Comparison of Activity Type Identification from Mobile Phone GPS Data Using Various Machine Learning Methods

Lei GONG*, Toshiyuki YAMAMOTOb, Takayuki MORIKAWAc

a Department of Civil Eng., Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Japan 464-8603, Japan; E-mail: Leigongchi@gmail.com
b Institute of Materials and Systems for Sustainability, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan; Email: yamamoto@civil.nagoya-u.ac.jp
c Institute of Innovation for Future Society, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan; Email: morikawa@nagoya-u.jp

Abstract: Ever since global positioning system (GPS) modules have been attached to smart phones, much research has focused on how to obtain personal trip (PT) information from them. One of the challenges is identifying activity type (or inferring the purpose of the trip) from these continuous GPS data. This paper focuses on obtaining the type of activity using several machine learning methods and comparing the results. The comparison is implemented from the perspective of accuracy and time cost in the phases of data training and prediction. After applying four machine learning methods to the data set obtained from 30 individuals in Nagoya, Japan, a classification tree method demonstrates superiority over support vector machine (SVM), neural network (NN), and discriminant analysis methods.

Keywords: Activity Type Identification, Mobile Phone GPS Data, Support Vector Machine, Discriminant Analysis, Classification Tree, Neural Network

1. INTRODUCTION

Traditional interview-based surveys that gather personal trip (PT) data have several disadvantages, such as inaccurate departure/arrival time, unreported trips, and so on. So when global positioning system (GPS) techniques became available in the civilian field, GPS modules became popular on smart phones. This led more researchers to focus on how to derive PT data from these GPS data. When the GPS signal is good, GPS data can provide accurate spatial and temporal information about object persons. However, information about trip segmentation, activity location, transportation mode, and trip purpose need to be obtained from the GPS data by methodologies or algorithms with satisfactory accuracy. As a result, many researchers have contributed to this field and tried to achieve better accuracy by advancing and testing many methods. Activity type identification (or trip purpose inference) is one of the topics in this field and different types of methodologies have been applied. However, most of these existing methodologies are rule-based methods, and machine learning methods have been rarely applied in this field (Gong et al., 2014).

One disadvantage of rule-based methods is that the rules may vary depending on features of data sets; rules suitable for one data set may lead to low accuracy of another data set. In contrast, machine learning is believed suitable for classification based on the features and potential algorithms learned from a given training data set. It achieves satisfactory results when applied to transportation mode identification. In this paper, we tested several machine learning
methods for identifying activity types from mobile phone GPS data. To be specific, support vector machines (SVMs), discriminant analysis (DA), classification trees, and neural networks (NNs) are applied to the mobile phone GPS data collected in the Nagoya metropolitan area in Japan.

A person’s whole day consists of trips and activities, and the notion of trip purpose is almost equivalent to activity type. Moreover, a trip is actually a product of an activity, and the purpose of stopping at a trip end is to do a certain kind of activity. Since the results of this paper are one of the preliminary steps for activity-based model analysis, here we use “activity type” rather than “trip purpose”.

The rest of the paper is organized as follows: part two describes the existing machine learning methods used for identifying activity type. It is followed in part three by a description of the data set used for machine learning methods. Part four describes the methodologies used in this paper, which is followed in part five by a description of the attributes selected as independent variables. In part six, the results of each methodology are interpreted and possible reasons for low accuracy for some activity types are discussed. The last part documents the conclusions and future research trends.

2. LITERATURE REVIEW

Gong et al. (2014) summarized the existing methodologies utilized for identifying activity type or inferring trip purpose from GPS data. Methods can be categorized into three groups: rule-based methods, probabilistic methods, and machine learning methods.

Rule-based methods are believed to be the most popular in this field (Bohte and Matt, 2009; Chen et al., 2010; Pereira et al., 2013; Stopher et al., 2005, 2008ab; Wolf et al., 2001). Methods in this category match the selected information from GPS, GIS, and respondents with a series of predefined heuristic rules to infer the trip purpose. The second category is probabilistic methods (Axhausen et al., 2004; Chen et al., 2010). The probability of each purpose is calculated based on the different values of information from GPS, GIS, and respondents’ characteristics. The estimated trip purpose is decided by the calculated probability. Machine learning methods, which have been widely used for transportation mode detection, however, are as yet rarely used for activity type identification or trip purpose inference. Only discriminant analysis and decision trees have been applied. Deng and Ji (2010) used a decision tree method to infer trip purpose and obtained an accuracy of 87.6%. Their independent variables included trip ending time, speed, mode, trip distance, trip duration, occupation, income, family structure, age, and land use. McGowen and McNally (2007) used discriminant analysis and a classification tree model to identify activity type using land use data and demographic data of the object person. They obtained 73% and 74% accuracy, respectively, with these two methods.

Wolf et al. (2014) also summarized the methods used for trip purpose identification from GPS data. Two methods are related to machine learning methods. One was by McGowen and McNally (2007) mentioned above; the other is a decision tree method implemented by Griffin and Huang (2005). They applied the C4.5 algorithm to build a decision tree capable of classifying trip ends into multiple trip purposes.

In the existing research, few machine learning methods have been tried for identifying activity types. In particular, a comparison of utilizing machine learning methods for activity type identification or trip purpose inference is yet to be examined. Moreover, among several types of machine learning methods, the one that is most suitable for identifying activity type based on accurate results and time cost is still unknown. In this paper, we test several machine
learning methods on the GPS data from mobile phones in Japan in an endeavor to fill this gap and answer these questions.

3. DATA SET DESCRIPTION

The GPS data utilized in this research were collected by 30 volunteers in the Nagoya area, in Japan over 5 weeks (from Sept. 24 to Oct. 30) in 2008. Each volunteer was assigned a mobile phone with GPS module that can record and send GPS information to a server every 10 seconds. However, sometimes the GPS module sent GPS information in an interval longer than 10 seconds when there was signal delay/loss in tunnels, subways, etc. Overall, 97.4% of the GPS sending intervals were less than 20 seconds.

The GPS information sent back to the server included longitude, latitude, time stamp, signal quality, etc. Methodological methods advanced in Gong et al. (2015a) were implemented to identify the activity location for trip end from continuous GPS trajectories. To obtain ground truth for our research, volunteers were required to mark information by inputting the start, end, mode, and purpose of each trip through a software app installed on the smart phones.

In addition, sociodemographic information about each volunteer was collected through questionnaires, including addresses of home and workplace, occupation, annual household income, driving license, usage frequency of auto/public transit, and so on. Figure 1 demonstrates the basic aggregated statistical information of the data set used in this research. Almost all volunteers were aged 20–65 years, which is equivalent to the work force age in Japan; almost all had a full-time or part-time job, which means they were active trip makers. The age of the volunteers lay mostly between 20 and 34 years, which is skewed compared with the general population; this is partly because the survey format entailed using a mobile phone. Based on our previous study (Gong et al., 2015b), it is suggested that the demographic variables have a limited influence on the accuracy of activity type estimation compared with trip and activity characteristics, so the influence of having an excess of younger volunteers is not addressed in this study. Auto, walking, and rail are the main modes of transport, while the three main trip purposes in this data set are for business, returning home, and going to or from work/school.
Since the basic unit for analysis in this paper is activity, activities between two consecutive trips are generated and in total we logged 1913 activities divided among five activity types. The return from company/school type is combined with the work/school type, since these two trip purposes have the same activity at the location of trip ends.

Detailed information relating to activity is interpreted as follows. While work and business activities have a relationship with each other, the definitions are different. In our data sets, work is the activity that a volunteer going to the work place does, no matter whether he/she travels from home or from somewhere else. In this study, business is work-related business and it can start from the work place or from other locations, such as other business locations (in this case, the volunteer visits a series of places to carry out business activities). Business may even originate from the volunteer’s home directly (in this case, the volunteer does not need to go to the work place first, or he/she is self-employed). Recreation includes shopping, having a meal at a restaurant, or pursuing amusement options. Finally, “others” consists of activities that are not included in the former four types of activities, such as medical care, pick-up/drop-off, strolling, etc. Note that while it may be better to use more detailed and domain-clear categories of activity, the activity category in this paper is limited by the survey design. Because of this, some activity may be aggregated improperly and some activities in different categories may have homogeneity to some extent. For example, it may be better to split the three types of activities in recreation activity and recategorize them with activities in others. We will discuss this in more detail in the Results and Discussion section.

All activities were randomly stratified into two subsets: 70% of the data went into a training set and 30% into a test set. The training set was used to train the model of machine learning methods and the test set was used to test the effect of what had been learned in the training set.

4. METHODOLOGY

4.1 Support Vector Machines

SVMs are a supervised machine learning method that can be used for classification or regression analysis. It was invented by Vladimir N. Vapnik and the current standard incarnation (soft margin) was proposed by Corinna Cortes and Vapnik in 1993 and published in 1995 (Cortes and Vapnik, 1995).

As for classification, SVMs divide the training data set by a hyperplane, which maximizes the margin between two classes in the first step. In the second step, SVMs apply what they learned in the training data set to the predicting data set and implement classification. This hyperplane can be linear or nonlinear, depending on whether the input data are linear or nonlinear. For nonlinear data, the kernel function is used in SVMs to map the points in the data set that are linearly separable in the higher dimensions by a nonlinear mapping function \(\phi\). The separating hyperplane in the higher dimensions can be represented by the following formula:

\[
\omega^T \phi(x_i) + b = 0
\]

where \(\omega\) is the weight vector normal to the hyperplane, \(x_i \in R^n, i = 1, \ldots, l\) are n-dimensional vectors needed to be trained or predicted into two classes, and \(b\) is the intercept associated with decision boundaries. A Gaussian kernel function is used in this paper.

When an SVM is applied to multiclass classification, one solution is to combine several binary classifiers. Depending on the combining strategy, several methods, such as one-against-one SVM, one-against-rest SVM, and directed acyclic graph support vector machine (DAG SVM) can be used. Since a one-against-rest SVM shows a relatively high accuracy compared
with the other two methods (Gong et al., 2015b), in this paper, we use one-against-rest SVM.

One-against-rest SVM constructs \( k \) SVM classifiers where \( k \) is the number of classes. As for the \( i \)-th SVM classifier, items in the \( i \)-th class will be marked as positive labels while items in all other class will be marked as negative labels. The \( i \)-th SVM classifier solves the following problem (Hsu and Lin, 2002).

\[
\min_{\omega^i, b^i, \xi^i} \quad \frac{1}{2} (\omega^i)^T \omega^i + C \sum_{j=1}^{l} \xi^i_j \\
\quad \quad \quad \quad \left( (\omega^i)^T \phi(x_j) + b^i \geq 1 - \xi^i_j, \text{if } y_j = i, \right) \\
\quad \quad \quad \quad \left( (\omega^i)^T \phi(x_j) + b^i \leq -1 + \xi^i_j, \text{if } y_j \neq i, \right) \\
\quad \quad \quad \quad \xi^i_j \geq 0, j = 1, \ldots, l,
\]

where \( l \) is the sample size of the data set; for each item \((x_i, y_i), \quad y_i \in \{1, \ldots, k\}\) is the class label of \( x_i \); \( \phi \) is the function mapping data \( x_i \) to a higher dimensional space; \( C \) is the misclassification cost parameter (or penalty parameter); \( b \) is the intercept associated with decision boundaries; and \( \xi^i_j \) is slack variables.

After solving problem (2), there are \( k \) decision functions as follows.

\[
(\omega^1)^T \phi(x) + b^1, \\
\vdots \\
(\omega^k)^T \phi(x) + b^k.
\]

Then, \( x \) is assigned into the class that has the largest value of the decision function.

\[
\text{class of } x \equiv \arg \max_{i=1, \ldots, k} (\omega^i)^T \phi(x) + b^i
\]

In practice, the dual problem of (2) is solved and it has the same number of variables as the number of data points in (2). Consequently, \( k l \)-variable quadratic programming problems are solved.

Before the one-against-rest SVM is used for training on and predicting from our data sets, optimal parameters of cost \( C \) in the optimization function and gamma \( \gamma \) in the Gaussian kernel function are tested in pairs by iteration as follows: from \(-1\) to \(30\) with a step of \(1\) for \( C \) and from \(-30\) to \(1\) with a step of \(-1\) for \( \gamma \). Finally, the optimal pair of \((C, \gamma)\), \((222, 2^{-15})\) is used for training and predicting.

### 4.2 Neural Networks

Neural networks are believed to be capable of handling nonlinear pattern recognition, but are difficult to interpret. A neural network consists of three layers: input layer, hidden layer, and output layer. With the known correct output values, the network can learn the unrevealed patterns between input and output values by cycles of training in the hidden layers. Then, this trained network can be used to predict the output with new input values.

Backpropagation (BP) neural networks are a popular kind of neural network used in practice and are usually considered to represent a supervised machine learning methodology. Backpropagation means “backward propagation errors”. It is usually used in conjunction with an optimization method such as gradient descent. In an attempt to minimize the loss function, for each training iteration, the current gradient of a loss function with respect to all the calculated weights in the network is evaluated. Then, the gradient is fed to the optimization method, which employs it to update the weights (Sayed and Baker, 2015).

Before training the network, the topological structure of the network needed to be designed. The topological structure includes a number of hidden layers and a number of neurons in each hidden layer. There is a consensus that the performance difference from adding
additional hidden layers—the situations in which performance improves with a second (or third, etc.) hidden layer—is very small (Hassen, 2013). One hidden layer is sufficient for the large majority of problems. Regarding the number of points in a hidden layer, there are several formulas (Sheela and Deepa, 2013; Stathakis, 2009) or rules (Karsoliya, 2012) that can be used for calculation, but they generate widely differing results. In this paper, the designed neural network structure contained one hidden layer and the number of neuron points in the hidden layer were tested from 1 to 100 (this range can satisfy almost all the formulas or rules) in order to reach a satisfactory accuracy. The points in the hidden layer achieving the highest accuracy were used for training the network and for prediction. Finally, we found that 31 neuron points in the hidden layer could achieve the best accuracy of 89.2%. The neural network structure utilized in this research is shown in Figure 2.

4.3 Discriminant Analysis

DA is usually used in statistics, pattern recognition, and machine learning to characterize or separate two or more classes of items or examples by a combination of attributions. DA is closely related to multivariate analysis of variance (MANOVA) and shares the assumptions of MANOVA (Poulsen and French, 2008). One of the vital assumptions is related to normal distribution: it is assumed that the data represent a multivariate normal distribution, and different classes generate data based on different normal distributions. However, violations of the normality assumption are not “fatal” and the resultant significance test is still reliable as long as non-normality is caused by skewness and not outliers (Tabachnick and Fidell, 1996).

DA can be linear or quadratic depending on whether variance–covariance matrices are homogeneous or heterogeneous. That is, linear DA requires different classes to share the same covariance structure (homogeneous), while quadratic DA does not have this constraint. Based on the features of the data set employed in our research, quadratic DA is used in this paper.

DA creates a classifier that will minimize the possibility of misclassifying cases into their respective groups or categories. The classifier is trained by estimating parameters of a normal distribution for each class. Then, to predict the classes of new data, the trained classifier finds the class with the smallest misclassification cost.

\[
\hat{\gamma} = \arg \min_{y=1,\ldots,K} \sum_{k=1}^{K} \hat{P}(k|x)C(y|k)
\]

where \(\hat{\gamma}\) is the predicted classification; \(K\) is the number of classes; \(\hat{P}(k|x)\) is the posterior probability of class \(k\) for observation \(x\); and \(C(y|k)\) is the cost of classifying an observation as \(y\) when its true class is \(k\),
where $\hat{P}(k|x)$ is calculated as follows.

$$\hat{P}(k|x) = \frac{P(x|k)P(k)}{P(x)}$$

$$P(x|k) = \frac{1}{2\pi|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)\right)$$

where $|\Sigma_k|$ is the determinant of $\Sigma_k$, and $\Sigma_k^{-1}$ is the inverse matrix.

### 4.4 Classification Tree

A classification tree is a method based on a series of rules in an optimal order. Rules are tested in the form of a chain of simple questions. After replying to the question at each split node, the answer decides the direction of which child node will be chosen as the next split node. To be more specific, a classification tree predicts a category of an item by following a series of decisions in the tree from the root (beginning) node down to a leaf node (the final child node). Leaves represent class labels and branches represent conjunctions of features that lead to those class labels. These questions-and-answers are actually a set of logical if–then conditions for classifying items into their corresponding categories.

A basic way to build a classification tree from labeled examples proceeds in a greedy manner: the most informative questions are asked as near to the root as possible in the hierarchy. In a greedy algorithm, the first question is designed to get the two children subsets consisting of sample cases of the same class in each subset as purely as possible. After the first subdivision is done, one proceeds in a recursive manner, by using the same method for the left and right child sets, designing the appropriate questions; this procedure is repeated until the remaining sets are sufficiently pure to stop the recursion. At this juncture, the leaves are reached by following the chain of questions descending from the root, and then the final remaining set in the leaf should be almost pure, i.e., consisting of sample cases of the same class (Battiti and Brunato, 2014).

Algorithms for constructing trees usually work top–down, by choosing a variable at each step that best splits the set of items. Different algorithms use different metrics for measuring “best” or purity. In this research, we use Gini Impurity (GI) to measure “best”. GI (Battiti and Brunato, 2014) is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it were randomly labeled according to the distribution of labels in the subset. It is computed as the expected value of the mistake probability; as usual, the expectation is given by adding, for each class, the probability of mistaking the classification of an item in that class (i.e., the probability of assigning it to any class but the correct one: $1 - p_i$) times the probability for an item to be in that class ($p_i$). Suppose that there are $m$ classes, and let $f_i$ be the fraction of items labeled with value $i$ in the set. Then, by estimating probabilities with frequencies ($p_i \approx f_i$):

$$\text{GI}(f) = \sum_{i=1}^{m} f_i (1 - f_i) = \sum_{i=1}^{m} f_i - \sum_{i=1}^{m} f_i^2 = 1 - \sum_{i=1}^{m} f_i^2$$

GI reaches its minimum (zero) when all cases in the node fall into a single target category; otherwise GI is positive.

### 5. ATTRIBUTE SELECTION

Variables related to activity and trip which first come to mind are used as input variables. In
addition, demographic information about volunteers and about land use at trip end may also have an influence on activity types. However, we do not currently have access to detailed land use information, so only demographic information is added as input variables along with activity- and trip-related variables. Variables from these three dimensions are interpreted in detail as follows.

The first dimension is related to activity features, including activity duration, periods of activity start and end, distance from activity location to home and work place, day characteristics of activity start and end. The second dimension is related to trip features and time cost of the trip before the current activity is used. The third dimension is related to demographic information about volunteers, including gender, age, annual household income, frequency of using auto and public transit. A question about driving licenses was also included in the questionnaire; however, since every volunteer had a license, this variable was finally eliminated from the independent variables for machine learning application. Specification of the selected attributes can be found in Table 1.

### Table 1 Attribute selected for analysis

<table>
<thead>
<tr>
<th>Dimensions</th>
<th>ID</th>
<th>Attribute Name</th>
<th>Numeric/Descriptive</th>
<th>Values of attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activity Dimension</td>
<td>1</td>
<td>Duration</td>
<td>Numeric</td>
<td>In seconds</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Period of activity start</td>
<td>Descriptive</td>
<td>7 categories&lt;sup&gt;1)&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Period of activity end</td>
<td>Descriptive</td>
<td>7 categories&lt;sup&gt;1)&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Distance from activity location to home</td>
<td>Numeric</td>
<td>In meters</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Distance from activity location to work place</td>
<td>Numeric</td>
<td>In meters</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>Weekday/weekend of activity starts</td>
<td>Descriptive</td>
<td>2 types: weekday and weekend</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>Weekday/weekend of activity ends</td>
<td>Descriptive</td>
<td>2 types: weekday and weekend</td>
</tr>
<tr>
<td>Trip Dimension</td>
<td>8</td>
<td>Time cost of trip before activity</td>
<td>Numeric</td>
<td>In seconds</td>
</tr>
<tr>
<td>Demographic Dimension</td>
<td>9</td>
<td>Gender</td>
<td>Descriptive</td>
<td>2 genders: male and female</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>Age</td>
<td>Descriptive</td>
<td>8 categories&lt;sup&gt;2)&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>Annual household income</td>
<td>Descriptive</td>
<td>10 categories&lt;sup&gt;3)&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>Occupation</td>
<td>Descriptive</td>
<td>8 categories&lt;sup&gt;4)&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>Auto usage frequency</td>
<td>Descriptive</td>
<td>5 categories&lt;sup&gt;5)&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>Public transit usage frequency</td>
<td>Descriptive</td>
<td>5 categories&lt;sup&gt;5)&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

Note: 1) 7 categories of time periods: NP-b-MP: non-peak before morning peak; MP: morning peak; NP-b-MP-NOON: non-peak between morning peak and noon; Noon: NP-b-NOON-EP: non-peak between noon and evening peak; EP: evening peak; NP-a-EP: non-peak after evening peak. Peak periods and the noon period are defined as follows: morning peak is from 7 am to 9 am on weekdays and weekends; evening peak is from 4 pm to 7 pm on weekdays and from 3 pm to 6 pm on weekends; and the noon period is from 12 noon to 1 pm on weekdays and weekends. Time periods for morning peak and evening peak are equivalent to those given in the report of current traffic characteristics in Chukyo Area in Japan.

2) 8 categories of age are as follows: 0–10 years old; 11–20 years old; 21–30 years old; 31–40 years old; 41–50 years old; 51–60 years old; 61–70 years old; 71–80 years old; more than 80 years old.

3) 10 categories of annual household income have the following intervals: [0,2), [2,3), [3,4), [4,5), [5,6), [6,7), [7,8), [8,10), [10,15), [15, +∞); unit is million Japanese yen.

4) 8 categories of occupation include: employee; self-employer or management; part time or freelance; government/school related; student; housewife; other; no occupation.

5) 5 categories of auto/public transit usage frequency cover: more than 5 days per week; 3–4 days per week; 1–2 days per week; 2–3 times per month; less than once per month.
6. RESULTS AND DISCUSSION

6.1 General Accuracy and Time Cost

We utilized Matlab 2014a as the platform for implementing empirical analysis of the four machine learning methods above and obtained accuracy and time elapse results for comparison.

Table 2 shows the general accuracy and time cost of each machine learning method utilized in this research. Numbers in bold italic show the highest accuracy or largest accuracy difference among methods, while underlined numbers have the lowest accuracy or smallest accuracy difference among methods. The accuracy difference, defined as the accuracy difference between the test set and the training set, is a reflection of the transferability of the method. A method with a larger accuracy difference can achieve more stable and transferrable parameters from the training set to other sets.

The classification tree obtains the highest general accuracy of 96.7% for training on the training set and 89.2% for predicting on the test set. One-against-rest SVM achieves the second-best general accuracy, followed by neural network and discriminant analysis. Discriminant analysis obtains the largest accuracy difference, whereby the accuracy of predicting is only 3.2% lower than that of the training; the other three methods have accuracy differences of 7.3%~7.5%.

Table 2 General accuracy and time cost of each method (unit of time: seconds)

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>Time (second)</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training data</td>
<td>Predicting data</td>
<td>Difference</td>
<td>Optimizing Parameter</td>
<td>Training</td>
<td>Predicting</td>
</tr>
<tr>
<td>One-against-rest SVM</td>
<td>91.8%</td>
<td>84.3%</td>
<td>-7.5%</td>
<td>21249.717</td>
<td>36.302</td>
<td>0.044</td>
</tr>
<tr>
<td>Neural Network</td>
<td>87.8%</td>
<td>80.5%</td>
<td>-7.3%</td>
<td>50.243</td>
<td>0.597</td>
<td>0.021</td>
</tr>
<tr>
<td>Discriminant Analysis</td>
<td>86.0%</td>
<td>82.8%</td>
<td>-3.2%</td>
<td>---</td>
<td>0.319</td>
<td>0.024</td>
</tr>
<tr>
<td>Classification Tree</td>
<td><strong>96.7%</strong></td>
<td><strong>89.2%</strong></td>
<td>-7.5%</td>
<td>---</td>
<td><strong>0.301</strong></td>
<td><strong>0.003</strong></td>
</tr>
</tbody>
</table>

As far as time cost is concerned, among these four methods, the classification tree uses the least time of 0.301 seconds for training on the training set and 0.003 seconds for predicting on the test set. The most time consuming method is one-against-rest SVM, which needs 36.302 seconds for training and 0.044 seconds for predicting. The time cost of the neural network and discriminant analysis are between the time costs of the classification tree and the one-against-rest SVM. One disadvantage of one-against-rest SVM and neural network is the process of obtaining the optimal parameter value, which needs much more time. In particular, the one-against-rest SVM takes nearly 6 hours for optimal parameter calculation.

From the point of view of general accuracy and time cost, it seems that the classification tree is the most appropriate of these four machine learning methods for identifying activity type from our data set.

6.2 Accuracy for Each Activity Type among Methods

Table 3 shows the accuracy of identifying each kind of activity type in the training set and test set as well as the accuracy difference when applying each method between these two sets. Numbers in bold italic and those underlined have the same meanings as in Table 2.
For “business” activity, all these methods can achieve a high accuracy of around 90%. Although the classification tree achieves the highest accuracy on the training set, it also produces the smallest accuracy difference between the test set and the training set. However, one-against-rest SVM achieves the highest accuracy on the test set and the accuracy difference is only –1.9%. Hence, it seems that one-against-rest SVM is the most capable of identifying “business” activity.

Regarding “recreation” activity, none of these four methods can achieve a satisfactory accuracy on both training set and test set. The classification tree achieves the highest accuracy on both the training set and the test set, but the accuracy difference is also the smallest one. In contrast, discriminant analysis achieves the largest accuracy difference but with the lowest accuracy on the training set and the test set. Thus, the classification tree appears to be the most appropriate for identifying “recreation” activity among the four methods.

For “home” activity, all these methods can achieve a high accuracy of over 93% for both the training set and the test set. The classification tree gives the highest accuracy and biggest accuracy difference, which means it is the most appropriate method for identifying “home” activity.

As far as “others” activity is concerned, the result is similar to that of “recreation” activity; none of these four methods can achieve a satisfactory accuracy on the training set and the test set. Compared with the other three, the classification tree can achieve a higher accuracy on both the training set and the test set. Although the classification tree has the smallest accuracy difference, it is still the most appropriate for identifying “others” activity.

Concerning “work-or-school” activity, the classification tree and discriminant analysis perform much better than the other two. Notably, the classification tree reaches 100% accuracy on both the training set and the test set.

For more detailed information, ratios of each activity type identified correctly or as other types by each method, which can be found in the confusion matrix, are shown in Table 4.

In summary, these four methods are all good at identifying the activity of business, home, and work-or-school, and all are incapable of identifying “recreation” and “others” with a satisfactory accuracy. From the viewpoint of the exact accuracy value and transferability of the method, it seems that the classification tree is the most appropriate method for identifying activity type among these four when being applied to our data set.

The optimally estimated structure of the classification tree is shown in Figure 3 (for simplicity, we show and explain only the first three levels). A brief explanation of the first three
levels is given as follows. The structure starts with a question of the current distance to the work place at the root (first level of parent node, node 1): if it is less than 55 meters, it directs to the left child node, node 2; else it directs to the right child node, node 3. At the second level, node 2 and node 3 become parent nodes and split again, based on the question of the distance to home. At node 2, if the distance is less than 1535 meters, then the activity type is business, else it is work/school. At node 3, if the distance is less than 54 meters, then the current activity type might be home, directing to the new left child node, node 6; else it directs to a new right child node, node 7 and needs further questions in the next level. In a recursive manner, the parent node splits into child nodes, and each child node is treated as parent node for the next split in the lower level based on further questions until all the activity types can be identified.

![Figure 3 Estimated structure of the classification tree with data from Nagoya](image)

### 6.3 Discussion

From the results above, it is clear that none of these four methods is capable of identifying the activities of recreation and others with high accuracy. Low accuracy may be explained from the following two perspectives.

1) Homogeneity among activities exists to some extent. Some subactivities in different categories may happen at a similar period of the day, costing a similar duration and locating near home or the work place, etc. For example, having dinner at a restaurant near home and doing exercise at a stadium not far away from home may be difficult to differentiate in the current methods. Table 4 demonstrates more detailed results in the form of a confusion matrix where the ratio of one kind of activity is identified as each kind of activity. It shows that recreation activity is easily misidentified as business activity, while “others” activity is easily misidentified as recreation or business activity. However, business activity is not easily misidentified as recreation and “others” activity, and recreation activity is not easily misidentified as “others” activity. This means that this kind of homogeneity is unidirectional. Some features of business activity and recreation activity may prevent the homogeneity turn to bidirectional. Consequently, based on the current attributes we used, recreation activity and “others” activity are not as feature-distinctive as the other three activity types.
<table>
<thead>
<tr>
<th>Methods</th>
<th>Training set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Truth/predicted</td>
<td>business</td>
</tr>
<tr>
<td>One-against-rest SVM</td>
<td>business</td>
<td>97.1</td>
</tr>
<tr>
<td></td>
<td>recreation</td>
<td>16.2</td>
</tr>
<tr>
<td></td>
<td>home</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>other</td>
<td>14.5</td>
</tr>
<tr>
<td></td>
<td>work or school</td>
<td>6.0</td>
</tr>
<tr>
<td>Neural Network</td>
<td>business</td>
<td>96.0</td>
</tr>
<tr>
<td></td>
<td>recreation</td>
<td>20.1</td>
</tr>
<tr>
<td></td>
<td>home</td>
<td>3.4</td>
</tr>
<tr>
<td></td>
<td>other</td>
<td>38.7</td>
</tr>
<tr>
<td></td>
<td>work or school</td>
<td>10.6</td>
</tr>
<tr>
<td>Discriminant Analysis</td>
<td>business</td>
<td>91.1</td>
</tr>
<tr>
<td></td>
<td>recreation</td>
<td>49.7</td>
</tr>
<tr>
<td></td>
<td>home</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>other</td>
<td>22.6</td>
</tr>
<tr>
<td></td>
<td>work or school</td>
<td>0.0</td>
</tr>
<tr>
<td>Classification Tree</td>
<td>business</td>
<td>97.3</td>
</tr>
<tr>
<td></td>
<td>recreation</td>
<td>5.0</td>
</tr>
<tr>
<td></td>
<td>home</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>other</td>
<td>11.3</td>
</tr>
<tr>
<td></td>
<td>work or school</td>
<td>0.0</td>
</tr>
</tbody>
</table>
2) No effective attributes are available in the existing data sets. As explained in the first respective, independent variables in the current data sets are not distinctive. Consequently, additional features may be necessary to make these two kinds of activity less homogeneous. Any two kinds of activities can be distinguished by some distinctive features. If not, they should be allocated into the same category. The only question is whether the distinctive features can be obtained or not. Additional features that can be obtained without too much difficulty may include interperson-related information; land use information; point of interest (POI) information, such as kindergarten, elementary school, railway/subway stations, etc.; visit frequency; social network information, among others. Interperson-related information can be obtained using several persons’ GPS trajectory and their social relationship. In our data sets, we only included information about volunteers’ home and work place; geographical information about other places may also be helpful in improving the accuracy. As regards social network information, some people like to update their status in real time. Thus, including this kind of information also has the potential to improve the accuracy.

Based on the discussion above, one possible solution is to refine an activity type into more specific ones. It provides the precondition for separating those that may have homogeneity. The next step is to add more features, especially distinctive features, in the machine learning algorithms in order to avoid the homogeneity.

7. CONCLUSIONS

In this paper, we applied four machine learning methods to identifying activity types from mobile phone GPS data we collected in Japan. Based on metrics of accuracy and time cost, it seems that the classification tree shows superiority over one-against-rest SVM, neural networks, and discriminant analysis. Except for business activity, the classification tree identifies four other types of activity with higher accuracy and uses less time. Moreover, it also achieves a satisfactory accuracy in identifying business activity. However, none of the four kinds of methods can handle the identification of recreation activity and other activity at a satisfactory level of accuracy. This may be because of the homogeneity of features in these two types of activities and a deficiency of effective features that can identify these two activities.

Future research can be furthered by including interperson-related information, land use information, POI information, visit frequency, social network information, etc., as input variables to check the performance of improving the accuracy. If each type of activity can be properly identified with a satisfactory accuracy by machine learning methods, GPS modules on mobile phones can be used as a reliable alternative to collect personal trip data with a lower cost during a longer period. Identified activities can be used as the input data source of activity-based models. It could be considered as another trend of future research.

ACKNOWLEDGEMENT

This study was supported by Grant-in-Aid for Scientific Research (No. 25630215 and 26220906) from the Ministry of Education, Culture, Sports, Science and Technology, Japan and the Japan Society for the Promotion of Science.

REFERENCES

of the 83rd Annual Meeting of the Transportation Research Board, January 2004, Washington D.C.


Sheela, K.G., Deepa, S. N. (2013). Review on methods to fix number of hidden neurons in
Stopher, P., Clifford, E., Zhang, J., FitzGerald, C. (2008b). Deducing Mode and Purpose from GPS Data. Working paper ITLS-WP-08-06. Institute of Transport and Logistic Studies, the Australian Key Center in Transport and Logistic Management, the University of Sydney