Study of Estimate Human Demographic Attributes Using Person Flow Datasets and Apply It for GPS Log Data

Takahiro NISHIMURA[1], Yuki AKIYAMA[2], Ryosuke SHIBASAKI[3] and Yoshihide SEKIMOTO[4]

In recent years, we have high-function mobile device called smartphone. As a result, we can get various information; location, time, gyro and so on. To analyze mobile phone GPS log data is specially focused on but it is hard to get validation data. We focus on Person Flow dataset (PFlow dataset) made from Person Trip survey (PT) that has label of demographic attributes, transfer history and develop model from it is applied for GPS log data using Transductive Transfer Learning. We developed gender, age and work type classifier model from PFlow dataset. As a result, we can classify high accuracy for these demographic attributes. ZENRIN DataCom Co. LTD provided us comprehensive and analytical processing of the GPS data.

1. Introduction

In recent years, customer tendency has getting diversified in Japanese society. As a result, many companies in Japan have changed marketing method from mass marketing to targeting marketing. Therefore companies are making targeted marketing activities that are optimized for each customer by analysis of their activity log data. Especially, they can get various log data from mobile phones with permission from phone users e.g. GPS log data, gyro, SNS based data and so on, because of their high functionality. However it is needed to estimate demographic attributes of users by any methods to utilize log data for targeted marketing because they contain only time-series location and velocity information of users. Moreover, it is difficult to develop classification models of demographic attribute based on only log data because acquisition of location data with demographic attributes are not easy.

In this study, we develop a classification model of estimate attributes to use Person Flow Dataset (PFlow data) that is similar to GPS log data with permission from users. The Person Flow Dataset is high-resolution data of Person Trip Survey (PT) that is government census. It contains location data and demographic attributes of randomized persons. In addition, we estimated feature distribution of visitors in arbitrary times and areas to apply models from Person Flow Dataset to GPS log data.

2. Approach

2.1 Person Flow dataset

In this study, we estimated demographic attributes of users using the Person Flow data (PFlow data). This data has many attributes of each person: user id, location, purpose of movement, gender, age and work type. Table 1 shows demographic attributes of PFlow data.

2.2 Extract features, learning and validation

[1] Graduate School of Frontier Sciences, The University of Tokyo
[2] Earth Observation Data Integration and Fusion Research Institute, The University of Tokyo
[3] Center for Spatial Information Sciences, The University of Tokyo
[4] Institute of Industrial Sciences, The University of Tokyo

* This is an abstract peer-reviewed by the International Affairs Committee of the City Planning Institute of Japan for ISCP2014.

In order to estimate demographic attributes, we used the Support Vector Machine (SVM) that is very high classification ability because this method uses non-liner classification called the Kernel Trick. We used the Gaussian Kernel that is often used in the SVM to estimate attributes. Table 2 shows extracted variables by the SVM. This study developed 3 models: gender, age and work type, using same variable. In order to estimate them by the SVM, we need to decide sample sizes because the SVM needs many calculations if it uses all samples. . They are calculated by equation 1.

$$argmin(T(n)/E(n)) \quad (1)$$

Where *T* is processing time, *E* is generalization error and *n* is the number of sample. Stay points were decided based on the user ID in the PFlow data. Accumulating threshold of stay point is referred in Hadano et al. (2012). In order to validate models, this study developed validation dataset from the PFlow data and calculated accuracy, recall and precision of models.

Table 1. Demographic attributes of PFlow data

attribute No	Gender	Ages	Work type
1	male	-5	agricalture workers
2	female	10-15	Industiral workers
3		15-20	sales workers
4		15-20	Service position workers
5		20-25	Transportaion and Communications workers
6		25-30	Preservation workers
7		30-35	clerical workers
8		35-40	professional workers
9		40-45	manager
10		45-50	other occupation emplyee
11		50-55	Elementary / Junior Highschool student
12		55-60	Highschol Student
13		60-65	University Student
14		65-70	housewife
15		70-75	No jobs
16		75-80	Others
17		80-85	unknown
		85-	

Table 2. Extracted variables

extract variables	unit
departure time of estimate residential point	hour
arrival time of estimate residential point	hour
departure time of estimate work point	hour
arrival time of estimate work point	hour
visit times of residencital point	
visit times of work point	
total amount of movement distance in day	m/day
amount of movement distance per hour	m/hour
variances of time of stay point	
number of stay point	

2.2 Application of the Pflow model to GPS Log data

In order to apply models from the PFlow data to GPS log data, this study used the transfer learning. It is a research problem in machine learning that focuses on storing knowledge gained while solving one problem and applying it to a different but related problem. This method classified 4 types based on label existence of source data and target data. GPS log data does not have label though the PFlow data has it. Therefore, the method in this study is Transductive Transfer Learning. In addition, Transductive Transfer Learning needs to satisfy equation 2. This equation means conditional probability of both dataset is same value.

$$P[Y^{(s)}|X^{(s)}] = P[Y^{(t)}|X^{(t)}] \quad (2)$$

However it is expected to satisfy equation 2 by the law of large numbers because sample sizes of both data are about 0.6 million. Finally this study applied the model of the Pflow data to GPS log data and aggregated it into 1km square grid in arbitrary times.

3. Results and Discussion

3.1 Result sample size problems and learning use raw labels of PFlow data

Fig. 1 shows results of learning with different number of samples. This result shows that 5000 samples are adequate to develop the model. Table 3 shows learning result of gender, age and work type using extracted variables in table 2. Table 3 shows that classification ability of the

SVM is low. It is considered that learning of gender is difficult by variables in this study and classification abilities of age and work type are low because numbers of their class are large. Therefore, this study aggregated labels in age and work type and developed model again. Tables 4 shows aggregated labels.

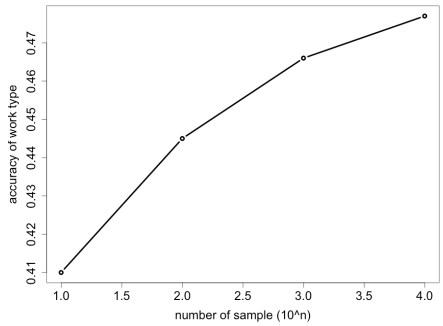

Fig 1. Result of learning change of number of samples

Table 3. Result of learning

Attribute	accuracy
Gender	0.67
Ages	0.25
Work type	0.4

Table 4 aggregate labels

raw label of ages	aggregate labels of ages	raw labels of work type	aggregate labels of work type
-5	under 10s	agricalture workers	Workers
10-15		Industiral workers	
15-20	10s	sales workers	
15-20		Service position workers	
20-25	20s	Transportaion and Communications workers	
25-30		Preservation workers	
30-35	30s	clerical workers	
35-40		professional workers	
40-45	40s	manager	
45-50		other occupation emplyee	
50-55	50s	Elementary / Junior Highschool student	Students
55-60		Highschol Student	
60-65	60s	University Student	
65-70		housewife	Housewife
70-75	70s	No jobs	Others
75-80		Others	
80-85	over 80	unknown	
85-			

3.2 Learning result after aggregation of labels

Table 5 shows learning result after aggregation of labels. We observed large f-measure value in work type. On the other hand, accuracy of age has improved.

Table 5 result of learning after class aggregation

Attribute	accuracy	recall	precision	f-measure
Gender	0.67	0.644	0.64	0.64
Ages	0.37	0.33	0.29	0.244
Work type	0.82	0.74	0.69	0.711

3.3 Result of application for GPS log data

The models from PFlow data after label aggregation was applied to GPS log data. Figure 2 and figure 3 show gird maps of gender ratio aggregated by 1km square grid in October 1, 2010 at 12:00 and 20:00. Grid color shows gender ratio. Male ratio is large in red grids and female is in green. ZENRIN DataCom Co. LTD provided us comprehensive and analytical processing of the GPS data.

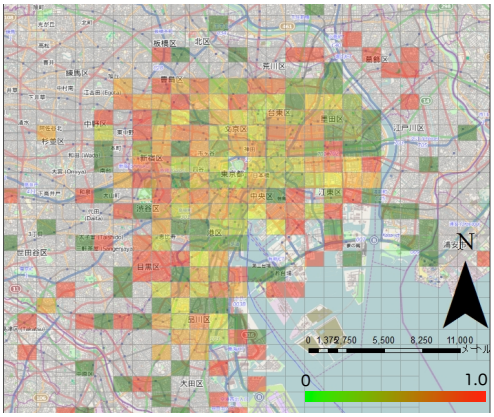

Fig 2. estimate gender ratio in 1/October/2010, 12 pm[1]

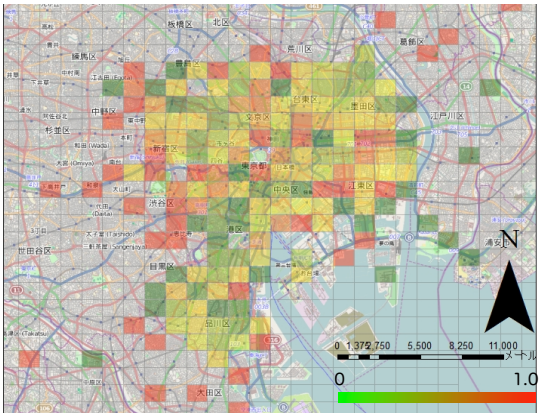

Fig 3. estimate gender ratio in 1/October/2010, 20 pm[1]

4. Conclusion

This study estimated human demographic attributes using user's movement features from the Person Flow Dataset that contains demographic attributes and location data of each person. In addition, this study considered a similarity between the Pflow data with GPS log data by mobile phones to apply the transductive transfer learning to it. It is revealed that learning is processed with adequate accuracy to use 5000 samples in the Pflow data in pre training. In addition, Result of learning by the SVM can estimate work types with high accuracy. On the other hand, it cannot estimate gender and age adequately. Finally, we can observe regional changes based on estimated features of persons in arbitrary times and areas to apply the model to GPS log data.This study uses only movement features of each user. Therefore, we try to improve the model to use detail feature of stay area, and to develop models in another area.

Keywords: GPS log data, pflow data, transductive transfer learning

Note:

1) Raw GPS data processing in all figures were provided by ZENRIN DataCom Co.LTD.

Acknowledgement:

The authors were given the "Konzatsu-Tokei Data®" from ZENRIN DataCom Co.LTD. The data is the person flow data developed by comprehensive and analytical processing of the GPS data with permission from phone users. The authors would like to thank CSIS for their contribution.

5. References

1) People Flow Project, http://pflow.csis.u-tokyo.ac.jp/index-j.html.

2) Akiyama, Y., Takada, T. and Shibasaki, R., 2013. Development of Micropopulation Census through Disaggregation of National Population Census, CUPUM2013 conference papers, 110.

3) Mayumi H., Ueyama S., Akiyama Y., Horanont T., Shibasaki R., 2012. Study of extracted method of visitor's behavior pattern for Commerce Accumulations, 21th Conference of Geographic Information System, F-3-4. (in Japanese)