Advanced Engineering Forum *Online: 2012-07-26*
ISSN: 2234-991X, Vol. 5, pp 3-8
doi:10.4028/www.scientific.net/AEF.5.3

Classification of mobility of cellular phone using linear classification and k-clustering

Shilong Wang, Hua Wang

School of Transportation Science and Engineering, Harbin institute of technology
School of Transportation Science and Engineering, Harbin institute of technology
wangshilongjms@126.com,wanghua@hit.edu.cn.

Keywords: multiple base station; cellular phone classification and recognition; linear classification; k–clustering.

Abstract: Road traffic data is a fundamental element of intelligent traffic system. However, due to the high investment of the road sensor, the availability of the traffic data is so limited that it can't satisfy the requirement of current situation. Using cellular phone information as road traffic data becomes an attractive alternative because of its low cost, widespread and high cover rate. Until now, there are several algorithms to process the cellular phone information and most of them present promising conclusion. In this paper, we proposed a process to collect the information of cellular phone based on the simulation of working mode of the real base station, i.e., putting an appropriate instrument on the side of the road to detect the cellular phone passing by. Using the data we got, then we proposed a method to classify the mobility of the cellular phone, which is the critical problem of the analysis of the cellular phone information. Two key attributes are the average vehicle velocity and the variance of the vehicle velocity.

I. Introduction

According to the up-to-date statistics, there are already 1.3 billion mobile phone users in China. The cellular phone penetration rate has reached 100% in some large and middle cities in China. The enormous mobile phone users can provide considerable potential traffic data for the location of mobile phone. And as the development of the 3g technique, the speed of wireless transmission is becoming quite fast[1]. These conditions put a solid foundation to the technique of traffic data detecting. However, since the work pattern of the base station, the data collected from cellular phones contain various carrier types such as pedestrians, travelers on bus, in car and by bicycle. Among these carrier types, the phones carried by travelers on buses and in cars are the objects we most care about. On the other hand, except for these two types, the rest of types (we can call them voice) can cause obvious error of analysis for traffic flow. If each carrier type of cellular phone can be accurately identified, cellular phone information can become essential source of road traffic data in the future. According to the literature, there are several algorithms to classify the mobility of cellular phone into pedestrian and vehicle, such as neural network[2], naive Bayes model and many other methods[3]. While, these methods either assume the voice has already been filter out or need a large amount real samples to train the model.

In this paper, we proposed a classification method which has two steps based on pattern recognition. And we got the cellular data from using an appropriate instrument with the similar function as the base station which can detect the cellular phone passed by it, and, at the same time, record the phone's ID and the passing time. The structure of this paper is as follows. Section 2 describes related work about the working pattern of base station. Section 3 presents preliminary process of the cellular phone data to get the velocity of the vehicle and classification of the cellular phone by two steps, linear classification and k-clustering. In section 4, analysis of the classification result is presented. We draw conclusion in section 5.

II. Related work

Cellular phones in work keep touch with the base stations cover it through GSM protocol by the way of high frequency wireless[4]. The radio frequency unit in the phone can transmit the phone's information we need to the base station. The information includes the ID of the phone and the register time. In the meanwhile, the phone can also receive signals from the base station. According to the strength of the signal, mobile phone registers to the nearest base station in order to make itself known to the network. When the phone moves from one cell into another cell, it needs to register again to the new nearest base station. The duration between the two operations of register is called cell dwell time (which is known as CDT) as shown in figure 1 and Eq.1.

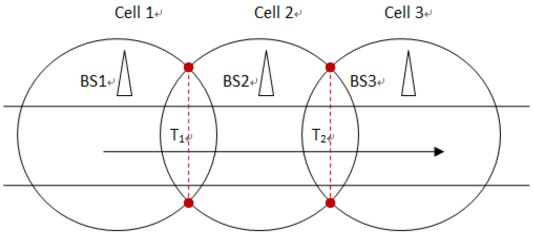

Fig 1 Principles of CDT

$$CDT=T_1-T_2 \tag{1}$$

III. Methodology

A. Data collection. Cellular phone data is collected by several instruments which have similar function as the base station. They are able to collect the ID and the corresponding register time of the phone passing by the instrument and store the data into data base. Every single instrument consists of radio frequency antenna and mainframe box and it can be powered by battery of car.

We choose the Bridge of Songhua River as the experiment location because of following reason. 1. The component of the traffic flow on the bridge is as same as the component of city area which consists of vehicle, bus, bicycle and pedestrian and there are some bus stops on the bridge.2.There is no traffic signal on the bridge. Therefore, the distributions of the velocity of each means of transportation conform to normal distribution which is easy to analyse.

The data collecting took place at day time from 7:30 to 10:00. The locations of the six instruments number 1 to number 6 are assigned as figure 3 shows.

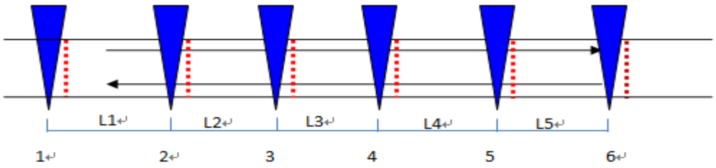

Fig 2 The locations of instruments

The triangle area represent the detect area of the instrument and L1-L5 represent the interval distance between the two close instruments. And the dotted lines represent the imaginary line to count the real traffic volume and the velocity of floating car.

Part of the original data we collected is shown in table 1.

Table 1 The original data

Instrument number	International identity number	Register time
1	460000334311710	25/03/2010 07:37:19
2	460021345529663	25/03/2010 07:48:07

Then we use each ID of the each phone to match its register time from each instrument by the software. During the experiment time, we set several floating cars to shuttle between the experiment area to get the real value of the velocity of cellular phones in it, and we've already known these phones' ID. Up to now, we have got the original data.

B.Preliminary data process. After the register time matching, we are able to work out the CDT of each phone between each two instruments. And the distance between each two close instruments we've already known. Therefore the estimated value of velocity of each phone can be calculated as Eq.2.

$$V=L/CDT \tag{2}$$

And the real values of velocity of some phones have been recorded by the method of floating car. Part of the consequence of contrast is listed in table 2.

Table 2 Error of estimation

Real value	Estimated value	Error
35.4 km/h	34.2 km/h	3.39%
25.7 km/h	25.2 km/h	1.95%
27.8 km/h	26.9 km/h	3.24%
40.4 km/h	42.6 km/h	5.45%

As shown in table 3, the maximum error of velocity estimation is lesser than 6%, which means the method reach a very high accuracy of velocity estimation. The high accuracy of velocity estimation put a solid foundation to the phase of data classification. 6 instruments have 5 intervals and we could get 5 CDTs from which the 5 values of velocity of each cellular phone calculated as shown in table 3.

Table 3 Result of preliminary process ,km/h

V_{1-2}	V_{2-3}	V_{3-4}	V_{4-5}	V_{5-6}	standard deviation	average
5.98	5.992786	6.055501	5.870521	5.752636	0.12	5.93
27.51	26.47069	25.47287	4.484469	4.680591	4.89	11.96
15.51	15.17947	15.7017	15.59274	15.5054	0.19	15.50
35.74303	35.62304	32.64834	5.178054	5.133856	15.21	24.73
24.71582	25.35427	26.28661	27.65894	27.55467	1.59	25.87
10.36	10.78132	10.557	10.29623	10.89844	0.26	10.58

Thus, every phone got through the experiment area can be denoted by follow vector as Eq.3.

$$\mathbf{P_i}= (V_{1-2}, V_{2-3}, V_{3-4}, V_{4-5}, V_{5-6}) \tag{3}$$

Where $\mathbf{P_i}$ is to represent the cellular phone passed through all the instrument.V_1-V_5 are the values of velocity of corresponding interval.

C. Classification of mobility. Consider the travel between instrument 1 and instrument 6 could be finished by 2 sorts of traffic modes, i.e., mode 1, the phone carried only by just one kind of means of travel within the whole journey, such as only by pedestrian, only by vehicle or only by bicycle. Mode 2, there is at least one time of switch of means of travel happened within the journey between instrument 1 and instrument 6. There are so kinds of switch within the journey that we can not use k-clustering directly. Here is a simplest instance, one phone carrier travels from instrument 1 to the instrument 3 by bus, and finish the rest of travel from instrument 3 to instrument 6 on foot. It is easy to draw a conclusion that the variance of series estimated velocity of mode 1 could be much lesser than mode 2 in the normal condition in which we took our experiment. Also consider that there are pedestrian, vehicle and bicycle, 3 types of phone carrier. And we can group these 3 types into two types for the reason that in traffic regulation and control, the flow of vehicle in mode 1 is what we most care about. So, the two new carrier types are motor vehicle type (such as car, bus, truck) and non-motor vehicle type (such as bicycle and pedestrian). And these two types' velocity has remarkable distinction from each other. Thus, the average of 5 values of velocity of one single phone is fit for being the characteristic value to involve in the classification. But, before we finish separating mode 1 from mode 2, we could not use this value to classify the types of carrier because of the average value of velocity of each phone can be full of the one dimension number axis and no obvious interval can be used to classify. Until now, we've got two characteristic value which can be applied to classify, i.e. σ_i=variance(P_i) and μ_i=average(P_i). Finally we rewrite the feature vector of $\mathbf{P}_i$ as Eq.4.

$$\mathbf{P}_i=(\sigma_i,\ \mu_i) \qquad (4)$$

The scatter diagram of $\mathbf{P}_i$ (about 500 points collected in 10 minutes)is shown in figure 4.

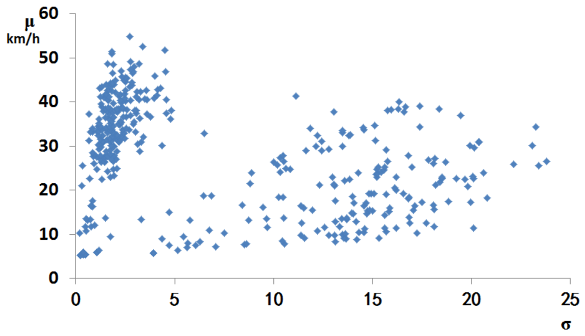

Fig 4 The scatter diagram of $\mathbf{P}_i$

As shown in figure 4, the two kinds of traffic mode are obviously grouped into two areas in the diagram. Considering the accuracy of using K-clustering right now is not enough, we take the method of linear classification as the first step processing.

$$w(k+1)\begin{cases} w(k) & ,\mathbf{w}^{\mathrm{T}}(\mathrm{k})\mathbf{x}_{\mathrm{k}} > 0 \\ w(k)+p\mathbf{x}_{\mathrm{k}} & ,\mathbf{w}^{\mathrm{T}}(\mathrm{k})\mathbf{x}_{\mathrm{k}} \le 0 \end{cases} \qquad (5)$$

Step 1: linear classification. In order to separate voice mode 1 from mode 2 and improve the accuracy of next classification, we use linear classification which perform perfectly in classifying two kinds of problem. The training procedure is processed by the principle above in Eq.5. The vector **w** is the augment weight vector. And the so called samples used to train the weight vector are extracted from data collected by the instruments, therefore, in fact, we didn't use any real sample to train the model. And by conscribing the longevity of samples extracting, we can also get a good real-time property. The result of linear classification is shown in figure 4 and figure 5.

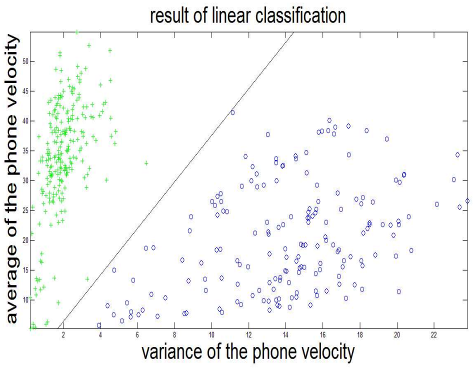

Fig 4 Result of linear classification

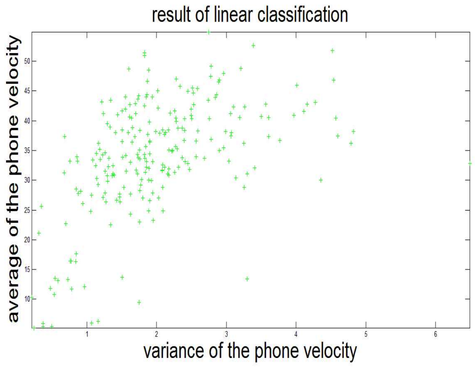

Fig 5 Result of linear classification

As shown in figure 4, the cycles under the line present the mode 2 and the crosses present the mode 1. The result of separating the cycle from the cross is shown in figure 5. Until now, the mode 1 has been separated out waiting for the next processing.

Sept 2: K-clustering. K-clustering is an unsupervised learning algorithm and when the different clusters in the scatter diagram are apparent, the effect of clustering could be accurate enough just like the condition in our case. The result of K-clustering is shown in figure 6.

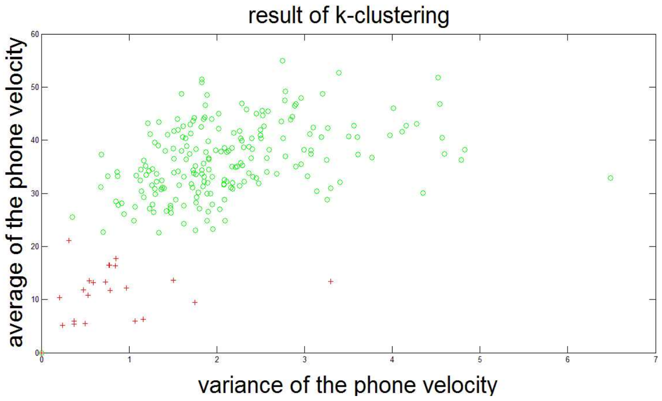

Fig 6 Result of k-clustering

Finally, the circles are the cellular phones of vehicle type in mode 1. The parameters of this phone type are what we need for the traffic regulation and control because that they can represent the parameter of real traffic flow.

IV. Analysis of result

In order to evaluate the result of classification, some statistics of the crosses data (mode 1, vehicle) and the circles data (mode 1, non-vehicle) in figure 6 have been done. The frequency distribution histogram of the crosses data and the probability density distribution of crosses data are shown in figure 7 and figure 8 respectively.

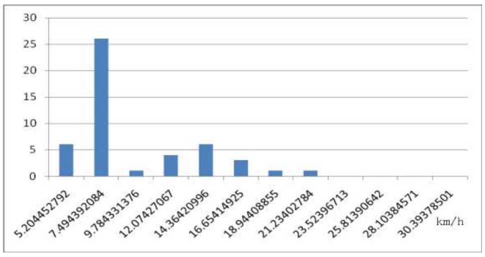

Fig 7 The frequency distribution histogram

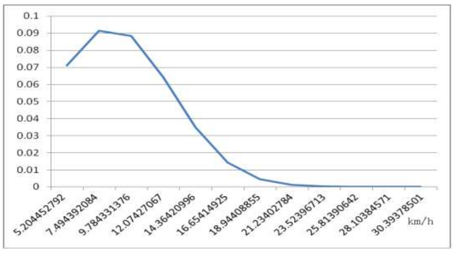

Fig 8 The probability density distribution

As shown in figure 7 and figure 8, no obvious regular pattern of distribution has been found. According to the knowledge of statistic, we can get the maximum value and the minimum value of velocity respectively is 21.07km/h and 5.2km/h and they are basically corresponded to the traffic characteristic of non-vehicle.

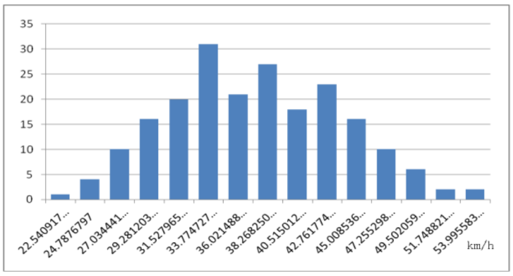

Fig 9 The frequency distribution histogram

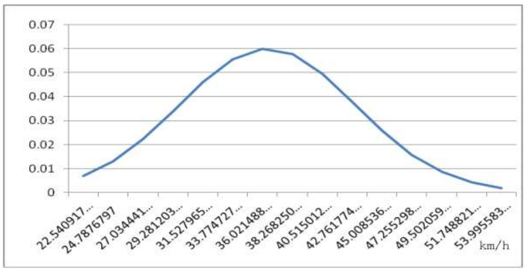

Fig 10 The probability density distribution

As shown in figure 9 and figure 10, the distribution of circles data accords with normal distribution, with μ=36.37km/h and σ=6.63, which is corresponded to the traffic characteristic of vehicle.

V. Conclusion

Cellular phone as the probe to detect the information of traffic flow has been becoming more and more attractive because of its advantage of high cover rate and low cost. And the essential problem of processing the cellular phone data is to separate the voice (non-vehicle phone) from the vehicle phone. In this paper, we proposed a new method to classify the mobility of cellular phone by two steps, i.e. 1) linear classification to separate the traffic mode with switches from the traffic mode with no switch. 2) K-clustering to classify the mode with no switch into two types of mobility, vehicle phone and non-vehicle phone. Based on the statistics of the classification result, we could believe that this method has an encouraging performance.

Acknowledgments

This research was supported by the China National Science Foundation Grant No. 51138003.

Reference

[1] Schneider W., Mrakotsky E. Mobile Phones as a Basis for Traffic State Information.Proceedings of the 8th International TB5.1 .IEEE Conference on Intelligent Transportation Systems Vienna, Austria, September 13-16, 2005.

[2] Wasan Pattara-atikom ,Ratchata Peachavanish. Estimating Road Traffic Congestion from Cell Dwell Time using Neural Network. 1-4244-11 78-5/07(C)2007 IEEE.

[3] Anum L.EnlilCorral-Ruiz,Felipe A.Cruz-Pérez,and Genaro Hernández-Valdez. Teletraffic Model for the Performance Evaluation of Cellular Networks with Hyper-Erlang Distributed Cell Dwell Time.

[4] Arwa Zabian.Mobile Cellular Networks and Traffic Road:a new Paradigm.Department of CIS Irbid National University Irbid-Jordan.