An Unlicensed Taxi Identification Model Based on Big Data Analysis

Wei Yuan, Pan Deng, Tarih Taleb, Senior Member, IEEE, Jiafu Wan, Member, IEEE, and Chaofan Bi

Abstract—Social networks and mobile networks are exposing human beings to a big data era. With the support of big data analytics, conventional intelligent transportation systems (ITS) are gradually changing into data-driven ITS (D2ITS). Along with traffic growth, D2ITS need to solve more real-life problems, including the issue of unlicensed taxis and their identification, which potentially disrupts the taxi business sector and endangers society safety. As a remedy to this issue, a smart model is proposed in this paper to identify unlicensed taxis. The proposed model consists of two submodel components, namely, candidate selection model and candidate refined model. The former is used to screen out a coarse-grained suspected unlicensed taxi candidate list. The list is taken as an input for the candidate refined model, which is based on machine learning to get a fine-grained list of suspected unlicensed taxis. The proposed model is evaluated using real-life data, and the obtained results are encouraging, demonstrating its efficiency and accuracy in identifying unlicensed taxis, helping governments to better regulate the traffic operation and reduce associated costs.

Index Terms—Big data, intelligent transportation systems, machine learning, data-driven ITS, unlicensed taxi.

I. INTRODUCTION

Along with the rapid growth of urban economies, cities are expanding rapidly and urban populations are constantly increasing. At the same time, urban population floating range are largely increasing and land resources are gradually becoming tense, which results in serious traffic congestions, traffic safety problems and environmental pollution. All these issues have become public concerns. At this point, Intelligent Transportation Systems (ITS) appear as a solution to some concerns [1], [33], [34]. In the past two decades, experts all over the world have made lots of research advancing the concept of ITS to ultimately better manage urban traffic and improve the quality of life of people. Some of the conducted research has largely contributed to ensuring traffic safety [34] and better use of road resources [2], [35], [36]. Although the application of ITS solves some urban transportation problems, to a certain extent, the constant increase in current transportation data does not meet up with the demand for data usage, which is crucial for transportation control. To fully utilize data, the concept of Data-Driven ITS (D2ITS) has emerged [3]. D2ITS has evolved from the conventional ITS but with some differences. First, the conventional ITS systems depend mainly on historical data and people’s real life experience [37]–[39] whereas D2ITS systems tend to use real-time data. Second, in conventional ITS systems, the source of data used is generally limited whilst D2ITS can make good use of all types of data, including vehicle information, sense coil and video images. Along with the ongoing advances in big data analytic techniques, D2ITS has become increasingly powerful providing new solutions to several problems, deemed unsolvable before.

Big data technologies could quickly extract valuable information from large amount and various types of data [4]. Big data are characterized by four typical features: 1) huge volume exceeding TB level to PB level; 2) various data types including videos, photos, location information, and sensor data; 3) low value density and high application value (e.g., within a video used to monitor road conditions, there may be only one or two minutes useful clips); and 4) fast processing speed. In the transportation field, there are four types of data that are frequently used: 1) sensor data (e.g., position data, temperature data, pressure data, images, and RFID); 2) system data (e.g., logs and equipment records); 3) service data (e.g., charging information, internet services, and some other information); and 4) application data (e.g., manufacturers, energy, and transportation). Transportation-relevant data consist indeed of various types and enormous volume which, in turn, need fast processing speed.

One type of transportation data is the pass-records data collected from the Electronic Traffic Bayonet Device (ETBD), which has high density and significant value for data mining and analysis. ETBD [5] integrates advanced optical technologies, image processing technologies and pattern recognition
technologies. It can take photos of each passing vehicle and automatically identifies the license plate, brand, type, speed and some other information relevant to the vehicle. It then stores the information into a database.

The pass-records data have very high values in the field of D²ITS. They can be used for traffic volume prediction, traffic trajectory tracking, and traffic behavior analysis. These data are crucial for the supervision and administration of traffic violations and tracking of criminals. In this vein, unlicensed taxis function like a proper taxi carrying passengers between different locations but without transport business licenses and tax registration certificates [6]. In other words, unlicensed taxis are legally not allowed to carry passengers. Impact of unlicensed taxis can be mainly summarized into four aspects [42], [43].

1. Unlicensed taxis seriously disrupt traffic operations and induce safety issues. Indeed, most unlicensed taxi drivers lack necessary security awareness. They tend to more likely ignore traffic regulations.
2. In case of unlicensed taxis, vehicle safety cannot be always guaranteed. Indeed, checkups of vehicle conditions are not carried out on a regular basis.
3. Given the above, passengers’ safety cannot be guaranteed either. Furthermore, because of unlicensed taxi drivers’ complicated background and relatively low level education, passengers may become subject to drivers’ inappropriate behavior.
4. In case of a traffic accident, passengers’ legal rights cannot be guaranteed as unlicensed taxis are not registered at the transportation department. Once an accident happens, it becomes difficult to trace back the cause.

Conventionally, the police department takes the following actions to identify unlicensed taxis and punishes the drivers.

1. Find out suspicious unlicensed taxis through manually watching amounts of videos in the “City Online Monitoring System”;
2. Collect the drivers’ on-site evidence by entrapment;
3. Make appropriate punishment on the suspects.

This process not only wastes a lot of manpower and time, but also results in having many unlicensed taxis go undetected. It is too passive, and a more proactive method is proposed in this paper.

The objective of this paper is to intelligently identify such unlicensed taxis conducting analysis and data mining on the pass-records data. The solution proposed in this paper will hopefully put an end to the unlicensed taxi phenomenon, saving manpower and time of police officers that, otherwise, have to stay for long durations monitoring for suspicious unlicensed taxis. Although some D²ITS technologies, such as image processing and cloud computing have set up a solid foundation for data generation, data acquisition, and data storage [7], review of existing literature does not reveal any approach suitable for unlicensed taxi identification. In this regard, this paper proposes an algorithm based on big data analysis, integrating Support Vector Machine (SVM) [8], [47] learning, whereby Hadoop [9] and Map/Reduce [10] technologies are adopted as well to generate data. Compared with traditional approaches whereby unlicensed taxis are identified by patrolling police, the proposed method achieves its design goals with high accuracy and efficiency. It shortens the unlicensed taxi identification cycle and saves important amount of manpower, eventually yielding important cost savings in the long run.

The paper is organized in the following fashion. Section II discusses some research work relevant to D²ITS. Section III portrays the approach adopted for data acquisition, data pre-processing, and data modeling. Our scheme proposed for unlicensed taxi identification is detailed in Section IV. Some test results are presented and discussed in Section V. Finally, the paper concludes in Section VI.

II. RELATED WORK

Being an essential field of transportation engineering, ITS consists of six fundamental components; namely advanced transportation management systems (ATMS), advanced traveler information systems (ATIS), advanced vehicle control systems (AVCS), business vehicle management (BVM), advanced public transportation systems (APTS), and advanced urban transportation systems (AUTS). As discussed in [3], whether these six components can be fully realized depends on how much data can be collected and processed into useful information. As social networks and mobile networks are developing at a fast speed, data start to overflow, evolving into the age of big data. With large amount of ITS-relevant data, ITS is gradually changing into the so-called D²ITS. There are two major differences between ITS and D²ITS. First, conventional ITS systems depend mainly on human experiences and historical data, while D²ITS systems are based on real time data [3]. Second, data used by conventional ITS systems are collected from limited sources, whilst D²ITS systems collect data from multiple sources. Based on big data, D²ITS systems exhibit better performance. They are straightforward and practical in real life.

A main component of D²ITS is the Vision-Driven ITS, which employs vision-based devices (e.g., cameras) to collect data for doing research and developing applications. In [11] and [12], a Pedestrian Tracking System is developed to detect the crossing characteristics of pedestrians at intersections. Background subtraction technique, tracked through an inherent cost characteristic function in conjunction with an $\alpha - \beta$ filter, is used. This advanced system introduces an automated computer-vision based approach for collecting pedestrian parameters in real-time. A similar approach dedicated for vehicle detection and tracking is proposed in [13]. In [14], another D²ITS system, called DRIVE Net, is developed for data sharing, visualization, modeling and analysis and that is using data extracted from different sources. The basic intention beneath this DRIVE Net platform is to offer a standard tool to incorporate more datasets from different fields and at the same time, and that is for real-time decision making. In [15], a Web-based truck performance measurement system is developed processing commercial GPS
data and quantifying truck travel characteristics and performance between zones. In [16], a Markov chain based Bayesian decision tree algorithm is devised to extract passengers’ origin information from recorded transit smart card systems, and that is for transit system planning and route optimization. In [17], pedestrian movement monitoring is conducted using static Bluetooth sensors, enabling an automated and cost-effective approach to acquire pedestrian data.

The latest developments in ITS pay more attention to proactive control as opposed to conventional passive control and management [18]. This has yielded Learning-Driven ITS which is based on different schemes. Among all of them, real-time traffic prediction schemes, such as those used for the prediction of travel time [37], [38], vehicle passage time [39], and hazardous locations are vital, and have been proven important for the support of efficient communications for highly mobile users in mobile networks [40], [41]. In [19], a state-space neural network is proposed to conduct short-term freeway travel time prediction. A real-time Kalman filter model to predict travel time by combining historical data with real-time measurements is proposed in [20]. Another cluster-based algorithm to identify the collision hotspots along a freeway is proposed in [21].

Thanks to big data analytic techniques, some problems which were impossible to solve before have solutions nowadays. For example, in populated countries such as China, a potential number of people tend to use fake vehicle licenses to avoid being fined for illegal travel behavior. In the past, data collection infrastructure was not available. Stopping the phenomenon of fake vehicle licenses therefore required important human labor, which was not always effective [22]. Nowadays, traffic flow collection system can be setup to capture the pictures of vehicles and image processing technologies can be subsequently used to extract useful information [22]. With the abundance of data, the fake vehicle license phenomenon can be mitigated to some degree. In this paper, we also aim at solving another “popular” transportation issue, namely unlicensed taxis, and that is using big data analytic technologies.

III. DATA ACQUISITION

A. Dataset Description

In this paper, we use vehicle trajectory, consisting of temporal and spatial characteristics of vehicle movements, to identify unlicensed taxis. The method adopted for getting vehicle trajectory is different from the conventional way which traditionally uses GPS [23]–[25], [46]. Indeed, the conventional way is relatively much more expensive as it requires all vehicles to be equipped with GPS devices. In this paper, vehicle trajectory is basically deduced from ETBD since vehicles have to continuously pass many ETBDs during the course of their motion. A vehicle passing record means a vehicle is captured by some ETBDs and at the same time one corresponding recognition dataset is generated. The dataset includes vehicle passing record ID, vehicle license, vehicle type, collection date, and ETBD ID. On average, there are almost 20 million records generated on a daily basis and each record may consist of 20 attributes. Given the big data nature of the dataset and in order to maintain acceptable data processing speeds, the dataset is filtered removing unnecessary data attributes and keeping only useful attributes as shown in Table I.

In Table I, vehicle license ID (referred to as nid throughout this paper) is the unique identification of vehicle passing record and is a 32 char string. Vehicle license (referred to as vLicense) is the license of the detected vehicle. Vehicle type (denoted by vType) indicates the type of vehicle and can be, amongst many types, police car, full-size car, or small car. Collection date (denoted by vDate) refers to when the corresponding record is generated. ETBD ID (denoted by etbdID) is the unique identifier of the corresponding ETBD device which captures the vehicle and generates the dataset. Vehicle speed (denoted as vSpeed) refers to the speed of the vehicle at the collection date time. Accordingly, one vehicle passing record can be formulated as follows:

\[
r = (\text{nid}, \text{vLicense}, \text{vType}, \text{vDate}, \text{etbdID}, \text{vSpeed}).
\]

ETBD devices are usually installed at fixed locations in the city and can capture all vehicles passing by one or more lanes. Another used dataset is the ETBD dataset which contains several location attributes, as shown in Table II.

In Table II, ETBD ID is the same as the one in Table I. Latitude, longitude and ETBD address indicate the relevant ETBD’s location. This dataset can be formulated as follows:

\[
etbd = (\text{etbdID}, \text{latitude}, \text{longitude}, \text{etbdAddress}).
\]

Merging the two datasets based on the common field etbdID, a vehicle passing record can be formulated as follows, giving all information required for the analysis.

\[
r'_i = (\text{nid}, \text{vLicense}, \text{vType}, \text{vDate}, \text{vSpeed}, \text{latitude, longitude, etbdAddress}).
\]
B. Data Preprocessing

In a single city, there could be several hundreds of thousands of vehicles on the move. They would generate 10 million records (i.e., roughly 30 GBytes) on a daily basis. Efficient pre-processing of this big data becomes then required. In this vein, and in addition to filtering out unnecessary fields from the original vehicle passing record, “bad data,” caused by ETBD deviation, should be eliminated. Bad data can lead to biased results, resulting in wrong conclusions [26]. Typical bad data are those with key fields missing or invalid. There are still less than 1% of vehicle-passing records, denoted by group R, which need more intelligent algorithms to detect. After the elimination of these bad data and on all vehicle passing records, whereby \( r_i = (\text{id}_i, \text{License}_i, \text{Type}_i, \text{cDate}_i, \text{etbdID}_i, \text{vSpeed}_i) \) five kinds of statistics are extracted using distributed systems [27], [44], [45].

i. Counting the number of vehicle passing records in time \( \tau \)

\[
C_{(vpr, \tau)} = \sum (r_i | \text{cDate}_i \in \tau). \tag{4}
\]

ii. For every ETBD, counting the number of vehicle passing records in time \( \tau \)

\[
C_{(vpr, etbdID, \tau)} = \sum (r_i | \text{cDate}_i \in \tau, \text{etbdID}_i = \text{etbdID}). \tag{5}
\]

iii. For every vehicle, counting the number of vehicle passing records in time \( \tau \)

\[
C_{(vpr, vLicense, \tau)} = \sum (r_i | \text{cDate}_i \in \tau, v\text{License}_i = v\text{License}). \tag{6}
\]

iv. Counting the total number of vehicles in time \( \tau \)

\[
C_{(vehicle, \tau)} = \sum (r_i | \text{cDate}_i \in \tau, \text{Group by } v\text{License}). \tag{7}
\]

v. Counting the number of ETBD devices crossed by a vehicle during a time interval \( \tau \)

\[
C_{(etbd, vehicle, \tau)} = \sum (r_i | \text{cDate}_i \in \tau, v\text{License}_i = v\text{License}, \text{Group by etbd}). \tag{8}
\]

vi. Counting the total speed for all vehicle-passing records in time \( \tau \)

\[
C_{(speed, \tau)} = \sum (r_i | \text{vSpeed}_i \in \tau). \tag{9}
\]

C. Data Modeling

In this paper, we use hotspot, namely the distribution of the number of vehicle passing record, as the research object. The location of each ETBD is defined by its latitude and longitude on a map. All ETBDs are classified based on the number of vehicle-passing records using a K-means [28] clustering algorithm based on \( C_{(vpr, etbdID, \tau)} \). All ETBDs are categorized into \( h \) grades, denoted by ETBD\textunderscore heat, whereby \( i \in [1, h] \). Based on the ETBD dataset, we define a new dataset as follows.

\[
etbd' = (\text{etbdID}, \text{latitude}, \text{longitude}, \text{etbdAddress}, \text{ETBD\textunderscore heat}). \tag{10}
\]

Generally speaking, the grade of each ETBD, ETBD\textunderscore heat, remains stable for relatively long periods of time, in the order of months. In order to mitigate data deviation, all ETBD-relevant data are stored in a Hadoop Distributed File System (HDFS) and we analyze them using Pig [29] and Perl [30], which are especially designed for big data processing.

As for traffic, its characteristics differ from a city to another, from a day to another and also within the same day. Indeed, morning and evening rush hours may be different for work days and holidays, and may even differ from a city to another and that is due to many factors such as the city location, residents’ behavior, and city regulation. To model morning and evening rush hours for weekdays and holidays with an acceptable level of accuracy, the following algorithm is devised based on big data analysis and statistics.

i. Count weekdays and holidays in time \( \tau \), denoted respectively by \( T_{weekday} \) and \( T_{holiday} \)

\[
\tau = T_{weekday} \cup T_{holiday}. \tag{11}
\]

ii. Count the number of vehicle passing records for each hour of the day during weekdays based on \( C_{(vpr, \tau)} \)

\[
C_{(vpr, T_{weekday}, hour)} = \frac{\sum_{\text{hour}=j} (r_i | \text{cDate}_i \in T_{weekday}, j \in [1, \ldots, 23])}{|T_{weekday}|}. \tag{12}
\]

iii. Count the number of vehicles for each hour of the day during weekdays based on \( C_{(vpr, vLicense, \tau)} \), i.e., (13), shown at the bottom of the page.

iv. Get the average speed for each hour of the day during weekdays to represent congestion rate using \( C_{(speed, T_{weekday}, hour)} \) divided by \( C_{(vpr, T_{weekday}, hour)} \). The lower the average speed is, the more congested the traffic is.

\[
\text{Average\textunderscore speed}_{weekday} = \frac{C_{(speed, T_{weekday}, hour)}}{C_{(vpr, T_{weekday}, hour)}}, \tag{14}
\]

v. The average speed during holidays can be calculated in the same fashion as \( \text{Average\textunderscore speed}_{holiday} \).
vi. Based on the computed average speed for each hour, we get the morning and evening rush hours for both weekdays and holidays as the periods of time in the morning and evening, respectively, when the average speed takes its lowest values, and that is for both holidays and weekdays.

\[
\text{morning\_rush\_hour} = [\text{mrh\_start\_hour}, \text{mrh\_end\_hour}] \\
\text{evening\_rush\_hour} = [\text{erh\_start\_hour}, \text{erh\_end\_hour}]
\]

\[
\text{mrh\_start\_hour} \text{ and mrh\_end\_hour} \text{ are the start and end time for moring\_rush\_hour, and the same meaning of erh\_start\_hour and erh\_end\_hour.}
\]

IV. UNLICENSED TAXI IDENTIFICATION MODEL

In this section, we will describe in details our proposed Unlicensed Taxi Identification Model (UTIM) based on statistics [31] and machine learning [32]. UTIM consists of two sub-model components, namely i) Candidate Selection Model (CSM) and ii) Candidate Refined Model (CRM). The former one is based on statistics and the latter is based on machine learning, as depicted in Fig. 1.

CSM screens out a coarse-grained suspected unlicensed taxi candidate list according to vehicles’ features, such as vehicle type, vehicle activity level during a day, traversed ETBDs, and days showing similar behavior which can directly get rid of the normal vehicles. This candidate list is subsequently used as input for CRM to get a fine-grained suspected unlicensed taxi candidate list. This model is trained based on the features of the confirmed unlicensed taxis and normal vehicles.

A. Candidate Selection Model (CSM)

Generally, unlicensed taxis keep running in the city for long time periods. They consequently pass high number of ETBDs, resulting in high vehicle passing records. Their vehicle trajectories are usually different from those of normal vehicles, which tend to be predictable [36]–[38]. To reduce false alarms, unlicensed taxis and normal vehicles should be distinguished based on statistics. Obviously, police cars, normal taxis, trucks and vehicles with specific roles can be easily deemed as not unlicensed taxis according to rules of the registered license plate. They can be therefore excluded from the operations of CRM. The CSM model is schematically portrayed in Fig. 2.

The CSM algorithm operates as follows.

a. During a time period \( \tau \), the system develops a vehicle list \( V = \{v_1, v_2, \ldots, v_n\} \) and the number of vehicle passing records of each day \( C_{(vpr, vLicense, day)} \) according to the aforementioned analysis;

\[
C_{(vpr, vLicense, day)} = \sum_{i \in V} (r_i \mid \text{eDate}_i \in \text{day} \cap v_{License}\_i = v\_License, \text{day}_i \in \tau).
\]

b. The system also computes the number of ETBDs traversed during each day \( C_{(etbd, vehicle, day)} \)

\[
C_{(etbd, vehicle, day)} = \sum_{i \in V} (r_i \mid \text{eDate}_i \in \text{day} \cap v_{License}\_i = v\_License, \text{Group by etbd}).
\]

c. The system then filters out specific vehicle types (e.g., police car) according to the rules of license plate from the vehicle list \( V \).

d. \( C_{(vpr, vLicense, day)} \) should be greater than the threshold \( \text{thre\_vpr} \)

\[
C_{(vpr, vLicense, day)} > \text{thre\_vpr}.
\]

e. \( C_{(etbd, vehicle, day)} \) should be greater than the threshold \( \text{thre\_etbd} \)

\[
C_{(etbd, vehicle, day)} > \text{thre\_etbd}.
\]

f. Through the above steps, the system can sort out a vehicle list for each day \( V_{\text{day}} \).

g. The system then counts the total number of times each vehicle appears in \( V_{\text{day}} \)

\[
C_{(vLicense)} = \sum_{j \in V_{\text{day}}} v_j \mid (v_{License} \in j, \text{day}_i \in \tau).
\]

h. \( C_{(vLicense)} \) should be greater than the threshold \( \text{thre\_day} \)

\[
C_{(vLicense)} > \text{thre\_day}.
\]
i. Finally, based on the conditions of inequalities (19), (20), and (22), the system sorts out a coarse-grained list of suspected unlicensed taxi candidates $V' = \{v'_1, v'_2, \ldots, v'_n\}$.

**B. Candidate Refined Model (CRM)**

After the creation of the coarse-grained unlicensed taxi candidate list $V'$, CRM comes into play and that is to further enhance the accuracy of the proposed scheme in unlicensed taxi identification. In CRM, the unlicensed taxi identification is transformed into a classification issue using data mining. For each vehicle, there are only two states; unlicensed taxi or normal vehicle.

1) Sample Selection: CRM is based on machine learning. For that purpose, we use both positive samples and negative samples which are inputs for model training. As normal taxis function similarly to unlicensed taxis, in regard of their time and spatial characteristics, they are used as positive samples of unlicensed taxis. The samples can be described as:

$$S_{ut} = \{v_1, v_2, \ldots, v_m\}. \quad (23)$$

As for the selection of negative samples, according to feedback and experience from police, most of unlicensed taxis are entry-level vehicles which are much cheaper. So, the high-end vehicles which are expensive are used as negative samples as:

$$S_{nv} = \{nv_1, nv_2, \ldots, nv_k\}. \quad (24)$$

2) Feature Extraction: Different features affect the accuracy of the proposed unlicensed taxi identification model. In this paper, features based on time and spatial characteristics are considered. This aims for distinguishing unlicensed taxis from normal vehicles to the maximum extent. Based on our experiments, the considered features are as follows.

i. Number of vehicle passing records averaged over predetermined time periods.

The average vehicle passing records is counted for different time periods as show in Tables III and IV, distinguishing between weekdays and holidays. The time periods are determined according to residents' daily activity times, morning rush hour, and evening rush hours. The feature corresponding to the number of vehicle passing records is then created based on the values obtained during each time period as follows:

$$F_1 = \{f_1, f_2, \ldots, f_{18}\}. \quad (25)$$

ii. Average vehicle passing records for each ETBD grade

As explained earlier, ETBDs have different grades. Unlicensed taxis and normal vehicles traverse different numbers of ETBDs. The former always run for longer times in the city, so their ETBD distribution is dispersed. Whereas, the latter travel within a small area and their ETBD distribution remains stable. The feature of average vehicle passing records for different ETBD\_heat is therefore worth considering. There are $h$ grades in ETBD\_heat, so there are $18 * h$ features in this dimension.

$$F_2 = \{f_{19}, f_{20}, \ldots, f_{18+h} \}. \quad (26)$$

iii. Number of passed ETBDs averaged over predetermined time periods.

Unlicensed taxis and normal vehicles take different trajectories in different time periods. The average number of ETBDs they traverse is also different. With regard to the average number of passed ETBDs over predetermined time periods, there are $18 * h$ features as follows:

$$F_3 = \{ f_{18+h+19}, f_{18+h+20}, \ldots, f_{2+18+h+18} \}. \quad (27)$$

iv. Average vehicle passing records of every day.
Average vehicle passing records of every day is computed for all samples and the features look like:

\[ F_4 = \{ f_{2+18h+19}, f_{2+18h+20} \}. \] (28)

As a result, a total of \( (2 \times 18 + h + 20) \) features are obtained. For instance, in case ETBDs are classified into three grades, there should be 128 features to train the proposed model. The features, \( F = \{ f_1, f_2, \ldots, f_{2+18h+20} \} \), can be further refined by discretizing their values as follows.

i. For a given feature \( f_i \), there is a value set for all samples as below.

\[ V(f_i) = \{ \text{value}_1, \text{value}_2, \ldots, \text{value}_{m+k} \}. \] (29)

ii. Cluster \( V(f_i) \) to create a p categories centroid set like \( C = \{ c_1, c_2, \ldots, c_p \} \).

iii. Refine the feature \( f_i \) into \( p + 1 \) features based on \( C = \{ c_1, c_2, \ldots, c_p \} \) and the new feature set becomes \( F_{f_i} = \{ f_i', f_i'', \ldots, f_i^{p+1} \} \) which includes \( p + 1 \) features. The new feature value \( V(f_i') \) is refined as below.

\[
V(f_i') = \begin{cases} 1; & \text{if } C_i - 1 < \text{value}_i \leq C_i, \text{ i }> 1 \\ 1; & \text{if } 0 \leq \text{value}_i \leq C_i, \text{ i } = 1 \\ 0; & \text{if } \text{value}_i > C_i, \text{ i } \geq 1. \end{cases}
\] (30)

iv. For all features, \( F = \{ f_1, f_2, \ldots, f_{2+18h+20} \} \), repeat the above steps till the new refined feature set \( F' = \{ f_1', f_2', \ldots, f_q' \} \) is obtained, whereby

\[ q = \sum_{i=1}^{2+18h+20} \text{Centroid}_i \] (31)

Centroid, denotes the centroid set by clustering \( V(f_i) \) and \( \text{Centroid}_i \) denotes its size.

3) Model Training: The proposed model is trained based on SVM to identify whether a vehicle is an unlicensed taxi or a normal vehicle. The basic idea is to map the low-dimensional space curve (or surface) to a high-dimensional space line or plane. After this step, the feature is linearly separable in high-dimensional space. It is trained by the cross-training model sample, as shown below.

i. Normalize the feature set \( F' = \{ f_1', f_2', \ldots, f_q' \} \) to remove the features-the value of which is zero and mark the feature sequence;

ii. Select the Radial Basis Function as the kernel function to train the model

\[
k(\|x - xc\|) = \exp \left\{ -\frac{\|x - xc\|^2}{2\sigma^2} \right\}
\] (32)

whereby \( xc \) denotes the kernel function center, and \( \sigma \) denotes the function’s width parameter;

iii. Train the optimal model parameters \( \varepsilon \) and \( \sigma \), which are the penalty factor and nuclear parameter. The two parameters enable the model to have a high precision for the model training.

V. RESULTS AND ANALYSIS

Having described in details our proposed model, we now concentrate on its evaluation. For this purpose, we use a dataset consisting of vehicle passing records generated by 1,193,281 vehicles over a period of 31 days. The number of vehicle passing records is 340,679,449, the size of which in HDFS is about 28 G. The records are made by 303 ETBDs installed along the main roads in the city. On average, each ETBD can generate 36,270 vehicle passing records every day. To train our proposed model, we use 6,868 normal taxis and 3,760 personal vehicles as positive samples and negative samples, respectively. As a computing platform setup, we use 20 physical servers which are virtualized to 80 virtual nodes. They are built in Hadoop and HBase. Table V shows other details about the computing platform setup.

![Fig. 3. ETBD classification on map.](image)

As explained earlier, vehicle passing records \( C_{\text{vpr.etbd1.month}} \), generated in one month, are used for ETBD classification. By clustering the data, we obtain three grades of ETBD for the model training. The clustering results are shown in Table VI and the corresponding locations on the map are illustrated in Fig. 3. As shown in Table VI, for the first ETBD grade \( \text{ETBD}_1 \), the number of vehicle
passing records exceeds 3.8 million and it has 8 ETBDs. As for ETBD_{heat2}, the number of vehicle passing records is between 1.5 million and 3.8 million. 66 ETBDs belong to this grade; for the last grade ETBD (ETBD_{heat3}), the number of vehicle passing records is less than 1.5 million and there are 229 relevant ETBDs. Fig. 3 shows the geographical distribution of ETBDs on the map. Different grades are highlighted using different colors. On the map, there are eight red points, 66 yellow points and 229 green points, referring to ETBD_{heat1}, ETBD_{heat2} and ETBD_{heat3}, respectively.

To compute the city traffic morning and evening rush hours for weekdays and holidays, one month vehicle passing record \( C(vpr, etbID, 1\text{ month}) \) is used. The results are shown in Figs. 4 and 5. Indeed, Fig. 4 shows the congestion rate during weekdays. “vpr” refers to vehicle passing record. For the figure, it can be deduced that the morning rush hour is from 6:00 to 8:00 and the evening rush hour is from 17:00 to 20:00. Fig. 5 shows the congestion rate during holidays. The morning rush hour is also from 6:00 to 8:00 and the evening rush hour is from 18:00 to 21:00.

### Table VII

<table>
<thead>
<tr>
<th>index</th>
<th>thre_vpr</th>
<th>thre_etbd</th>
<th>thre_day</th>
<th>Candidate number</th>
<th>Retrieved number of vut</th>
<th>vut</th>
<th>accuracy ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>70</td>
<td>30</td>
<td>15</td>
<td>4,532</td>
<td>56</td>
<td>136</td>
<td>41.18%</td>
</tr>
<tr>
<td>2</td>
<td>60</td>
<td>30</td>
<td>15</td>
<td>6,126</td>
<td>103</td>
<td>136</td>
<td>75.74%</td>
</tr>
<tr>
<td>3</td>
<td>50</td>
<td>25</td>
<td>15</td>
<td>8,256</td>
<td>117</td>
<td>136</td>
<td>86.03%</td>
</tr>
<tr>
<td>4</td>
<td>50</td>
<td>20</td>
<td>10</td>
<td>11,231</td>
<td>135</td>
<td>136</td>
<td>99.26%</td>
</tr>
<tr>
<td>5</td>
<td>50</td>
<td>15</td>
<td>10</td>
<td>32,416</td>
<td>136</td>
<td>136</td>
<td>100%</td>
</tr>
<tr>
<td>6</td>
<td>30</td>
<td>20</td>
<td>10</td>
<td>56,178</td>
<td>136</td>
<td>136</td>
<td>100%</td>
</tr>
</tbody>
</table>

### Table VIII

<table>
<thead>
<tr>
<th>Index</th>
<th>Weekday Time Periods</th>
<th>Index</th>
<th>Holiday Time Periods</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p_1 )</td>
<td>00:00 ~ 6:00</td>
<td>( p_{10} )</td>
<td>00:00 ~ 6:00</td>
</tr>
<tr>
<td>( p_2 )</td>
<td>06:00 ~ 8:00</td>
<td>( p_{11} )</td>
<td>06:00 ~ 8:00</td>
</tr>
<tr>
<td>( p_3 )</td>
<td>08:00 ~ 9:00</td>
<td>( p_{12} )</td>
<td>08:00 ~ 9:00</td>
</tr>
<tr>
<td>( p_4 )</td>
<td>09:00 ~ 12:00</td>
<td>( p_{13} )</td>
<td>09:00 ~ 12:00</td>
</tr>
<tr>
<td>( p_5 )</td>
<td>12:00 ~ 13:00</td>
<td>( p_{14} )</td>
<td>12:00 ~ 13:00</td>
</tr>
<tr>
<td>( p_6 )</td>
<td>13:00 ~ 16:00</td>
<td>( p_{15} )</td>
<td>13:00 ~ 17:00</td>
</tr>
<tr>
<td>( p_7 )</td>
<td>16:00 ~ 17:00</td>
<td>( p_{16} )</td>
<td>17:00 ~ 18:00</td>
</tr>
<tr>
<td>( p_8 )</td>
<td>17:00 ~ 20:00</td>
<td>( p_{17} )</td>
<td>18:00 ~ 21:00</td>
</tr>
<tr>
<td>( p_9 )</td>
<td>20:00 ~ 24:00</td>
<td>( p_{18} )</td>
<td>21:00 ~ 24:00</td>
</tr>
</tbody>
</table>

To evaluate the proposed UTIM model, we first evaluate its CSM component. As UTIM is based on machine learning, the more real unlicensed taxis it identifies, the better CSM becomes. For this purpose, we use a preliminary list of unlicensed taxis already verified by the local police department. CSM can tell how many real verified unlicensed taxis are identified in the candidate list using different model parameters. CSM is evaluated based on the accuracy ratio of the number of identified unlicensed taxis to the total number of unlicensed taxis.
taxis as verified by the local police department. Intuitively, thre_vpr, thre_etbd and thre_day are three key parameters that could impact the model quality.

Table VII shows the CSM evaluation results. “vut” denotes the total number of unlicensed taxis verified by the local police department. “Candidate number” indicates the total number of vehicles identified as unlicensed taxis by CSM. “Retrieved number of vut” indicates how many vehicles in the candidate list are indeed unlicensed taxis as verified by the police department. The accuracy ratio is the ratio of the number of vut identified by CSM to vut. Table VII shows that CSM functions best when the parameters thre_vpr, thre_etbd and thre_day are set to 50, 20 and 10, respectively, as it generates with high accuracy a list of only 11231 candidates that can be easily processed by CRM.

To evaluate the performance of CRM, 6,868 positive samples and 3,760 negative samples are used (i.e., m = 6,868 in $S_{ut}$ and k = 3,760 in $S_{nv}$). Furthermore, according to the weekday/holiday morning/evening rush hours determined from Figs. 4 and 5, the weekday/holiday time periods can be divided as shown in Table VIII. Additionally, as ETBDs are classified into three grades, the envisioned samples have 128 features—$F = \{f_1, f_2, \ldots, f_{128}\}$. Consequently, in the feature refinement process, there should be 10628 sample values for each feature $f_i, V_{f_i} = \{\text{value}_1, \text{value}_2, \ldots, \text{value}_{10628}\}$.

Fig. 6 shows the feature refinement results. Different color sections represent different feature groups. After feature refinement, the new feature set becomes—$F' = \{f'_1, f'_2, \ldots, f'_{1280}\}$ consisting of 1,280 features. The value of each feature and for each sample can be collected. Then, the final feature values form a matrix which has $1,280 \times 10,628$ elements that are used to train CRM. Fig. 7 shows the results of the CRM model training. The model accuracy for all samples is 98.6029%. The parameter $\varepsilon$ is 8 and $\sigma$ is 0.0078125. The candidate list $V'$ is used as input for CRM to identify the fine-grained suspected unlicensed taxis. Table IX shows the CRM results, indicating the number of unlicensed taxis (out of CSM’s candidate list) that are identified by CRM.

To validate the CRM results, trajectories of all unlicensed taxis identified by CRM were carefully checked. Figs. 8 and 9 show the example of a vehicle xxx0C38 among those identified by CRM. The vehicle kept running for long times in the city. It carried different passengers in short time as shown in the two figures. For instance, Fig. 8 shows the vehicle xxx0C38 carrying a female passenger wearing a red dress at 16:28:41 on Jan 22nd in 2014. Fig. 9 shows the vehicle carrying another female passenger wearing a grey dress 20 minutes later, at 16:48:13. During the same day, such situation happened several times, each time with a new passenger on board, confirming that this vehicle is indeed an unlicensed taxi. All 4,348 vehicles identified by CRM as unlicensed taxis were carefully investigated in the same manner along with the professional police officer and finally 3,768 unlicensed taxis were indeed confirmed, yielding an overall accuracy of 86.667%.

<table>
<thead>
<tr>
<th>(\varepsilon)</th>
<th>(\sigma)</th>
<th>Candidate number</th>
<th>Retrieved unlicensed taxi number</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>0.0078125</td>
<td>11,231</td>
<td>4,348</td>
</tr>
</tbody>
</table>
VI. CONCLUSION
As an important contribution to D^2ITS, this paper described
an efficient model for the identification of unlicensed taxis. The
efficiency of the model in accurately identifying unlicensed
taxis is demonstrated through real-life implementation of
the model and its application to real-life traffic in a medium size
Chinese city. Indeed, using the proposed model, unlicensed
taxi were identified from over 800,000 vehicles at an accuracy
equal to 86.667%, a value high enough to enable relevant
legal institutions (e.g., police department and public transporta-
tion department) to efficiently mitigate the phenomenon of
unlicensed taxis in China and other countries suffering from
the same issue. Whilst the obtained results are encouraging,
the learning phase of the model can be further enhanced.
Furthermore, admittedly, the characteristics used to distinguish
unlicensed taxis from normal ones may not be sufficient. Other
characteristics are yet to be defined and considered. This defines
one of the future research directions of the authors in this
particular area of research.

REFERENCES
[1] Y. Wang, W. Heng, and Q. Shi, “Current status for ITS national architec-
[2] K. Yin, “How data mining is applied in intelligent transportation field,”
[5] N. Shi, “Intelligent Detection of License Plate-Related Illegal vehi-
cules Using Bayonet Images,” Ph.D. dissertation, Nanjing Normal Univ.,
Nanjing, China, 2013.
data analytics: A technology tutorial,” IEEE Access, vol. 2, pp. 652–687,
Jul. 2014.
Faculty/teaching/726/spring11/svmguide.pdf
Media, 2012.
[12] Y. Malinovskiy, J. Zheng, and Y. Wang, “Model-free video detection and
[13] Y. Malinovskiy, Y. Wu, and Y. Wang, “Video-based vehicle detection and
platform for data sharing, visualization, modeling, and analysis,” Transp.
to develop a web-based truck performance measures program,” Transp.
[16] X. Ma, Y. Wang, F. Chen, and J. Liu, “Transit smart card data mining for
passenger origin information extraction,” J. Zhejiang Univ. Sci. C, vol. 13,
[17] Y. Malinovskiy, N. Saumier, and Y. Wang, “Pedestrian travel analysis
Wei Yuan received the B.S. degree from Wuhan University, Wuhan, China, in 2011. He is currently working toward the Ph.D. degree in the Parallel Computing Laboratory, Institute of Software, Chinese Academy of Sciences, Beijing, China. He has authored/coauthored four academic papers. His research interests include information retrieval, sponsored search, data mining, parallel computing, smart city, big data, and internet system.

Pan Deng received the Ph.D. degree in computer software and theory from Beijing University of Aeronautics and Astronautics, Beijing, China, in 2011. She is currently an Associate Professor with the Parallel Computing Laboratory, Institute of Software, Chinese Academy of Sciences, Beijing, where she is also the Executive Deputy Director. She is also currently the Executive Vice President of Guiyang Academy of Information Technology, Guiyang, China, and the Deputy Director of Guiyang Technology Bureau, Guiyang. She has directed over 19 science projects as a project leader and has authored/coauthored over 25 scientific papers, four Chinese patents, and 16 software copyrights. Her research interests include smart city, big data, cloud computing, parallel computing, and Internet of things.

Tarik Taleb (SM’10) received the B.E. degree (with distinction) in information engineering and the M.Sc. and Ph.D. degrees in information sciences from Tohoku University, Sendai, Japan, in 2001, 2003, and 2005, respectively. He is currently a Professor with the School of Electrical Engineering, Aalto University, Espoo, Finland. Prior to his current academic position, he was a Senior Researcher and a 3GPP Standards Expert at NEC Europe Ltd. He was then leading the NEC Europe Laboratories Team working on R&D projects on carrier cloud platforms. Before joining NEC and until March 2009, he was an Assistant Professor with the Graduate School of Information Sciences, Tohoku University, in a laboratory fully funded by KDDI. He has been also directly engaged in the development and standardization of the evolved packet system as a member of 3GPP’s System Architecture working group. His research interests lie in the field of mobile core, mobile cloud networking, network function virtualization, software-defined networking, mobile multimedia streaming, and social media networking. Prof. Taleb is an IEEE ComSoc Distinguished Lecturer. He is serving as a Chair of the Wireless Communications Technical Committee. He is on the editorial board of the IEEE Transactions on Wireless Communications, the IEEE Wireless Communications Magazine, the IEEE Transactions on Vehicular Technology, IEEE Communications Surveys & Tutorials, and a number of Wiley journals. He has been a recipient of numerous awards, including the prestigious IEEE ComSoc Asia-Pacific Best Young Researcher award and the TELECOM System Technology Award from the Telecommunications Advancement Foundation. Some of his research studies have been also awarded best paper awards at prestigious conferences.

Jiafu Wan (M’11) received the Ph.D. degree in mechatronic engineering from South China University of Technology (SCUT), Guangzhou, China, in June 2008. He is currently an Associate Professor at SCUT, where he held a postdoctorate position in Computer Science and Engineering from October 2008 to June 2012. He is a project leader of several projects (e.g., NSFC). He has authored/coauthored one book and over 80 scientific papers (with over 30 indexed by ISI SCIE and over 40 indexed by EI Compendex) and has been cited over 1200 times. His research interests include wireless sensor networks, cyberphysical systems, Internet of things, mobile cloud computing, and embedded systems. Dr. Wan is a Senior Member of CCF and a member of ACM. He was a Workshop Chair of M2MC2012, M2MC2013, and MCC2013. He is a Managing Editor of IJAACS (EI) and JJART (EI).

Chaofan Bi received the master’s degree from the University of Washington, Seattle, WA, USA. He is currently an Assistant Research Associate with the Parallel Computing Laboratory, Institute of Software, Chinese Academy of Sciences, Beijing, China. He has authored/coauthored seven academic papers. His research interests focus on traffic safety, transportation big data, and transit performance measurement.