considering its interaction with other towers and measuring the frequency of outgoing calls to them. While this feature is potentially correlated to *user communication entropy*, it is more informative about *where* calls are placed to or received from. For instance, one could expect that *Travel spots* or *Work* areas may have more calls towards many different towers, whereas in *Entertainment* areas people are often calling each other to synchronize before they meet and may constrain a lot of their calls to a subset of proximate towers. Formally, we measure the proportion of calls $p_{ij} = \frac{c_{ij}}{|C_i|}$, where c_{ij} is the number of calls initiated by tower i towards tower j and subsequently we define the corresponding entropy score for m outgoing towers as

$$E_{ij}^{towers} = -\sum_{j=1}^{m} p_{ij} \log p_{ij} \tag{2}$$

Remote Communications: In order to explicitly incorporate geographic distance as a feature, for each tower we measure the geographic distance to the contact tower where the outgoing (resp. incoming) call was initiated. Hence, for each tower, we have a set of distance samples which we employ to characterize it geographically. As we have shown in the analysis section, the distribution of distances follows a power-law trend so an average value would not correspond to a valid characterization. Instead, we define a threshold value d_0 and we measure the volume of calls above that threshold. Formally we define the distance of a call c_{il} at tower i as $dist(c_{il})$ and the corresponding feature as

$$\frac{|\{c_{il} \in C_i \ s.t. \ dist\,(c_{il}) > d_0\}|}{|C_i|} \tag{3}$$

Our goal has been to identify remote calls to other cities and we have found 10 km to be a good threshold that would allow us to distinguish between calls within the urban borders to those destined further away. *Travel* areas which attract visitors from other cities may be a good candidate activity class favored by this feature.

Weekend Calls: It is intuitively expected that *Nightlife* areas should be more prone to gather crowds during weekends, whereas *Work/Office* spaces may be, in most cases, weekday attractors. Consequently, the call volume distribution over time for these areas will also be analogous. We then define a feature that captures specifically this behavioral property of cell phone users. We define the day of a call l at BTS tower i as $day(c_{il})$ defined as

$$\frac{|\{c_{il} \in C_i \ s.t. \ day\,(c_{il}) \in \{Saturday, Sunday\}\}|}{|C_i|} \tag{4}$$

and measure the fraction of CDR records of the BTS tower that are observed during weekends.

Night Time Call Volume: Similarly, we are capturing the signal of night time calls by measuring the corresponding ratio of night time versus the total number of calls. This is formally achieved by setting the hour of a call c_{il} as $hour(c_{il})$ and defining the *nightntime call volume* as

$$\frac{|\{c_{il} \in C_i \ s.t. \ hour\,(c_{il}) > 6p.m. \wedge \ hour\,(c_{il}) < 4a.m.\}|}{|C_i|} \tag{5}$$

where by night time call we define any call that took place after 6pm and before 4am. We note that, alternatively to the above definitions, we could have an even more granular representation of the temporal communication signal, if for instance we were using vectors noting the frequency of calls at each hour slot of the week. Considering time with this level of accuracy, however, would dramatically increase the dimensionality of the input signal. When we informally experimented with this scenario we observed a large drop in the prediction accuracy of our classifiers.

Durations: The duration of calls may be another proxy to infer human activity. People tend to have long calls when they are talking to close friends and relatives or when discussing work related subjects. Aiming to separate short from long calls, assuming that this may reflect to different types of activity, we set a threshold dur_0 of *three* minutes and we define the corresponding feature as

$$\frac{|\{c_{il} \in C_i \ s.t. \ dur\,(c_{il}) > dur_0\}|}{|C_i|} \tag{6}$$

where $dur(c_{il})$ is the duration of the call.

User Return Times: Above, we have defined the entropy value of the frequency distribution of user calls. This feature represents the tendency of cell phone users to return to the area within the scope of the BTS tower, but it does not provide information about *when* users return there. Food and nightlife areas may be opportunistically visited by users, but another aspect of human behaviour is *habit*. Thus, despite not being bound to return to our favourite restaurant, we may still visit it frequently, if we appreciate its service, for instance. Thus we define the volume of return times over nighttime and weekends respectively, by noting a *return event* of a user at tower i as

Fig. 6. Depicting the affect of aggregating BTS towers to Super-Towers that cover larger geographic areas. We have clustered the geographic coordinates of BTS towers to yield spatial centroids covering larger patches of land. Each voronoi cell at the tessellations is colored with the most popular activity proxied by exploiting the popularity of nearby Foursquare places. This figure is better inspected in full color mode.

r and the time of the event as $time(r)$, as

$$\frac{|\{r \in R_i \ s.t. \ time(r) > 6p.m. \wedge \ \ hour(r) < 4a.m.\}|}{|R_i|} \quad (7)$$

and

$$\frac{|\{r \in R_i \ s.t. \ day(r) \in \{Saturday, Sunday\}\}|}{|R_i|} \quad (8)$$

where R_i is the total number of *return events* at tower i.

B. Tower Aggregation

In addition to the extraction of features required for the encoding of a tower's telecommunication signal, we have performed the pre-processing step of aggregating geographically nearby towers into virtual super-towers. The aim of this step is two-fold. First and most importantly, in very dense urban environments there is a very large number of telecommunication towers which segregates the geographic space into very small partitions. Sometimes two or three BTS towers may be almost stuck on each other to serve an area, as it has also been hinted in Figure 1. The effect of this is that for every tower there are very few Foursquare places associated to it, thus extracting labels from such a small sample has resulted to noise. Second, through aggregation, we combine the signal of nearby towers and enhance the quality of the input signal which is critical to the learning task.

In order to perform tower aggregation, we have exploited a density-based spatial clustering algorithm named DBSCAN [9]. The algorithm accepts as input the BTS towers in the city encoded through their latitude and longitude coordinates. After clustering the input points in the 2-dimensional space defined by geographic coordinates, it returns clusters of geographically proximate towers. We define the centroid of each cluster to be a super-tower, that is, a tower whose input signal $\mathbf{x_i}$ is defined by aggregating the call detail records of the towers that have been grouped together by the algorithm. In Figure 6 we show the effect of aggregating towers in the city of Madrid. On the left (Figure 6(a)), we observe the Voronoi tessellation generated the original set of BTS towers in the city without considering any aggregation mechanism. On the right hand side (Figure 6(b)), we show the tessellation after the aggregation step has been applied. There are two main observations. First and by definition, the areas emerging after aggregation are larger. Second, the distribution of popular activities at areas in the city changes (see different colors mapping urban activities). Indeed, the way we segment the geographic plane plays an important role on the characterization of areas, since the underlying distribution of places per BTS tower and geographic area alters. During evaluation (Section V), we will discuss the results for the urban activity prediction task with tower aggregation, as noise becomes an obstacle to the learning process when the aggregation mechanism is not enabled with accuracy dropping to the levels of a random guess.

V. Evaluation

Below we demonstrate the results obtained after exposing a diverse set of supervised learning algorithms to the training and test datasets obtained after the generation of features and labels shown previously. We consider two scenarios. First we examine how easy it is to predict whether a specific Foursquare activity is prominent in a geographic area, whereas in the second scenario we consider the potential existence of multiple urban activities and we seek to predict the most popular one.

A. Algorithms, Methodology and Metrics

We have chosen to evaluate a host of classification algorithms, each one being a representative of a different school of thoughts from the machine learning and artificial intelligence communities. In particular, we examine a simple Logistic

Regression model which assumes a linear relationship amongst the *training features* and performs regression analysis for predicting the outcome of a categorical variable. In addition, we employ a Support Vector Machine (SVM) classifier [10]. SVMs have been successfully employed in a number of machine learning tasks and they operate by maximizing the margin between the nearest data points of two classes and the hyperplane that separates them. From the field of neural networks we select one of the well-known representatives, the Multilayer Perceptron [25]. The perceptron is a network of artificial neurones, which exploits a supervised learning technique known as *backpropagation*, to adapt their weights with other nodes in the network. Moreover, we consider the class of Decision Tree learning and use Logistic Model Trees [18]. This supervised model combines logistic regression with tree induction. Each leaf in these trees corresponds to a linear model that is used for classification. Finally, we also exploit a Bayesian Network based classifier [28], referred here as *DMNBText*, which exploits discriminative learning to infer the parameters of a Bayesian Network that is used for classification. We have invoked the algorithms above through the weka machine learning framework [14] using default parameters. For each classification task we have performed a 10-fold cross validation on a balanced training set averaging the accuracy results obtained in Madrid and Barcelona. In each prediction task a random predictor would correspond to flipping a *coin* with probability $\frac{1}{K}$, where K is the number of classes.

B. Binary Urban Activity Inference

Formulation: As a first step we are interested in understanding whether we can predict a *single* most prominent activity for a given area in the city. More specifically we ask the following:

- *Is Food (resp. for every activity) the most popular activity in an area or some other activity?*

We formulate this question as a binary classification task and we generate a distinct supervised learning task for each of the eight activities we examine. In each prediction task, every area in the city where the activity in question is the most popular is marked with the corresponding class name (for instance *Food*), whereas all remaining areas are marked with the label *Other*. Then, a classifier's goal is set to predict accurately in which areas, the activity we are interested in is the most popular, and in which cases is not, based on the feature vector of the BTS tower input signal $\mathbf{x_i}$.

A potential example application of such an approach could be related to urban planning. For instance, if municipal authorities are willing to discover which areas are popular Nightlife hotspots they could exploit the existing telecommunication infrastructure to identify those and then build related services. Here, in terms of experimental evaluation, we would like to understand firstly whether certain activities are easier to predict than others, and secondly, how do machine learning classifiers perform when they are invoked to infer each of the individual activities.

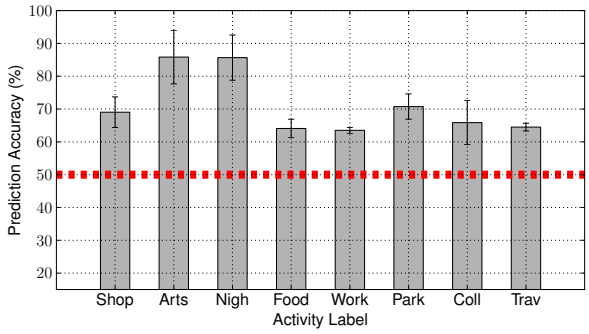

Fig. 7. Average Prediction Accuracy for different Foursquare activities. The error bars correspond to standard deviations in the performance of the five classifiers over each class.

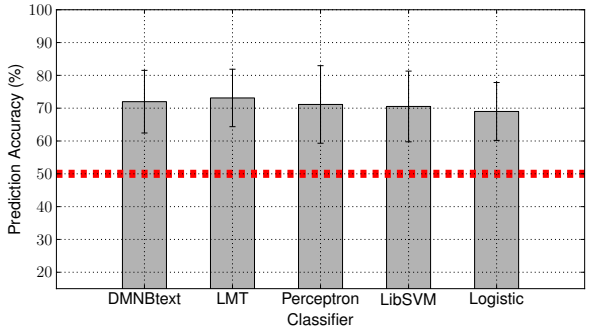

Fig. 8. Average Prediction Accuracy of the five classifiers for the binary classification task.

Results: In Figure 7 we present the prediction performance achieved for each one of the Foursquare activities. The best results have been achieved for *Nightlife* and *Arts and Entertainment* activities, both averaging 86%. They are followed by *Shopping* and *Parks and Outdoor* areas, where almost 70% of the test instances is predicted correctly. The accuracy scores obtained for the rest of the activities are in the range between 60% to 65%, yet offering at least 20% improvement over a random guess. Next, in Figure 8 we observe the performance of individual classifies averaged over all the prediction activity tasks (i.e., considering all different class labels). While it is not easy to identify a dominant algorithm, Logistic Model Trees (LMT) achieve peak performance followed closely by the *DBNMText* Bayesian Network approach. The Logistic Regression classifier trails behind with an accuracy score just below 70%.

An analytical overview of the 10-fold cross validation scores achieved by each classifier for every Foursquare activity is shown in Figure 9. It is interesting to observe that, for certain activities, supervised learning algorithms present high variability. Arts and Entertainment areas, which include Museums and Art Galleries have been predicted with perfect accuracy by the Multi-layer Perceptron (100%), when Logistic Model Trees for the same task fall below 80%. On the other hand,

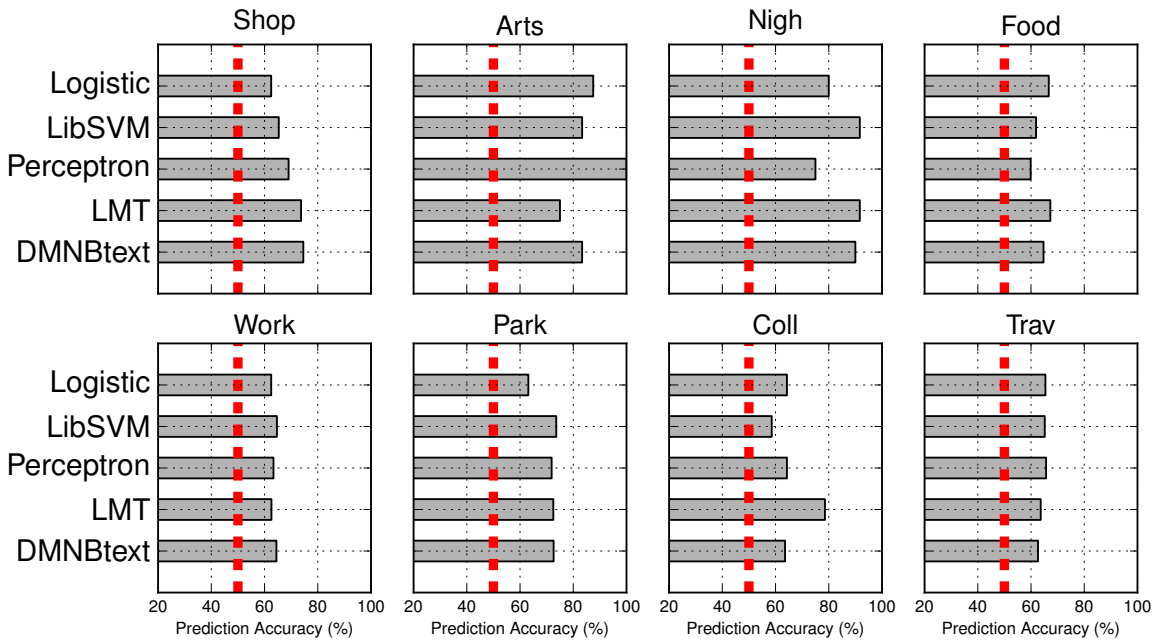

Fig. 9. Average Prediction Accuracy of the five classifiers across the different urban activity classes for the binary classification task considering a 10-fold cross validation approach.

classifier scores for *Work* and *Travel* areas are quite stable at almost 62%. Notably, Logistic Model Trees do very well on the *College and Education* category, reaching an 80% accuracy when the rest of the algorithms are bound well below 65%. Overall, it is interesting to observe that different urban activities may be more predictable by different classifiers, although there is no clear winner across all cases. A potential explanation for this behavior could be deduced if one bears in mind that different supervised learning models make different assumptions about the relationships of the input features (for instance linear versus non-linear models). It could be the case then that for certain types of activities, different model assumptions may prove to be better than others. Moreover, from an application point of view, it appears that the type of activity one is willing to infer may greatly affect the expected accuracy. Whether some types of areas are easier to predict than others may depend on a number of factors, such as the fact that the information retrieved by the telecommunication signal as seen through Call Detail Records has inherent limitations with respect to the inference task, or, from a data mining perspective, that the *features* we have introduced in the present work are biased towards a specific set of classes. Our approach to exploit the quantitive signal of user communication activity to infer qualitative types of urban activities may at first glance appear ambitious, however, the prediction accuracy of the classifiers we have experimented with demonstrates that the inference task is feasible to solve and could have interesting practical applications.

C. Multi-Class Urban Activity Inference

Formulation: In the previous section we have illustrated experimental results for the prediction task on whether a certain urban activity is the most popular in an area or not. An alternative and perhaps harder exploration would be to infer which activity is prominent in an area, by considering a potential candidate set of those. We pose the following question:

- *Which is the most popular activity in a given area in a city?*

This question can be formulated in machine learning terms by designing the corresponding multi-class classification version of the problem, where a classifier is trained to infer one class amongst a set of K potential classes $\{C_1, \ldots, C_K\}$. We consider different values for K ranging from 2 to 8 classes and we are interested in comparing the performance of the different classifiers for each K. For each prediction task we elect the top-K categories in terms of the number of areas in the city that a corresponding category is most popular in, and subsequently we train and test our classifiers to predict across those. We note that the instance of the binary classification task examined here (for $K = 2$) is different from the ones examined in the previous section: now we aim to assess the discriminative power of the supervised learning algorithms amongst the two most popular classes, instead of picking a single class out of eight and asking whether an area features this class or *any* other.

Results: In Figure 10 we demonstrate how the prediction accuracy of the supervised learning algorithms evolves with respect to the number of classes K incorporated in the

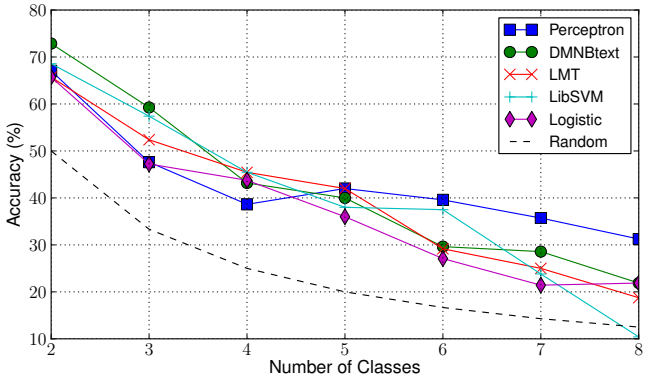

Fig. 10. Performance of the five supervised learning classifiers in the multi-class classification task. Here we depict the accuracy of the algorithms for varying number of classes K, using a 10-fold cross validation approach.

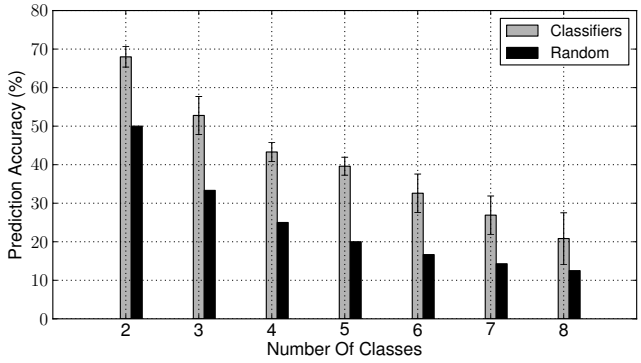

Fig. 11. Average accuracy of the five classification algorithms compared to randomly guessing the class label. It can be observed that despite the significant drop in prediction accuracy, the input telecommunication signal offers clear gains over a random choice.

classification task. It can be observed that accuracy is high and above 65% when $K = 2$. In that case *DMNBText* is leading in performance with at least 5% margin with respect to the rest of the classifiers. The dominance of *DMNBText* remains also for $K = 3$ where *Support Vector Machines* are performing well too, with the *Multi-layer Perceptron* and *Logistic Regression* to drop heavily in performance, scoring below 50%. For number of classes $K = 4$ or more, all classifiers can predict successfully only less than half of the instances.

For large K values, the *Multi-layer Perceptron* presents high resilience in accuracy terms, maintaining a score much higher to the rest of the algorithms that is only being matched by that of the *Support Vector Machines* when $K = 6$. For a maximum number of classes $K = 8$ which is the total number of Foursquare place categories, the perceptron improves performance almost by 50% when compared to the rest of the algorithms. To sum up, it is in principle expected that a classification task becomes harder as we increase the number of classes K. The number of outlier instances increases (noise) together with the complexity of the problem of identifying effectively class boundaries. Despite that, the performance of the classifiers drops significantly for large Ks, as we demonstrate in Figure 11, the gain achieved from the algorithms and telecommunication signal features is significantly higher over a random guess in all cases.

VI. Related Work

Due to the pervasiveness of cell phones and the popularity of mobile social media a variety of works that focus on characterzing urban environments using Call Detail Records (CDR) or mobile social networks have emerged.

Focussing on CDR, a seminal work by Ratti et al. [23] used aggregated cell-phone data to analyze urban planning in Milan with an interest in location-based services and related applications. The work presented in [24] monitored the urban dynamics in the city of Rome and obtained clusters of geographical areas measuring cell phone towers activity using Erlangs. Another study is described in [27], where CDR is used to characterize and cluster land uses based on cell phone activity. A similar work is presented in [15], where the authors analyze four different geographical spots at different times in Bangkok and identified correlations between land use and CDR activity. Finally, CDR has been extensivelly used to characterize mobility within urban areas, as in Isaacman *et al.* [16], which analyzed and compared daily movements of people in New York City and Los Angeles.

Regarding social media, services such as Twitter, Flickr or Foursquare, have been used to study and characterize urban environments exploiting geo-tagged information. Focusing on Twitter, Wakamiya *et al.* [31] and Fujisaka *et al.* [13] have used geo-tagged Twitter datasets and its semantic content to study and characterize crowd mobility. Similarly, Kinsella *et al.* [17] used geo-located tweets, together with their content, to create geographic language models at varying levels of granularity (from zip codes to countries). The authors use these models to predict both the location of the tweet and the user based on linguistic content changes. Foursquare has been used by Noulas *et al.* [21] to model crowd activity patterns in London and New York City using spectral clustering. Finally, Flickr has been the focus of the work by Crandall *et al.* [11], which used a dataset of geotagged photos from Flickr to perform world landmark localization using the mean-shift algorithm.

Although the literature presents a large variety of studies, to the best of our knowledge, no previous work combines both CDR and mobile social media to characterize urban environments. From this perspective, the results we have presented combine the capabilities of both datasets for studying and characterizing urban landscapes.

VII. Conclusion

In this paper we have proposed two approaches for the characterization of urban environments. Our framework leverages the telecommunication signal of BTS towers as seen through Call Detail Records in conjunction with information about the

categories and popularity of Foursquare places.

We have formalized this problem as a supervised learning classification task where the goal is to infer a learning function that associates the BTS tower signal to Foursquare semantic labels distributed geographically in the nearby area. We have considered two alternative approaches of classification: a binary classification task where a classifier is called to discriminate whether a given urban activity is the most popular at an area of the city, and a multi-class classification task where the most popular activity in a neighbourhood has to be elected amongst a set of a many. In both cases we have evaluated the performance of five different supervised learning classifiers considering two Spanish cities, Madrid and Barcelona.

In addition we have performed an analysis on the telecommunication usage patterns of millions of cell phone users and offered insights about the distribution of Foursquare categories and their popularity in terms of check-in number. We have also built a set a machine learning features that may be exploited to infer the qualitative characteristics of urban areas through cellular data.

In terms of future work, we are aiming to exploit cellular data to infer a mixture of activities in a neighbourhood rather than a single one. In addition, the prominence of user activities may change over time in a given area, thus exploring the temporal dynamics of urban activity evolution is another interesting direction to be explored in the future.

REFERENCES

[1] CASA Smart Cities: bridging physical and digital. http://www.bartlett.ucl.ac.uk/casa/events/2012-04-20-Conference.

[2] European Smart Cities. http://www.smart-cities.eu/.

[3] Foursquare application. http://www.foursquare.com/.

[4] Foursquare Venue Categories. http://aboutfoursquare.com/foursquare-categories/.

[5] IBM Smart Cities. http://www.ibm.com/smarterplanet/us/en/smarter_cities/overview/index.html.

[6] London to test 'smart city' operating system. http://www.bbc.co.uk/news/technology-17940797.

[7] MIT Smart Cities. http://cities.media.mit.edu/.

[8] Smart Santander. http://www.smartsantander.eu/.

[9] A density-based algorithm for discovering clusters in large spatial databases with noise, 1996.

[10] C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27, May 2011.

[11] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 761–770, 2009.

[12] Facebook. Building better stories with location and friends. http://developers.facebook.com/blog/post/2012/03/07/building-better-stories-with-location-and-friends, March 2012.

[13] T. Fujisaka, R. Lee, and K. Sumiya. Exploring urban characteristics using movement history of mass mobile microbloggers. In *Proceedings of the Eleventh Workshop on Mobile Computing Systems & Applications*, pages 13–18. ACM, 2010.

[14] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, Nov. 2009.

[15] T. Horanont and R. Shibasaki. Evolution of urban activities and land use classification through mobile phone and gis analysis. In *CUPUM*, 2009.

[16] Isaacman, S. and Becker, R. and Cáceres, R. and Kobourov, S. and Rowland, J. and Varshavsky, A. A tale of two cities . In *Proceedings of the Eleventh Workshop on Mobile Computing Systems & Applications*, pages 19–24, 2010.

[17] S. Kinsella, V. Murdock, and N. Oare. I am eating a sandwich in glasgow modeling locations with tweets. In *Proc. of the 3rd Workshop on Search and Mining User-generated Contents, Glasgow, UK*, 2011.

[18] N. Landwehr, M. Hall, and E. Frank. Logistic model trees. *Mach. Learn.*, 59(1-2):161–205, May 2005.

[19] N. Lathia, S. Ahmed, and L. Capra. Measuring the impact of opening the London shared bicycle scheme to casual users. 22, 2012.

[20] N. Lathia and L. Capra. Mining mobility data to minimise travellers' spending on public transport. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 1181–1189, New York, NY, USA, 2011. ACM.

[21] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. In *3rd Workshop Social Mobile Web (SMW 2011)*.

[22] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. An empirical study of geographic user activity patterns in foursquare. *ICWSM2011*, 2011.

[23] C. Ratti, R. M. Pulselli, S. Williams, and D. Frenchman. Mobile landscapes: using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design*, 33(5):727–748, 2006.

[24] J. Reades, F. Calabrese, A. Sevtsuk, and C. Ratti. Cellular census: Explorations in urban data collection. *IEEE Pervasive Computing*, 6(3):30–38, 2007.

[25] F. Rosenblatt. Principles of neurodynamics: Perceptrons and the theory of brain mechanisms. 1961.

[26] C. Roth, S. M. Kang, M. Batty, and M. Barthlemy. Structure of urban movements: Polycentric activity and entangled hierarchical flows. *PLoS ONE*, 6(1):8, 2011.

[27] V. Soto and E. Frias-Martinez. Robust land use characterization of urban lanscapes using cell phone data. In *1st Workshop on Pervasive Urban Applications, in conjunction with 9th Int. Conf. Pervasive Computing*, 2011.

[28] J. Su, H. Zhang, C. X. Ling, and S. Matwin. Discriminative parameter learning for bayesian networks. In *Proceedings of the 25th international conference on Machine learning*, ICML '08, pages 1016–1023, New York, NY, USA, 2008. ACM.

[29] Techcrunch. Facebook has acquired gowalla. http://techcrunch.com/2011/12/02/report-facebook-has-acquired-gowalla/, December 2011.

[30] The Next Web. Foursquare hits 20 millions users and 2 billion check-ins. http://thenextweb.com/socialmedia/2012/04/16/foursquare-hits-20-million-users, April 2012.

[31] S. Wakamiya, R. Lee, and K. Sumiya. Urban area characterization based on semantics of crowd activities in twitter. In C. Claramunt, S. Levashkin, and M. Bertolotto, editors, *GeoSpatial Semantics*, volume 6631 of *Lecture Notes in Computer Science*, pages 108–123. Springer Berlin / Heidelberg, 2011.