Journal of Ambient Intelligence and Humanized Computing
https://doi.org/10.1007/s12652-021-03346-y

ORIGINAL RESEARCH

Trip purpose inference for tourists by machine learning approaches based on mobile signaling data

Haodong Sun[1] · Yanyan Chen[1] · Yang Wang[1] · Xiaoming Liu[1]

Received: 6 October 2020 / Accepted: 7 June 2021

Abstract
It has been gradually recognized that mobile phones can be used as a practical and promising way to identify individual travel trajectories. Researchers have developed various approaches to detecting human mobility and trip characteristics including trip origin–destination, travel modes, trip purposes based on mobile phone data. Among these researches, trip purpose detection has drawn less attention from researchers. This paper presents our work to investigate a set of machine learning approaches to identifying the trip purposes for tourists based on mobile signaling data combined with sampling surveys and point of interest (POI) data. Five machine learning algorithms, including support vector machine, decision tree, random forest, artificial neural network, and deep stacked auto-encoded (DSAE), have been employed to infer trip purposes under multiple scenarios. Four scenarios have been designed by considering the POI information around trip end [a 500 m buffer or Thiessen polygon (the coverage of the base station theoretically)] and training dataset selection (equal probabilities selection or equal proportion selection). The accuracy of trip purpose classification with machine learning algorithms has compared under different scenarios. The highest accuracy of 93.47% for the test dataset is achieved based on DSAE model under the scenario of a trip end 500 m buffer and equal probabilities selection. The experimental results indicate that the methodology developed with machine learning algorithms based on mobile signaling data combined with sample travel survey is expected as an alternative way to traditional travel surveys for trip purposes.

Keywords Trip purpose · Machine learning · Mobile signaling data · Point of interest data

1 Introduction

Understanding, estimation, and prediction of trip purposes on urban transportation is essential for government agencies, decision makers, planning and design experts to implement planning, designing, management, and distribution of urban resources. Traditional methods for the research of trip purposes largely rely on travel surveys (household or personal) (Ni et al. 2018). Travel surveys are used to collect individuals information (age, gender, and status), their household (size, structure, and relationships), and their trips characteristics (time, location, mode, and purposes) during the survey period (Xiao et al. 2016). However, travel surveys are not without limitations: (a) the usage of small sample; (b) the high cost of travel surveys; (c) the limitation in spatial and temporal scales; (d) the low update frequency; (e) the heavy burden on respondents (Calabrese et al. 2013).

In the past 20 years or so, along with the development of information and communication technologies, the modern traffic and living conditions have greatly improved (Cui and Zhao 2021). Mobile phone signaling data are particularly promising to identify human mobility and apply for transportation management since the high penetration rates for mobile phones, and they are nearly always at hand (Ahas et al. 2015). Indeed, mobile phone data provide a tremendous opportunity to revolutionize the transportation field that goes beyond behavior analysis, traffic planning, operation management and safety analysis (Wang and Chen 2018). In the region of activity-based models and travel demand modelling are being the focal point of consideration for past many years (Shanmugam and Ramasamy 2021). The researches cover a wide range of interesting topics including, for example, estimation of individual human mobility patterns (Csáji et al. 2013; González et al. 2009; Song et al. 2010) and

✉ Haodong Sun
cdyan@bjut.edu.cn

1 Beijing University of Technology, Beijing, China

Published online: 28 June 2021

origin–destination (OD) matrix (Calabrese et al. 2011; Iqbal et al. 2014), and inferring trip purposes (Alexander et al. 2015; Jiang et al. 2017; Widhalm et al. 2015) and travel modes (Qu et al. 2015; Wang et al. 2010).

In contrast to the research on human mobility and travel modes, few works have been devoted to the area of the trip purposes identification (Gong et al. 2014; Montini et al. 2014). The primary problem with identifying trip purpose or activities at destination has been the need for multifactorial approaches, which require additional data sources, such as external land use or POI data, combined with the basic trip characteristics computed directly from Global Positioning System (GPS) and sensing data (Oliveira et al. 2014). Besides, with reliance on external land use or POI data, existing research on trip purpose has concentrated on developing probabilistic methods or machine learning algorithms that identify trip purposes (Ermagun et al. 2017). However, there are three aspects of limitations on trip purpose identification in existing literature, namely, data validation, classification methods related problem, and method comparison under multiple scenarios (Xiao et al. 2016).

To the best knowledge of the authors, few existing studies have focus on the trip purposes or activities for tourists. Nowadays, tourists are becoming the most active group in the city during holidays, and it is important for city manager and transportation planner to understand the purpose of their activities. In addition, compared with local residents, tourists might visit many different places and be more random in spatiotemporal. Meanwhile, there are complex and strong relationships between tourist activities and land use. Consequently, it's challenging and meaningful for researchers to infer the trip purposes of tourist. In this paper, we adopted a set of machine learning algorithms to identify the activity purpose of tourists, using massive mobile signaling data collected from millions of mobile phone users in Xiamen, China, during June 1–30, 2015, combined with small sampling surveys and POI data.

The contributions of this work are as follows:

- Firstly, this work proposes a systematical approach to identifying the trip purposes or activities of tourists based on mobile signaling data, combined with a small sampling surveys and POI data.
- Secondly, we have employed a set of machine learning algorithms to identify trip purposes and discover the potentially complex relationships between features and trip purposes using these multi-source data.
- Thirdly, we have designed multiple scenarios that considering the POI information around trip end (a 500 m buffer or Thiessen polygon) and training dataset selection (equal probabilities selection or equal proportion selection) to infer trip purposes, and the prediction accuracy under different scenarios is compared.

The rest of the paper is organized as follows: Sect. 2 summarizes methodologies and findings from the previous literature. Section 3 briefly descripts the multi-source data and feature description. In Sect. 4, we introduce multiple scenarios and the methods, such as support vector machine (SVM), decision tree (DT), random forest (RF), artificial neural network (ANN), and deep stacked auto-encoded (DSAE) used in our study. In Sect. 5, we present the case study in Xiamen, a popular tour city, and a discussion on results with different methods and scenarios. Section 6 concludes the whole paper and discusses future research.

2 Literature reviews

Trip purposes or activity type identification has not drawn significant attention in the past. Gong et al. (2014) illustrated the information of categorized methodologies utilized in the existing researches and grouped those methodologies into four broad categories:

- Rules-based methods that match the selected information from location [e.g. GPS, Geographic Information System (GIS)] and the respondent's personal information with a series of predefined heuristic rules to infer the trip purpose.
- Probabilistic methods that calculate the probability of each trip purpose based on the different values of the locational information and the respondent's personal information.
- Machine learning methods that facility complicated, computationally intensive classification, and pattern recognition models associated with the locational information, respondent's personal information, transportation information (e.g. start time, activity duration) and trip purpose.
- Deep learning methods is a multi-layer neural network, which uses the analysis and calculation methods of multi-layer networks to obtain results (e.g. SAE/DBN/RNN).

The most popular method for trip purpose identification adopted in the early existing researches is the rules-based method. Zou et al. (2016) focused on detecting the home location and trip purposes for subway passengers (cardholders), based on the internal temporal–spatial rule approach within multi-day smart card transaction data. Wolf et al. (2007a) processed GPS data collected from GPS logger in personal vehicles with a geographic information system to derive most of the traditional travel diary elements. The proposed approach established the relationships between land use and trip purposes and assigned the appropriate trip purpose for each trip based on its destination land use with 79% accuracy. Wolf et al. (2007b) presented automated methodologies for inferring trip purposes while the trip destinations are identified. The input variables of those methodologies are POI, activity duration, time of day, day of week, socio demographics, home addresses and working hours. Stopher et al. (2008) provided a software to improve the analyzing results of trip identification and detecting travel mode and trip purpose collected from GPS devices. They used a set of heuristic rules deduce the probable trip purposes based on the information of categorized land uses and the additional information on personal occupation. Bohte and Maat (2009) presented an innovative method that combines GPS logs, GIS technology and an interactive web-based validation application to derive and validate trip purposes and travel modes using reliable multi-day data collection. Wu et al. (2011) developed and evaluated two automated models to classify major time activity patterns (i.e., indoor, outdoor static, outdoor walking, and in-vehicle travel) based on GPS time activity data collected under 47 participants with 131 person days from the Harbor Communities Time Location Study in 2008 and 21 person days' data in 2010 from 3 participants. Bao et al. (2017) conducted the Latent Dirichlet Allocation (LDA) analysis model to discover the hidden bike-sharing travel patterns and trip purposes using the station types and smart card data. Lu et al. (2018) applied multinomial logit models as well as mixed logit models with considering the station category, trip time, trip sequence, and alighting station frequency during five weekdays to infer passengers' trip purpose. Chen et al. (2018b) simply adopted the common clustering algorithm (i.e., K-means) to aggregate trips with similar latent representation, then conduct trip purpose interpretation based on the clustering results, followed by understanding the time-evolving tendency of trip purpose patterns (i.e., profiling) in the city-wide level. Zhao et al. (2019) presented an unsupervised two-phase framework for inferring multiple trip purposes (i.e. loading, unloading, in-yard, and other stops) based on the passive global positioning system (GPS) data during the hazmat-transportation process.

In last decade, the probabilistic methods and machine learning method gradually emerged and with high accuracy. Wang et al. (2018) developed a general probabilistic framework for spotting trip purposes from massive taxi GPS trajectories. Chen et al. (2018a) proposed a probabilistic two-phase framework named *TripImputor*, for making the real-time taxi trip purpose imputation and recommending services to passengers at their drop off points. Cui et al. (2018) employed a Bayesian neural network (BNN) to model the trip dependence on each individual's daily trip chain and infer the trip purpose. Alsger et al. (2018) investigated the potential of the smart card data to infer passengers' trip purpose, thereby reducing the use of the expensive and time-consuming Household Travel Surveys. Chen et al. (2010) used a multi-modal transportation network, a set of rules to achieve both complexity and flexibility for trip end clustering and trip purpose prediction. They applied the hierarchical clustering method to cluster trip ends into activity locations and then adopted a probabilistic model to evaluate how various factors determine the purpose of a trip probabilistically. Li et al. (2020) proposed a rigorous method combined the Gaussian mixture model and the hidden Markov model to interpret the purpose of each leg of the trip-chain. Oliveira et al. (2014) presented two methods, choice modeling and decision tree analysis to develop models for identifying trip purpose based on the 2011Atlanta Regional Commission household travel survey data. The developed models produced encouraging results with overall accuracy greater than 70% across all purposes and around 90% for mandatory activities (i.e., work and school). Liao et al. (2016) proposed a novel approach which takes the high-level context into consideration to identify the significant places and adopted hierarchically structured conditional random fields to infer trip purposes for each person. The model could be applied to different persons successfully after trained, achieving more than 85% in inferring low-level activities and above 90% accuracy in determining and identifying significant places. Deng and Ji (2010) presented a decision tree method which engages a number of attributes (i.e., time stamp and land-use type of trip ends, a set of spatiotemporal characteristics of travel, and additional personal information) to deriving trip purpose from GPS track data coupled with other relevant data sources. The results of the method seemed rather promising, with an overall classification accuracy of 87.6%. Montini et al. (2014) applied random forest to improve trip purpose identification based on GPS tracks and accelerometer data collected by 156 participants for a week travel survey from Switzerland in 2012. The results showed that the share of correct predictions was between 80 and 85%. Lu and Zhang (2015) developed machine learning methods (including decision tree and meta-learning) to infer trip purposes for long-distance passenger travel and validated by the 1995 American Travel Survey associated with land use data. The accuracy of the trip purpose imputation methods decreases from 95% with two purposes (business and non-business) to 77% with four purposes (business, personal business, social visit, and leisure). Xiao et al. (2016) proposed an artificial neural networks associated with particle swarm optimization to identify trip purposes from the GPS data combined with land use information and POI. The highest accuracies of 97.22% for the training dataset and 96.53% for the test dataset are achieved under the scenario of polygon-based information and equal proportion selection.

Recently, deep learning methods were proposed by various research studies to minimize the issues inherent in conventional feature learning methods (Alo et al. 2020). Nweke et al. (2018) provided in-depth summaries of deep learning methods for mobile and wearable sensor-based human activity recognition. They not only categorized the studies into generative, discriminative and hybrid methods but also highlighted their important advantages. Alawneh et al. (2021) investigated the benefits of time series data augmentation in improving the accuracy of several deep learning models on human activity data gathered from mobile phone accelerometers. Hassan et al. (2018) presented a smartphone inertial sensors approach based on a Deep Belief Network (DBN) for successful human activity recognition. The result showed that the proposed approach was outperformed than traditional expression recognition approaches such as typical multiclass Support Vector Machine (SVM) and Artificial Neural Network (ANN). Suto and Oniga (2018) investigate the performance of more ANN structures with different hyper-parameters and inputs on two public databases to recognition human activity. Sansano et al. (2020) conducted an extensive analysis among the most suited deep learning architectures, such as convolutional neural networks (CNN), long short-term memory networks (LSTM), bidirectional LSTM (bi-LSTM), gated recurrent unit networks (GRU), and deep belief networks (DBN), for activity recognition. Results showed that CNNs are efficient at capturing local temporal dependencies of activity signals, as well as at identifying correlations among sensors and performance in activity classification better than recurrent models. Almaslukh et al. (2018) proposed a deep convolution neural network (CNN) model that provided an effective and efficient smartphone-based human activity recognition system. The results showed that the proposed model established the state-of-the-art performance using these datasets. Peng et al. (2019) proposed a complex activity recognition using acceleration, vital sign, and location data to recognize complex activities from wearable sensors. The results of experiments showed that the proposed approach could effectively utilize location data to improve complex activity recognition performance. Buijs et al. (2021) explored a relatively new method that can contribute to behavioral analysis of transport movements, using a novel data set collected in Amsterdam with an app on smartphones that automatically recognizes activity signals. Thakur and Biswas (2020) presented a comprehensive survey of smartphone sensor based human activity monitoring and recognition using various ML and DL techniques to apprehend varied completely different activities with high precision. Alo et al. (2020) proposed a deep stacked autoencoder algorithm, and orientation invariant features, for complex human activity identification. The results showed that the proposed model can accurately recognize complex activity details using only smartphone accelerometer data and achieved 97.13% identification accuracy compared to the conventional machine learning methods and the deep belief network algorithm.

In this paper, we attempt to investigate the use of mobile signaling data, associated with POI data to infer trip purposes or activities of tourists. We have employed five trip purpose identification approaches (the SVM model, the DT model, the RF model, the ANN model and the DSAE model) to evaluate the utility of mobile signaling data in trip purpose identification. In addition, because of the correlation between the trip purpose and land use, we designed different scenarios for testing. For example, on the one hand, we calculate the POI information (e.g., restaurants, grocery stores, hotel, park, greenbelt, government buildings, etc.) in the Thiessen polygon of the trip end; on the other hand, we calculate the POI information within a buffer of 500 m from the trip end. In the result analysis, accuracy under different scenarios is compared. Table 1 summarizes various models were recently implemented for trip purposes inference, and their strengths, and weaknesses.

3 Data

In this section, we describe the use of data (e.g. mobile phone signaling data, POI data, and survey data), feature description (e.g. trip related characteristics, the POI information around the destination of trip).

3.1 Data collection

3.1.1 Mobile signaling data

In this study, the mobile signaling data was collected from June 1 to June 30, 2015 in Xiamen. There are 5555 telecommunication base stations in Xiamen, mainly distributed in Xiamen Island (Fig. 1). The data mainly contains the seven fields, and explanations for each field are shown in Table 2. There are about 10 million mobile phone users in Xiamen. In our previous work, we have extracted users' behavior characteristics and then employed unsupervised clustering techniques to identify tourists. In this work, we sift through all the trajectory data of the tourists and transfer these trajectory data into trip data. Spatial and temporal thresholds were adopted to generate the locations where tourists stop for specific activities (Zhong et al. 2018).

3.1.2 POI data

The POI data, usually used to infer trip purpose for each trip, was collected and aggregated from the Xiamen electronic map in the year of 2015. The figure below shows the complicated POI data in Xiamen Island (Fig. 2). In total, there are roughly 15,000 points of interest in Xiamen, Fujian province. The descriptions of each type POI are shown in Table 3.

Table 1 Summary of various methods for trip purposes inference

Methods	Descriptions	Strengths	Weaknesses
Rules-based methods	Matching the selected information from location and the respondent's personal information with a series of predefined heuristic rules to infer the trip purpose (e.g. heuristic rules/multi-data matching rules)	Simple and requires fewer data sources	Mainly based on subjective experience, the classification accuracy is not high
Probabilistic methods	Calculated the probability of each trip purpose based on the different values of the locational information and the respondent's personal information (e.g. multinomial logit model)	Simple, and the classification results based on statistical probability distribution are relatively high	It is difficult for the model to deal with complex classification task, and the classification accuracy will decrease
Machine learning methods	Facility complicated, computationally intensive classification models associated with the locational information, respondent's personal information, transportation information and trip purpose (e.g. DT/RF/SVM/ANN)	Learns features automatically, high recognition accuracy, and widely used	Requires large amount of training data to obtain discriminant features. In addition, it requires a high number of parameter optimization
Deep learning methods	Deep learning is a multi-layer neural network, which uses the analysis and calculation methods of multi-layer networks to obtain results (e.g. SAE/DBN/RNN)	Reduces high-dimensional data to low dimensional feature vectors. This helps to reduce computational complexity	Lack of scalability to high-dimensional data. It is difficult to train and optimize, especially for one layer auto-encoder

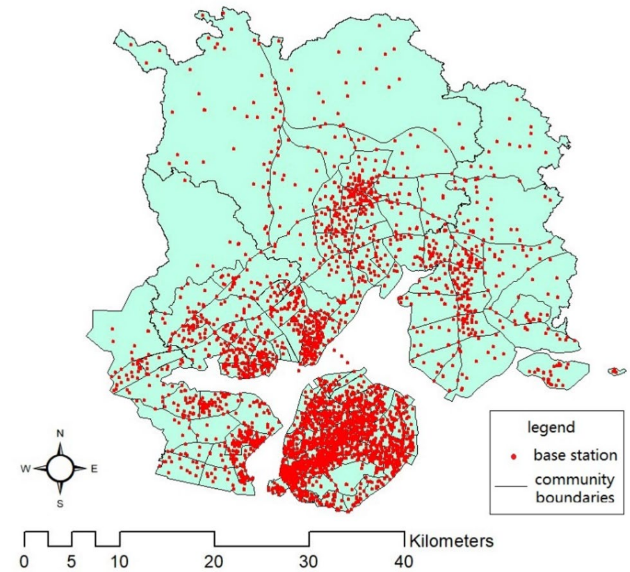


Fig. 1 Telecommunication base station distribution in Xiamen

3.1.3 Survey data

It is important to understand the ground truth of the tourists' trip purposes or activities which are used to train the models. In this study, we organized surveys for tourists in Xiamen from June 1 to June 30, 2015. We surveyed whether the volunteers would authorize us to use their mobile signaling data in Xiamen and called them back for a telephone interview. In the telephone interview, we first told the volunteers about the derived trip information (e.g. trip origin, trip destination, time of origin and destination) and asked them whether the information was correct and their trip purpose.

3.2 Feature description

The sample size in the survey was 350 respondents, 24 of whom failed because their mobile signaling data records were too few to extract the trip characteristics and matched the surveyed trip information. In total, mobile data and trip characteristics of 1059 person days of 326 respondents were collected from June 1 to June 30, 2015, in Xiamen. According to existing literature, we classified four trip purposes [i.e., tour, accommodation, catering and recreation and return back (back to home)] which were the most of all tourists' activities. A total of 3079 trips with these four trip purposes were acknowledged in the survey. The points of interest analysis in ArcGIS 10.2 allowed us to obtain the statistical information on POI within the buffer of trip end. The number of trips with each trip purpose is shown in Fig. 3. The highest percentage was taken by tour trips (1288), followed by catering and recreation trips (747), accommodation trips (689) and return back trips (355).

Table 2 Descriptions of mobile signaling data

Field names	Field explain
DATE	The date when user triggers the telecommunication base station
MSISDN	The abbreviation of Mobile Subscriber International ISDN/PSTN number
REGION	The area where the phone number is displayed
LONGITUDE	Longitude of telecommunication base station
LATITUDE	Latitude of telecommunication base station
START_TIME	The start time of triggering the telecommunication base station
END_TIME	The end time of triggering the telecommunication base station

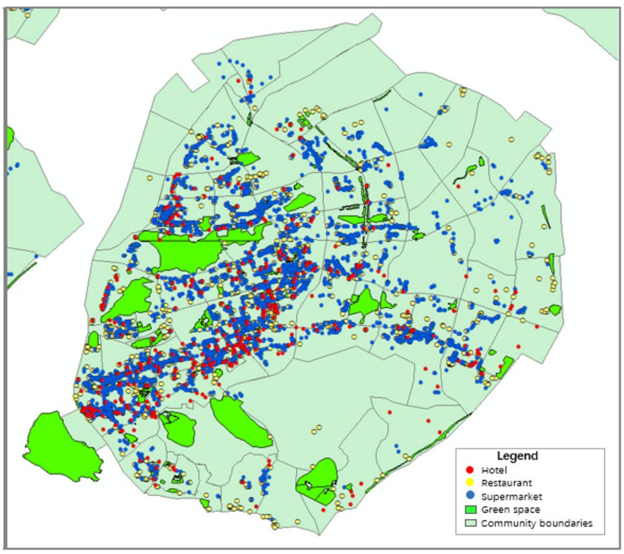


Fig. 2 Distribution of points of interest in Xiamen Island

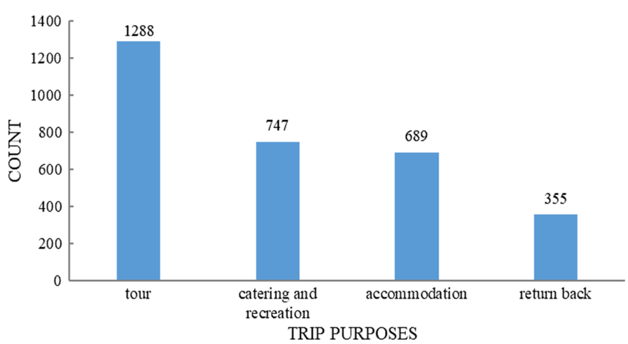


Fig. 3 The number of trips with different purposes

A proper feature set can be conducive to improving classification accuracy. Individual characteristics (i.e., age, gender, education, income) and household characteristics (i.e., the presence of children) are used to infer trip purpose in most of existing literature. Considering about the difficulty of getting the individual and household characteristics, we used the mobile signaling data to extract the trip characteristics (i.e., location of origin and destination, time of origin and destination, activity duration) and POI information of trip end (i.e., the amounts of hotel, restaurant, supermarket, green space, park).

According to the existing researches (Xiao et al. 2016), analyzing the activity duration was helpful for estimating different trip purposes (Fig. 4a). Activity duration of tour had a similar distribution to that of catering and recreation in that a single peak occurred in "less than 1 h" and then the percentage of those activities decreased monotonically. In addition, the distribution of accommodation activity duration exhibited high percentage in "longer than 10 h". On the contrary, return back activity duration presented high peak in "less than 1 h" and with declined sharply after.

Similar to activity duration, analyzing the activity start time also contributed to identifying different trip purposes (Fig. 4b). The distribution of tour activity start time exhibited double peaks. The former peak during 07–10 a.m. was obvious since tourists often start their tours at this time, whereas the latter peak during 12–14 p.m. might result from tourists toured after lunch or a noon break. Similarly, double peaks presented in the distribution of the start time in catering and recreation activity. The first one during 10–11 a.m. might result from tourists toured in 07–10 a.m. and then have lunch that early than locals. The last peak during 15–18 p.m, which had the same reason that tourist had

Table 3 Descriptions of points of interest in Xiamen

No.	Type	Amount	No.	Type	Amount
1	Hotel	523	6	School	966
2	Restaurant	2955	7	Drug store	667
3	Supermarket and shopping mall	6099	8	Medical	576
4	Mansion	400	9	Bank	2053
5	Park and green buffer	234	10	Government agency	510

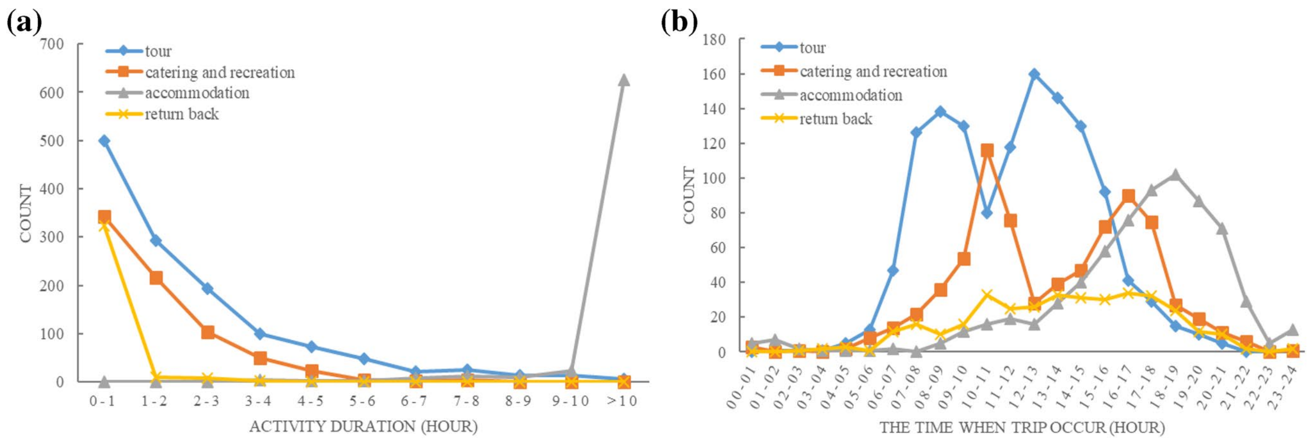


Fig. 4 Distribution of features with different purposes. **a** Distribution of activity duration with different purposes. **b** Distribution of activity start time with different purposes

dinner after toured. Unlike the former activities, the accommodation activity appeared single peak during 17–20 p.m. might infer tourist went back hotel to sleep after toured and dinner in daytime, besides, there had a little rise during 11 a.m.–12 p.m. which might be associated with a short noon break. The distribution of return back presented a relatively stable state during 10 a.m.–19 p.m. probably because tourist returned randomly in daytime.

The features used to infer trip purposes are shown in Table 4. Ten features (o_zoneid, d_zoneid, o_time, d_time, trip time, duration, No. trip, last trip and weekend) of trip characteristics were extracted from mobile signaling data associated with survey data. In addition, one trip purpose was related to trip end corresponded to multiple land use types (POI). In this paper, we attempt to infer tourist trip purposes, therefore, 5 variables (hotel, restaurant, supermarket, park and green buffer) were used to represent the land use type of the trip end. In addition, considering that the level of park and green buffer attracts different amount tourists, we set different weights according to the level of park and green buffer.

At the same time, we used the random forest classifier in ensemble learning to rank the importance of features statistically (Fig. 5). The results show that tourists' duration, departure time and arrival time are of significant importance to the activity identification of tourists, among which the importance of the duration can reach 30.89%. As we mentioned above, the analysis of duration is of great help to the identification of tourist activities (Fig. 4a).

Table 4 Features used to infer trip purposes

Feature category	Feature name	Description
Trip characteristics	O_zoneid	Origin zone id of trip
	D_zoneid	Destination zone id of trip
	O_time	Origin time of trip
	D_time	Destination time of trip
	Trip_time	Time inner trip
	Trip_distance	Distance inner trip
	Duration	Duration of the activity
	No._trip	Number of trip in a day
	Last_trip	Last trip in whole tour
	Weekend	Trip made at weekend
POI information	Hotel_num	Number of hotels around the destination
	Restaurant_num	Number of restaurants around the destination
	Supermarket_num	Number of supermarkets around the destination
	Park_num	Number of parks around the destination
	Green_buffer	Contain green buffer around the destination

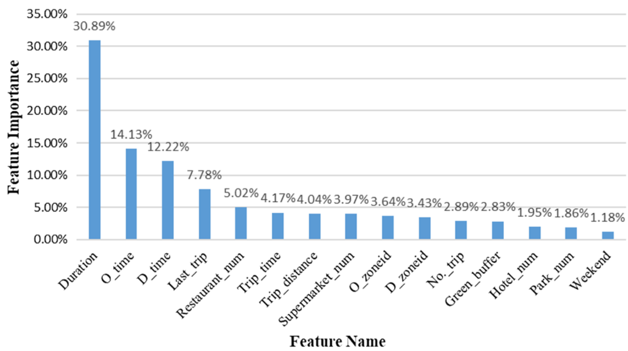


Fig. 5 The ranking of feature importance

4 Methodologies

In this section, we introduce multiple designed scenarios (e.g. POI information and training dataset) and machine learning algorithm (SVM, DT, RF, ANN, DSAE) for trip purposes identification.

4.1 Multiple scenarios

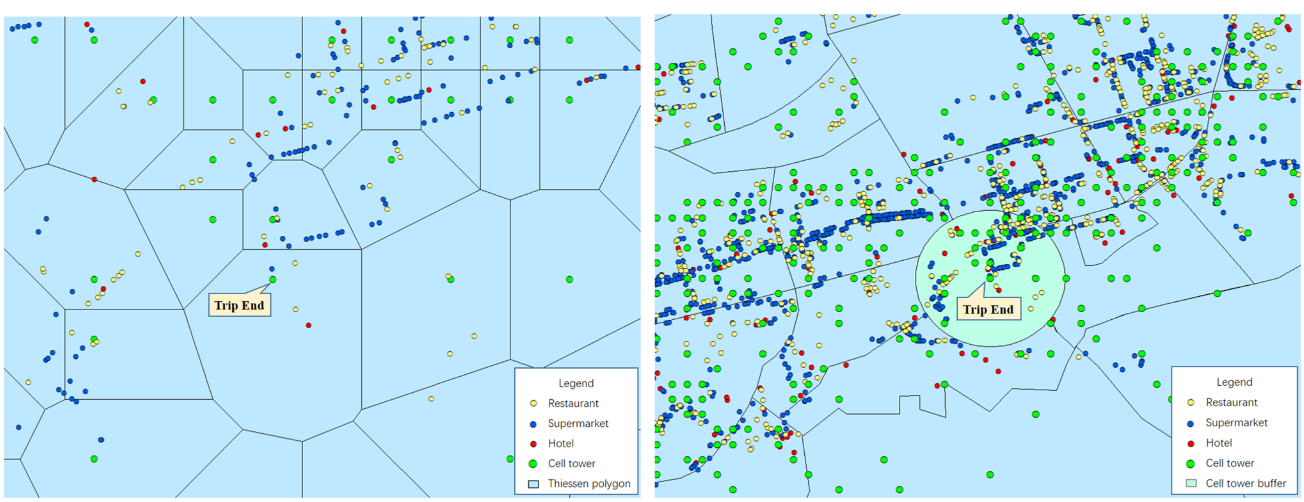
To identify the best scenario for inferring trip purposes, it's necessary to compare classification results under each scenario. In this section, we propose two considerable aspects that trip end POI information and training dataset selection to design multiple scenarios.

4.1.1 Trip end POI information

POI information was usually used to detect trip purposes. A trip end is supposed to be at home/work/education if the trip end is within the buffer of 500 m from home/work/education (Lu et al. 2013). In this paper, we designed two scenarios to calculate trip end POI information. The first one was based on Thiessen polygon that we obtain the amount of each type POI information in the trip end Thiessen polygon. Figure 6a shows an example of how to calculate the amount of each POI information. In the trip end Thiessen polygon, we might find there are two restaurants, one supermarket and one hotel. Therefore, the values of "Hotel_num" and "Supermarket_num" are 1, "Restaurant_num" is 2, and other POI type are 0. On the other hand, we calculate the amount of each type POI information within a 500 m buffer of a trip end. Figure 6b depicts the POI information in trip end 500 m buffer. There are 13 hotels, 51 restaurants, 84 supermarkets in the trip end 500 m buffer. So the values of "Hotel_num" is 13, "Restaurant_num" is 51, "Supermarket_num" is 84, and other POI type are 0.

4.1.2 Training dataset selection

A training dataset is applied to determine potentially predictive relationship between trip purposes and features, and a test dataset is used to evaluate the accuracy and applicability of the predictive relationship (Xiao et al. 2016). Since the different number of samples with each trip purpose, the training dataset can be casually selected in two common ways: equal probabilities selection and equal proportion selection. Equal probabilities selection is a sampling method in which each individual in a population has an



(a) Thiessen polygon **(b)** 500 m buffer

Fig. 6 POI information around the trip end. **a** Thiessen polygon. **b** 500 m buffer equal probability of being selected. In this paper, the sampling rate is 80%, so training dataset have 2487 samples. On the other hand, equal proportion selection ensures that the training dataset is proportional to the test dataset for each trip purpose. In this paper, 80% of samples are randomly selected as the training dataset, and the rest is test dataset. As a result, the amounts of samples in the training dataset with the four purposes are 1030, 598, 551 and 284.

In summary, we test four scenarios for inferring trip purposes:

- Scenario 1: 500 m buffer of trip end and equal probabilities training dataset selection
- Scenario 2: 500 m buffer of trip end and equal proportion training dataset selection
- Scenario 3: Thiessen polygon of trip end and equal probabilities training dataset selection
- Scenario 4: Thiessen polygon of trip end and equal proportion training dataset selection.

4.2 Machine learning algorithms

There are many machine learning algorithms, among which the algorithms used to realize classification can be divided into logic-based (decision tree and random forest), perceptron based (ANN and deep learning), statistical learning (Bayesian and SVM) and instance-based (k-NN) algorithms (Praveen Kumar et al. 2019). In this study, we considered the support vector machine algorithm, because it was the best classification algorithm before the emergence of deep learning. It could be applied to both linear (regression problem) classification and nonlinear classification. In addition, we chose decision tree and random forest based on logical rules for feature classification, because they could make feasible and effective results for large data samples in a relatively short period of time. At the same time, we also introduced the artificial neural network (ANN) algorithm, because it had a strong nonlinear fitting ability and strong feature extraction ability, and could automatically extract the reasonable rules between the data through learning. Meanwhile, we also employed the deep sparse auto-encoder (DSAE) algorithm. Because of its sparsity, the automatic feature selection could be realized, and could give a better feature description than the original data.

4.2.1 SVM model

The trip purposes identification is actually a process of data classification and support vector machine (SVM) is very powerful for dealing classification problems. Since SVM classification can obtain adaptive nonlinear boundary for optimizing the classification boundary. In this paper, the SVM classifier is trained and tested with a Radial Basis Function (RBF) kernel, and a one-versus-rest mode is employed to enable multiclass classification. The RBF kernel can achieve nonlinear mapping and has less numerical difficulties. This one-versus-rest mode only needs an optimal hyperplane between a class of samples and the remaining multi-class samples. Therefore, the number of generated hyperplanes is less and the prediction speed is faster than one-versus-one mode. In this paper, we applied 4 SVMs to deal the 4 trip purposes classification problems.

4.2.2 DT model

This study employed a decision tree-based approach, which is a simple, yet powerful, suitable, non-parametric supervised learning method for classification. DT adopted an effective and straightaway approach to understand the relationships between independent and dependent variables. DT produced results in rule expressions or graphical form, where tree induction is a recursive top-down process. We might find a number of classification DT algorithms in literature, such as ID3, CART, and C4.5. In this paper, we employ a CART algorithm in classification model.

4.2.3 RF model

Similar to DT model, the random forest (RF) consisted of numerous decision trees, each of which predicting a trip purpose grounded on the given independent variables (Ermagun et al. 2017). Each tree in the RF is built from the training set. Besides, when splitting a node during the construction of the tree, the split that is picked is the best split among a random subset of the features not all the features. As a result of this randomness, the bias of the forest usually slightly increases, however, because of averaging, its variance also decreases, usually more than compensating for the increase in bias, so the model is overall better. At last, the final result combines classifiers by averaging their probabilistic prediction, instead of letting each classifier vote for a single class.

4.2.4 ANN model

An artificial neural network (ANN) is developed to find potentially predictive relationships between the output and the inputs (Xiao et al. 2016). ANN builds a mathematical or computational model which is similar to human brain to deal the problem in the prediction and decision making methods. An ANN model generally consists of minimum three layers such as an input layer, hidden layer, and an output layer. In the input and output layers, the number of neurons is equal to the number of input and output variables, while the number of neurons in the hidden layer depends on the type of problem. In this study, we used five layers (including three hidden layers) ANN to infer trip purposes. The ANN includes 15 input neurons and 4 output neurons. The number of neurons in the hidden layers is decided by the classification performance of ANNs. In addition, we employed a feed-backward neural network with a soft-max activation function.

4.2.5 DSAE model

Deep stacked auto-encoder is an unsupervised deep learning method that uses the sparse term of the model loss function to learn complete feature representation from the original sensor data, which can give a better feature description than the original data. Meanwhile, the use of sparsity enables the model to learn feature representations that are robust, linearly separable, and invariant to changes, distortions, and displacements, as well as learning applications. These salient features of the sparse auto-encoder ensure effective extraction of low-dimensional features from high-dimensional input data. In this research, we have implemented a seven-layer sparse auto-encoder trained with a greedy layer method. Therefore, the output of the first layer is used as the input of the second layer. In each training iteration, the weight matrix, deviation vector and loss function of the model are iteratively updated. The sparsity term is added to the cost function of the automatic encoder by using the regularization term. Using greedy training helps to determine the average output activation value of the network to minimize overfitting.

4.3 Accuracy assessment measures

To evaluate the performance of the proposed tourist trip purpose identification framework based on machine learning, we calculated four performance metrics that include accuracy, recall, precision, and f-measure. For each trip purpose class, the predicted value was measured with ground truth labels. After testing, we used the confounding matrix for each prediction to calculate the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).These performance metrics were selected based on their wide application in evaluating the performance of the trip purpose identification framework.

Accuracy calculates the ratio of trip purpose classes that are correctly classified from the total number of trip purpose instances. The accuracy is calculated using the following equation:

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N}\frac{(TP+TN)_i}{(TP+FP+TN+FN)_i}$$

where N is the total number of classes in the training sample.

Precision represents the proportion of positive instances predicted to be correct. Precision is calculated by the following equation:

$$Precision = \frac{1}{N}\sum_{i=1}^{N}\frac{(TP)_i}{(TP+FP)_i}.$$

Recall represents the ratio of the positive instances predicted to be correct to the actual positive instances. Recall is calculated by the following equation:

$$Recall = \frac{1}{N}\sum_{i=1}^{N}\frac{(TP)_i}{(TP+FN)_i}.$$

F-measure represents the harmonic mean of precision and recall rates. F-measure is calculated by the following equation:

$$f\text{-}measure = 2\times\frac{Precision\times Recall}{Precision+Recall}.$$

To better show the overview of the proposed method, we adopted a flowchart to display the proposed method (Fig. 7). It shows that the flow chart mainly consists of three parts: data preprocessing, machine learning algorithm pre-training, and machine learning algorithm test and output results.

5 Case study and discussion

In this section, the prediction results of trip purposes for tourists are presented to assess the accuracy of machine learning algorithms under different scenarios.

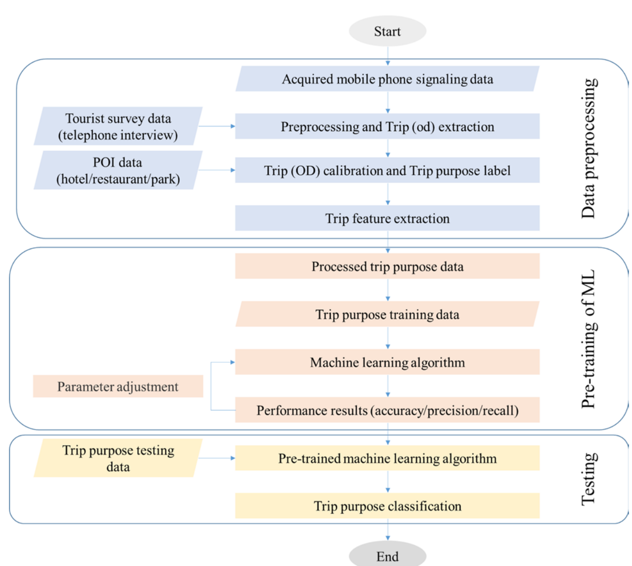


Fig. 7 Flowchart of the proposed method

5.1 The performance results of machine learning algorithms under different scenarios

In this paper, we employed five representative machine learning algorithms to infer the classification issue of trip purposes for tourists under four scenarios. These algorithms are SVM, DT, RF, ANN and DSAE, as described in Sect. 4.2. These four scenarios are also introduced in Sect. 4.1. Consequently, the performance results of those algorithms under different scenarios are presented in Fig. 8.

Among these results of prediction accuracy for machine learning algorithms under multiple scenarios in test datasets. In scenario 1, the average performance of each algorithm is generally higher than that of the other three scenarios. The performances of DSAE had achieved was the highest among these machine learning algorithms under multiple scenarios. The performance accuracy of DSAE achieves was on average about 93.47% among different metrics under scenario 1, followed by ANN is (90.54%), RF is (89.68%), SVM (is 88.15%) and DT (is 83.98%). The average prediction time of DT model is 0.01 s, RF model is 0.04 s, and SVM model is 0.29 s, ANN model is 0.47 s for each epoch, and the total testing time took 47.5 s, DSAE is 0.22 s for each epoch, and the total testing time took 21.9 s to complete. In summary, the performance results indicate that DSAE model performs better classification for trip purposes than other methods.

5.2 Classification results of trip purposes with different machine learning algorithms under scenario 1

In this study, we use the confusion matrix to evaluate the performance of the algorithm on each trip purposes class. Table 5 presents the confusion matrix analysis for DSAE model under scenario 1. In confusion matrix, each row shows the actual number of trips in the original data, each column indicates the number of trips that were predicted by the model, precision and recall are used to assess the accuracy each class in the model. For example, looking at the first row of Table 5 we can see that, of the 258 actual “Tour” trip purposes are predicted that 242 were “Tour”, 15 were “Catering and recreation”, and 1 was “Accommodation”. It shows that 93.80% (i.e., 242/258) of the actual “Tour” trips are correctly predicted by DSAE model under scenario 1. Similarly, looking at the first column of Table 5 we can see that the DSAE model predicted 271 activities to be “Tour”

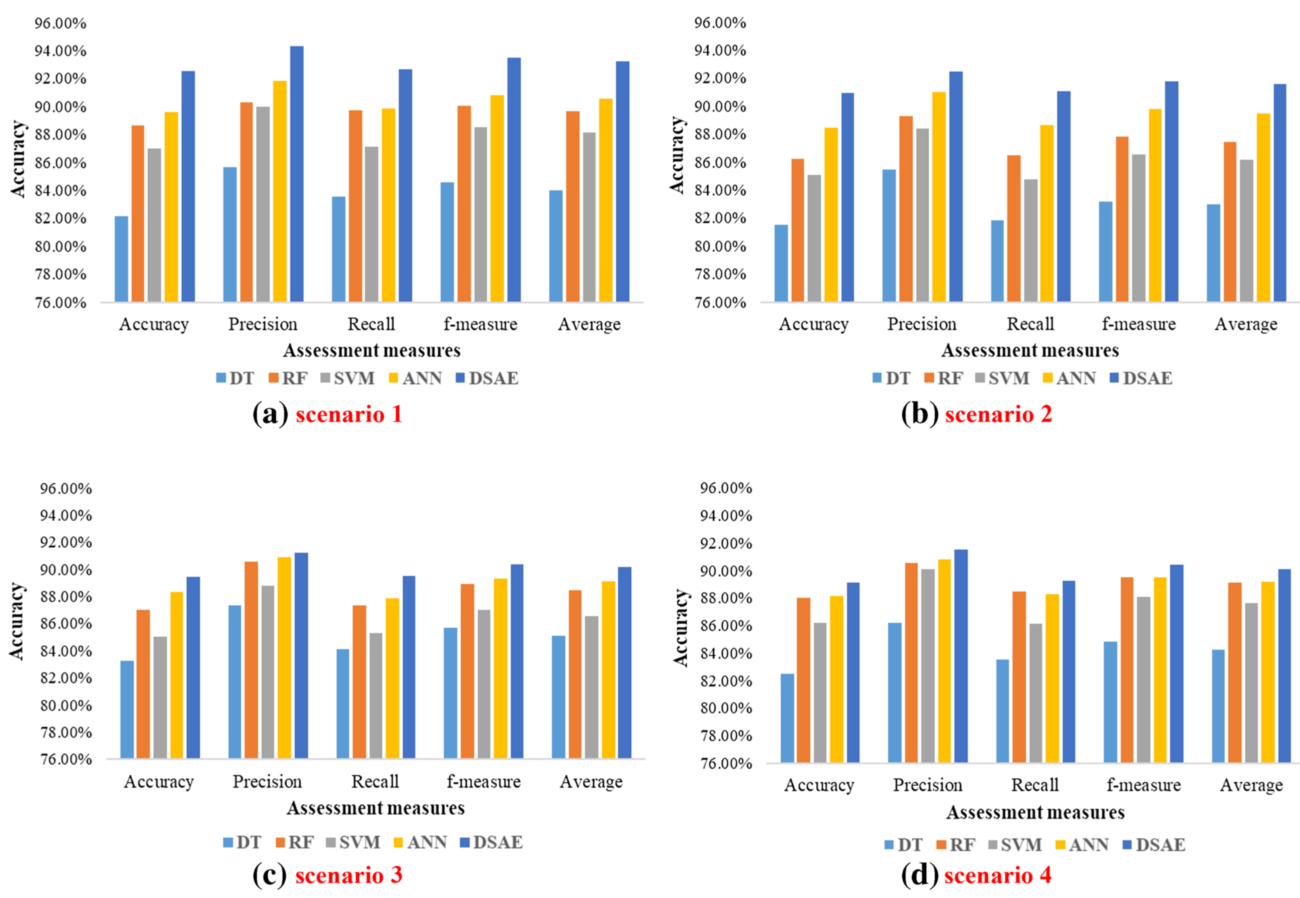


Fig. 8 Performance results for machine learning algorithms under different scenarios

Table 5 Performance results for DSAE algorithms under scenario 1

Trip purposes	Predicted				Recall (%)	Average recall (%)
	Tour	Catering and recreation	Accommodation	Return back		
Actual						
Tour	242	15	1	0	93.80	92.65
Catering and recreation	24	125	0	0	83.89	
Accommodation	1	1	136	0	98.55	
Return back	4	0	0	67	94.37	
Precision (%)	89.30	88.65	99.27	100.00	92.53	93.47
Average Precision (%)	94.31					

activities, out of which 242 were truly tour, 24 were catering and recreation, 1 was accommodation, and 4 were return back. This corresponds to 89.30% (i.e., 242/271) precision performance of the DSAE model for "Tour" trip purpose under scenario 1. In summary, The average precision is 94.31%, average recall is 92.65%, accuracy is 92.53%, and f-measure is 93.47%.

In addition, we visualized confusion matrixs with confusion matrix diagram (Fig. 9). From the confusion matrix diagram, we can find that machine learning algorithms in identifying tourists' "Tour" and "Catering and recreation" trip purposes accuracies are generally lower than in the "Accommodation" and "Return back" trip purposes, in addition, DT in the "Tour" and "Catering and recreation" trip purposes performance are poorer, DSAE model performed significantly better than other algorithms.

To sum up, we investigated the effectiveness of mobile phone signaling data combined with other travel survey data on the accuracy of the machine learning models for tourists' trip purposes recognition. The results showed that the DSAE model and the ANN model performed better than other machine learning models. The performance accuracy of DSAE model achieves was about 93.47% on average. The results showed that using deep learning model on recognition of trip purposes is promising and improves the accuracy significantly. As was mentioned before, other authors also worked on the trip purposes recognition but they measured smaller accuracies. For example, Wolf et al. (2007a) proposed approach established the relationships between land use and trip purposes with 79% accuracy. Deng and Ji (2010) presented a decision tree method with an overall classification accuracy of 87.6%. Montini et al. (2014) applied random forest to improve trip purpose identification and the correct predictions was between 80 and 85%. Xiao et al. (2016) proposed an artificial neural networks associated with particle swarm optimization to identify trip purposes and achieved 96.53% accuracy for the test dataset. Alsger et al. (2018) applied spatial and temporal attributes to infer passengers' trip purposes with 78% correct inference. Li et al. (2020) proposed a rigorous method combined the Gaussian mixture model and the hidden Markov model to interpret the trip purpose with 77% accuracy.

6 Conclusions

In this paper, we investigated the mobile signaling data combined with a small sampling survey data on the accuracy of the SVM, DT, RF, ANN, and DSAE models under different scenarios for tourists' trip purposes recognition. The results showed that the DSAE model presents best classification performance and achieves classification accuracy of 93.47% in the test data. Meanwhile, the results showed that using mobile signaling data combined with sample travel survey is promising way to infer trip purposes for tourist and deep learning algorithms improves the accuracy of results significantly. As part of our future directions, we intend to apply the DSAE models on more datasets to take advantage of the generalization of this approach. We also aim to explore and compare the effectiveness of other techniques such as the Bayesian Optimization, the Meta Learning, and the Reinforcement Learning approach. Finally, we intend to apply the models on a Cloud Computing environment and study different machine learning approaches such as supervised learning, semi-supervised learning and federated leaning.

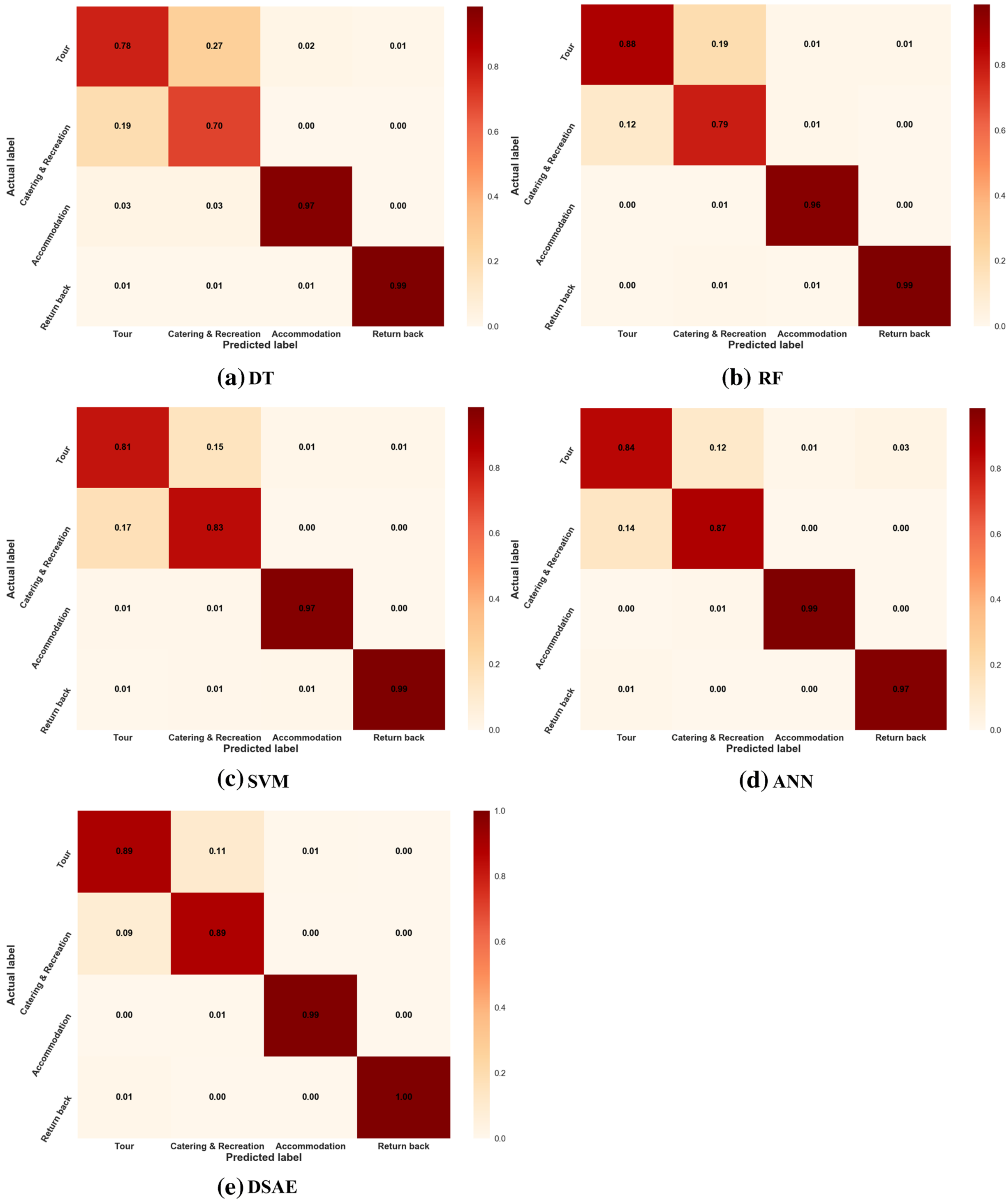


Fig. 9 Confusion matrix showing the performance results of each machine learning method under scenario 1

Acknowledgements The authors would like to acknowledge the financial support for this study provided by the National Key Research and Development Plan of China (no. 2016YFE0206800) and Science and Technology Project of Beijing (no. Z181100003918011). Meanwhile, we would like to thank Prof. Jianming Ma at Texas Department of Transportation for his great comments and assistance on our experiments.

Author contributions YC: Conception of the study and Constructive discussions. HS: Data analyses, Literature Search and Review, Manuscript Writing. YW: Manuscript editing and Constructive discussions. XL: Constructive suggestions.

Data availability statement Some or all data, models, or code generated or used during the study are proprietary or confidential in nature and may only be provided with restrictions. (1) Mobile phone signaling data of Xiamen: This data is in cooperation with Xiamen communication operator, and our permission is only allowed to deploy the algorithm on their data platform and calculate the results. Meanwhile, the data cannot be token out. Therefore, this data is provided with restrictions. (2) POI data and survey data: These data are available from the corresponding author if requested. (3) Machine learning method: This related codes are also available from the corresponding author if requested.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Ahas R, Aasa A, Yuan Y, Raubal M, Smoreda Z, Liu Y, Ziemlicki C, Tiru M, Zook M (2015) Everyday space–time geographies: using mobile phone-based sensor data to monitor urban activity in Harbin, Paris, and Tallinn. Int J Geogr Inf Sci 29(11):2017–2039

Alawneh L, Alsarhan T, Al-Zinati M, Al-Ayyoub M, Jararweh Y, Lu HT (2021) Enhancing human activity recognition using deep learning and time series augmented data. J Ambient Intell Humaniz Comput. https://doi.org/10.1007/s12652-020-02865-4

Alexander L, Jiang S, Murga M, González MC (2015) Origin–destination trips by purpose and time of day inferred from mobile phone data. Transp Res Part C Emerg Technol 58:240–250

Almaslukh B, Al Muhtadi J, Artoli AM, Tiwari S, Trivedi M, Kohle ML (2018) A robust convolutional neural network for online smartphone-based human activity recognition. J Intell Fuzzy Syst 35(2):1609–1620

Alo UR, Nweke HF, Teh YW, Murtaza G (2020) Smartphone motion sensor-based complex human activity identification using deep stacked autoencoder algorithm for enhanced smart healthcare system. Sensors (basel) 20(21):6300. https://doi.org/10.3390/s20216300

Alsger A, Tavassoli A, Mesbah M, Ferreira L, Hickman M (2018) Public transport trip purpose inference using smart card fare data. Transp Res Part C Emerg Technol 87:123–137

Bao J, Xu C, Liu P, Wang W (2017) Exploring bikesharing travel patterns and trip purposes using smart card data and online point of interests. Netw Spat Econ 17(4):1231–1253

Bohte W, Maat K (2009) Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: a large-scale application in the Netherlands. Transp Res Part C Emerg Technol 17(3):285–297

Buijs R, Koch T, Dugundji E (2021) Using neural nets to predict transportation mode choice: Amsterdam network change analysis. J Ambient Intell Humaniz Comput 12:121–135. https://doi.org/10.1007/s12652-020-02855-6

Calabrese F, Lorenzo GD, Liang L, Ratti CJIPC (2011) Estimating origin–destination flows using mobile phone location data. IEEE Pervasive Comput 10(4):36–44

Calabrese F, Diao M, Di Lorenzo G, Ferreira J, Ratti C (2013) Understanding individual mobility patterns from urban sensing data: a mobile phone trace example. Transp Res Part C Emerg Technol 26:301–313

Chen C, Gong H, Lawson C, Bialostozky E (2010) Evaluating the feasibility of a passive travel survey collection in a complex urban environment: lessons learned from the New York City case study. Transp Res Part A Policy Pract 44(10):830–840

Chen C, Jiao S, Zhang S, Liu W, Feng L, Wang Y (2018a) TripImputor: real-time imputing taxi trip purpose leveraging multi-sourced urban data. IEEE Trans Intell Transp Syst 19(10):3292–3304

Chen C, Liao C, Xie X, Wang Y, Zhao J (2018b) Trip2Vec: a deep embedding approach for clustering and profiling taxi trip purposes. Pers Ubiquit Comput 23(1):53–66

Csáji BC, Browet A, Traag VA, Delvenne J-C, Huens E, Van Dooren P, Smoreda Z, Blondel VD (2013) Exploring the mobility of mobile phone users. Phys A Stat Mech Appl 392(6):1459–1473

Cui J, Zhao J (2021) Optimal route planning of traffic multi-source route based on granular computing. J Ambient Intell Humaniz Comput. https://doi.org/10.1007/s12652-020-02815-0

Cui Y, Meng C, He Q, Gao J (2018) Forecasting current and next trip purpose with social media data and Google Places. Transp Res Part C Emerg Technol 97:159–174

Deng Z, Ji M (2010) Deriving rules for trip purpose identification from GPS travel survey data and land use data: a machine learning approach. In: International conference on traffic & transportation studies 2010. https://ascelibrary.org/doi/10.1061/41123(383)73

Ermagun A, Fan Y, Wolfson J, Adomavicius G, Das K (2017) Real-time trip purpose prediction using online location-based search and discovery services. Transp Res Part C Emerg Technol 77:96–112

Gong L, Morikawa T, Yamamoto T, Sato H (2014) Deriving personal trip data from GPS data: a literature review on the existing methodologies. Procedia Soc Behav Sci 138:557–565

González MC, Hidalgo CA, Barabási A-L (2009) Understanding individual human mobility patterns. Nature 458(7235):238–238

Hassan MM, Uddin MZ, Mohamed A, Almogren A (2018) A robust human activity recognition system using smartphone sensors and deep learning. Future Gener Comput Syst 81:307–313

Iqbal MS, Choudhury CF, Wang P, González MC (2014) Development of origin–destination matrices using mobile phone call data. Transp Res Part C Emerg Technol 40:63–74

Jiang S, Ferreira J, Gonzalez MC (2017) Activity-based human mobility patterns inferred from mobile phone data: a case study of Singapore. IEEE Trans Big Data 3(2):208–219

Li W, Ji Y, Cao X, Qi X (2020) Trip purpose identification of docked bike-sharing from IC card data using a continuous hidden Markov model. IEEE Access 8:189598–189613

Liao L, Fox D, Kautz H (2016) Extracting places and activities from GPS traces using hierarchical conditional random fields. Int J Robot Res 26(1):119–134

Lu Y, Zhang L (2015) Imputing trip purposes for long-distance travel. Transportation 42(4):581–595

Lu Y, Zhu S, Zhang L (2013) Imputing trip purpose based on GPS travel survey data and machine learning methods. In: Transportation research board meeting 2013

Lu K, Khani A, Han B (2018) A trip purpose-based data-driven alighting station choice model using transit smart card data. Complexity 2018:1–14

Montini L, Rieser-Schüssler N, Horni A, Axhausen KW (2014) Trip purpose identification from GPS tracks. Transp Res Rec J Transp Res Board 2405(1):16–23

Ni L, Wang X, Chen X (2018) A spatial econometric model for travel flow analysis and real-world applications with massive mobile phone data. Transp Res Part C Emerg Technol 86:510–526

Nweke HF, Teh YW, Al-garadi MA, Alo UR (2018) Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: state of the art and research challenges. Expert Syst Appl 105:233–261

Oliveira MGS, Vovsha P, Wolf J, Mitchell M (2014) Evaluation of two methods for identifying trip purpose in GPS-based household travel surveys. Transp Res Rec J Transp Res Board 2405(1):33–41

Peng L, Chen L, Wu M, Chen G (2019) Complex activity recognition using acceleration, vital sign, and location data. IEEE Trans Mob Comput 18(7):1488–1498

Praveen Kumar D, Amgoth T, Annavarapu CSR (2019) Machine learning algorithms for wireless sensor networks: a survey. Inf Fusion 49:1–25

Qu Y, Gong H, Wang P (2015) Transportation mode split with mobile phone data. In: 2015 IEEE 18th international conference on intelligent transportation systems - (ITSC 2015). IEEE. https://doi.org/10.1109/ITSC.2015.56

Sansano E, Montoliu R, Belmonte Fernández Ó (2020) A study of deep neural networks for human activity recognition. Comput Intell 36(3):1113–1139

Shanmugam L, Ramasamy M (2021) Study on mode choice using nested logit models in travel towards Chennai metropolitan city. J Ambient Intell Humaniz Comput. https://doi.org/10.1007/s12652-020-02868-1

Song C, Qu Z, Blumm N, Barabasi AL (2010) Limits of predictability in human mobility. Science 327(5968):1018–1021

Stopher P, FitzGerald C, Zhang J (2008) Search for a global positioning system device to measure person travel. Transp Res Part C Emerg Technol 16(3):350–369

Suto J, Oniga S (2018) Efficiency investigation of artificial neural networks in human activity recognition. J Ambient Intell Humaniz Comput 9(4):1–12

Thakur D, Biswas S (2020) Smartphone based human activity monitoring and recognition using ML and DL: a comprehensive survey. J Ambient Intell Humaniz Comput 11:5433–5444. https://doi.org/10.1007/s12652-020-01899-y

Wang F, Chen C (2018) On data processing required to derive mobility patterns from passively-generated mobile phone data. Transp Res Part C Emerg Technol 87:58–74

Wang H, Calabrese F, Lorenzo GD, Ratti C (2010) Transportation mode inference from anonymized and aggregated mobile phone call detail records. In: International IEEE conference on intelligent transportation systems IEEE, 2010. https://doi.org/10.1109/ITSC.2010.5625188

Wang P, Liu G, Fu Y, Zhou Y, Li J (2018) Spotting trip purposes from taxi trajectories. ACM Trans Intell Syst Technol 9(3):1–26

Widhalm P, Yang Y, Ulm M, Athavale S, González MC (2015) Discovering urban activity patterns in cell phone data. Transportation 42(4):597–623

Wolf J, Guensler R, Bachman W (2007a) Elimination of the travel diary: experiment to derive trip purpose from global positioning system travel data. Transp Res Rec 1768(1):125–134. https://doi.org/10.3141/1768-15

Wolf J, Schönfelder S, Samaga U, Oliveira M, Axhausen KW (2007b) Eighty weeks of global positioning system traces: approaches to enriching trip information. Transp Res Rec 1870(1):46–54. https://doi.org/10.3141/1870-06

Wu J, Jiang C, Houston D, Baker D, Delfino R (2011) Automated time activity classification based on global positioning system (GPS) tracking data. Environ Health 10(1):1–13. https://doi.org/10.1186/1476-069X-10-101

Xiao G, Juan Z, Zhang C (2016) Detecting trip purposes from smartphone-based travel surveys with artificial neural networks and particle swarm optimization. Transp Res Part C Emerg Technol 71:447–463

Zhao H, Qian D, Lv Y, Zhang B, Liang R (2019) Development of a global positioning system data-based trip-purpose inference method for hazardous materials transportation management. J Intell Transp Syst 24(1):24–39

Zhong G, Zhang J, Li L, Chen X, Yang F, Ran B (2018) Analyzing passenger travel demand related to the transportation hub inside a city area using mobile phone data. Transp Res Rec J Transp Res Board 2672(50):23–34

Zou Q, Yao X, Zhao P, Wei H, Ren H (2016) Detecting home location and trip purposes for cardholders by mining smart card transaction data in Beijing subway. Transportation 45(3):919–944