Short-Term Traffic Prediction Using Fuzzy C-Means and Cellular Automata in a Wide-Area Road Network

Hitoshi Kanoh, Member, IEEE, Takeshi Furukawa, Souichi Tsukahara, Kenta Hara, Hirotaka Nishi, and Hisashi Kurokawa

Abstract—In this paper, we propose a short-term prediction method for forecasting traffic in a time-series manner for up to one hour ahead for all roads in a wide-area road network. The results of our research enable traffic to be simulated for a wide-area road network based on actual traffic data by combining fuzzy clustering and cellular automata. On application to an actual road network with 3,405 links, the proposed technique was found to be superior to the nearest neighborhood method for traffic prediction at times of congestion outbreak and alleviation and heavy congestion.

I. INTRODUCTION

Recent advances in two-way communication between on-board car navigation equipment and key stations have made it possible to provide drivers with real-time traffic information on principal roads. We have been researching route-guidance systems that make use of this information for predicting traffic conditions [1]–[3]. In this paper, we propose a short-term prediction method for forecasting traffic in a time-series manner for up to one hour ahead for all roads in a wide-area road network.

Because time-series traffic data is highly nonlinear, it is not easy to make accurate predictions using a linear model like the autoregressive model, and other types of methods have come to be studied as a result. These include (a) nonlinear models such as the local linear regression model and neural networks [4]–[8]; (b) methods that use average values of historical data for each day of the week and time of day as predicted values [6], [9], [10]; (c) methods that apply the nearest neighborhood approach to real-time data and historical data [4]; and (d) methods that incorporate actual data in traffic simulators [11].

Our research is based on methods of type (d). We point out here that the target of past research was usually a specific road interval with attention paid to the time-series variation of traffic flow in that interval. In that research, little consideration was given to traffic conditions in neighboring roads making it extremely difficult to predict an outbreak or alleviation of traffic congestion. In contrast, the results of our research enable traffic to be simulated for a wide-area road network based on actual traffic data by combining fuzzy clustering and cellular automata (CA). This enables the temporal and spatial distribution of traffic to be reproduced with good accuracy on a computer, which, in turn, enables time-series traffic data to be computed for all roads in the target road network and the outbreak and alleviation of congestion to be predicted.

Though there are some applications of type (d) methods to a wide-area freeway network [11], we here target all drivable roads. This is real-world research using digital maps employed by actual car navigation equipment and actual traffic data provided by Japan’s Vehicle Information and Communication System (VICS). This VICS data is collected every five minutes by more than 20,000 traffic meters installed along principal roads throughout the country.

In the following, we begin with an overview of this research field. We then describe a method that uses fuzzy clustering to spatially interpolate traffic on roads having no traffic meters and a method for incorporating VICS data in a traffic simulator based on cellular automata. We next present experimental results comparing this technique with the nearest neighborhood method in an actual wide-area road network.

II. OVERVIEW OF RESEARCH FIELD

A. Use of traffic data in Japan

In Japan, traffic meters are installed at more than 20,000 locations along principal roads throughout the country. These meters measure the average travel time of cars and the number of cars passing through specific road intervals (links) in 5-minute intervals. The data so obtained is collected at a VICS center and provided to VICS subscribers in real time. A driver of a car can make use of VICS data in two ways. First, a driver can obtain data for several tens of kilometers ahead of the car’s current location. This data would be received via on-board car navigation equipment from optical or radio beacons set up along the road. Second, a driver can obtain a wide range of data independent of the car’s current location.
by two-way communication with a key station. In this case, the driver would be provided with time-series data or predicted traffic values obtained by processing that data, or optimal roads computed from such data. There are presently more than 10 million vehicles in Japan using VICS data.

B. Problems with route guidance

Some of the car navigation equipment currently available in Japan use VICS data for route guidance, but the searches performed for this purpose do not take predicted values into account. As a result, the subsequent occurrence of congestion on a road that was not congested when giving route guidance will result in a delay in the originally estimated arrival time. Furthermore, considering that the travel time for links not installed with meters is calculated by standard times established beforehand, the occurrence of congestion on such a link may generate a discrepancy between actual and computed travel time.

To solve these problems, a method for making spatial interpolations of VICS data must be developed and a more accurate method for predicting time-series data must be found. The former method must be able to estimate traffic on links not installed with meters from the traffic on links installed with meters in a wide-area road network. The latter method must be able to make short-term predictions of traffic—especially outbreaks and alleviations of congestion—from current VICS data for five minutes to one hour into the future.

C. Fuzzy c-means

In ordinary clustering, a certain target of classification must belong to one class. But in fuzzy clustering, the degree to which a certain target of classification belongs to a certain class is expressed by a membership function. Here, we consider that \( x_k \) \((k=1, \ldots, n)\) in \( p \) dimensional Euclid space are classified into \( c \) cluster centers \( v_i \)(\( i = 1, \ldots, c \)), where membership function \( u ik \) indicates the degree to which \( x_k \) belongs to \( v_i \). Fuzzy c-means (FCM) is a classification method that minimizes the following target function where \( \text{||x}_k - v_i \text{||} \) denotes a Euclid norm and \( m > 1 \) is a constant.

\[
f = \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ik})^m \|x_k - v_i\|^2 \tag{1}
\]

Subject to \( u_{ik} \in [0,1], \sum_{i=1}^{c} u_{ik} = 1 \) for all \( k \)

A solution can be obtained by applying iterative improvement to the following expressions.

\[
u_{ik} = \left[ \frac{1}{\sum_{j=1}^{c} \left( \frac{\|x_k - v_j\|^2}{\|x_k - v_i\|^2} \right)^{\frac{1}{m-1}}} \right]^{1-1} \tag{2}
\]

D. Traffic-flow simulation using cellular automata

We have implemented Nagel’s multi-speed model [13], [14] in cellular automata. In this model, a road is defined as a one-dimensional array of cells. The length of each cell is defined as the minimum distance between two cars. Each site (cell) of the array may be empty or may be occupied by a single car that has an integer velocity \( V \). This integer velocity represents the number of sites through which the vehicle will move on each step. The choice of maximum velocity is somewhat arbitrary, but may be justified by comparison between the model and real-world measurements. One iteration consists of the following steps, which are simultaneously applied to all of the vehicles:

1) Free vehicles are accelerated, vehicles that need to slow down because of other cars do so.
2) Each vehicle reduces its speed by one with probability 1/2.
3) Each vehicle advances \( V \) sites.

Our simulator [3], [12] uses map databases in the standard format which has been developed and established by the Navigation System Researchers’ Association in Japan. The maps used in actual car-navigation devices are of this type and areas where spontaneous traffic congestion occurs are indicated on these maps. This simulator is capable of simulating any road in Japan. The map includes indicators of such characteristics of roads as speed limits, distances, road classes, numbers of lanes, and the existence of traffic lights.

III. SPATIAL INTERPOLATION OF TRAFFIC DATA BY FUZZY C-MEANS

A. Method of interpolating link traffic data

The following discussion assumes a wide-area road network. Here, a link on which traffic meters have been installed is referred to as a “metered link” and a link having none is referred to as an “ordinary link.” We describe in particular a method for estimating traffic data of ordinary links from traffic data of metered links. Since no computation is, of course, required for traffic data of metered links, no such interpolation method is considered here for metered links. We treat the average speed of cars passing though a link as traffic data of that link.

This method applies fuzzy c-means to an actual road network based on the following guidelines and assumptions.

[Guideline 1] Assume a metered link to be cluster center \( v_i \) and an ordinary link to be classification target \( x_k \).

[Guideline 2] Use link dissimilarity (defined in the next
section) instead of the Euclid norm. It is known that an equation similar to that of (2) can be obtained using any norm instead of the Euclid norm of (1).

[Guideline 3] Letting $c=2$ from a practical perspective, classify each ordinary link into the closest metered link ($i=1$) and the second closest metered link ($i=2$). Although an evaluation experiment with $c=3$ was actually performed, no significant difference with respect to $c=2$ could be observed.

[Assumption 1] Focus only on (2) since actual meters have already been installed whereas (3) indicates an optimal arrangement of meters. In other words, perform approximate calculations assuming that (3) always holds.

[Assumption 2] Average speed $V(x_i)$ of cars passing through ordinary link $x_i$ can be calculated by the following equation using the average speed of cars $V(v_i)$ passing through metered link $v_i$ that ordinary link belongs to and membership function $u_{ik}$.

$$V(x_i) = \sum_{i=1}^{c} u_{ik} V(v_i)$$  \hspace{1cm} (4)

Here, denoting the dissimilarity between links $x_i$ and $v_i$ as $d_{ik}$, we obtain (5) below from (2) and (4) in conjunction with the above guidelines and assumptions.

$$V(x_i) = \sum_{i=1}^{2} \frac{d_{ik}d_{2k}}{d_{ik} + d_{2k}} V(v_i)$$

$$= \frac{d_{2k}V(v_i)}{d_{ik} + d_{2k}} + \frac{d_{ik}}{d_{ik} + d_{2k}} V(v_2)$$  \hspace{1cm} (5)

B. Calculation of link dissimilarity

It is empirically known that “if two links are close and of the same class of roads, there is generally some type of correlation between the traffic of those links.” We assume here that “close” means within about one kilometer, and that “class of roads” might be an expressway, national highway, principal local road, prefectural road, etc. Now, letting $L$ denote the fuzzy set of link pairs that are close to each other, $B$ the fuzzy set of link pairs of similar class of roads, and $S$ the fuzzy set of link pairs exhibiting similar traffic, the following can be written based on the above heuristics:

[heuristics] $S = L \cap B$

Accordingly, fuzzy set $D$ of link pairs exhibiting dissimilarity can be written as follows:

$$D = \overline{S} = L \cup B$$  \hspace{1cm} (6)

From this, we define the dissimilarity between links $x_i$ and $v_i$ in the following way:

$$d_{ik} = l_{ik} + b_{ik}$$  \hspace{1cm} (7)

Here, $l_{ik}$ is the length of the path that connects the midpoints of the two links and $b_{ik}$ is a constant set according to the class of roads of those two links.

C. Interpolation for remote links

As described in the previous section, dissimilarity between two links can be calculated based on heuristics that call for an inter-link distance within about one kilometer. Equation (5) is consequently not appropriate for links located more than one kilometer from each other. In this regard, it is generally difficult to interpolate traffic between such remote links with good accuracy, and we therefore add a previously determined standard speed to (5) to prevent deterioration of interpolation accuracy in such cases.

Specifically, we apply the following fuzzy inference rule where $V_{FCM}$ is average car speed calculated from (5) and $V_{STD}$ is the standard speed:

[Rule 1] if interpolation is “appropriate,” then adopt calculated value $V_{FCM}$$$

[Rule 2] if interpolation is “inappropriate,” then adopt standard speed $V_{STD}$

Now, if $r$ denotes the membership function of fuzzy set $R$ of link pairs exhibiting the “inappropriate” property, final interpolation value $V$ can be calculated by the expression shown below. Note that the fuzzy set exhibiting the “appropriate” property can be treated as the complement of $R$.

$$V = (1-r)V_{FCM} + rV_{STD}$$  \hspace{1cm} (8)

IV. SHORT-TERM TRAFFIC PREDICTION BY CELLULAR AUTOMATA

A. Calculation of traffic density

With the aim of reflecting VICS data in a traffic simulator, we first present a method for calculating link traffic density $K$ (number of cars per kilometer) from average car speed $V$ calculated by (8). In this study, we use a CA model that we previously developed as a traffic simulator [3], [12]. In this model, car speed takes on six levels of values, that is $V=0$ to 5, based on a multispeed model from Nagel et al [13], [14]. The following Greenshields equation, where $K_J$ denotes car jam density and $V_J$ free speed, is commonly used to describe the relationship between $V$ and $K$:

$$V = V_J \left(1 - \frac{K_J}{K} \right)$$  \hspace{1cm} (9)

We here use the following empirical formula as an amendment to (9).

$$K_{VICS} = \begin{cases} 
K_J \times 0.7 & \text{for} \quad V = 0 \\
K_J (1 - V / V_J - 0.1) & \text{for} \quad 0 < V < V_J \\
K_J \times 0.05 & \text{for} \quad V = V_J
\end{cases}$$  \hspace{1cm} (10)

We determined this equation by empirical means since the simulator is to be applied to a wide-area road network. The subscript in $K_{VICS}$ indicates traffic density calculated from VICS data. In this simulator, 1 cell = 5.5 m and $K_J=181$ (cars per kilometer). Free speed is the speed limit of the road in question.

B. Correction of traffic density by VICS data

To reduce prediction error that arises after running the traffic simulator, we correct traffic density obtained from (10)
by using VICS data from the next time interval. For metered links, we simply use traffic density $K_{\text{VICS}}$ as calculated from (10), but for ordinary links, we correct traffic density by the following formula using VICS data usage rate $\alpha \in [0, 1]$.

$$K = \alpha K_{\text{VICS}} + (1 - \alpha) K_{\text{CA}}$$

(11)

Here, $K$ denotes traffic density after correction and $K_{\text{CA}}$ traffic density obtained as a result of simulation. An optimal value for $\alpha$ is determined by experiment.

C. Prediction procedure

The procedure for predicting traffic is given below. The distribution of traffic flow on the simulator should approach that of actual traffic flow by repeating steps 6 to 8.

Step 1: Input map data and current VICS data for the road network targeted for prediction.
Step 2: Calculate average speed of cars on all metered links on the map directly from VICS data.
Step 3: Calculate average speed of cars on all ordinary links on the map using (8).
Step 4: Calculate traffic density of all links on the map using (10).
Step 5: Begin traffic flow simulation.
Step 6: Input VICS data from the next time interval.
Step 7: Calculate traffic density for metered links using (10) and that for ordinary links using (11).
Step 8: Jump to step 6 until $|K_{\text{VICS}} - K_{\text{CA}}|$ reaches a specific value or the current number of loops reaches a certain value.

V. EXPERIMENT

A. Experimental method

To evaluate the proposed technique, we forecast the average speed of cars on a link including crossroads in the area of Ogikubo Station in Tokyo, Japan on June 17 (Tuesday), 2003 for up to one hour ahead. We compared the results so obtained with those predicted by the nearest neighborhood method. Note here that the link targeted for prediction was a metered link, and while we did not, of course, use VICS data for that link in the experiment, we did use that data as true values for evaluating prediction results. The target map included 1,223 intersections and 3,405 links, and 22,385 cars were initially arranged on the road network.

As experimental conditions, $\alpha$ in (11) was set to 0.4 and the maximum loop number in step 8 was set to 7. One year’s worth of historical VICS data was used as search data for the nearest neighborhood method, and the number of neighborhoods was set to 4. Predictions were made for 12 5-minute intervals into the future, i.e., for up to one hour ahead. The prediction values obtained by the proposed technique were the average of five trials with respect to a stream of random numbers.

B. Experimental results

Figures 1 to 4 shows experimental results. In each of these figures, the horizontal axis represents prediction period (in 5-minute intervals into the future) and the vertical axis represents error rate (%) as well as actual speed (km/h) calculated directly from VICS data. From Figs. 1 to 4, the results shown correspond to times of congestion outbreak, congestion alleviation, heavy congestion, and no congestion, respectively.

The results shown in these figures reveal the following. In Fig. 1, car speed decreased 10 km/h from the 5-minute to the 10-minute time points. Here, error rate by the nearest neighborhood method was 43% whereas that by the proposed technique was lower at 17% indicating that the latter could predict an outbreak of congestion. Next, in Fig. 2, speed gradually increased corresponding to an alleviation of congestion, whereas in Fig. 3, speed ranged only between 4 and 7 km/h corresponding to heavy congestion. In both of these cases, the validity of the proposed technique is clear. Finally, in Fig. 4, speed is nearly constant (about 35 km/h) throughout the entire prediction period with no significant error exhibited by either the proposed technique or the nearest neighborhood method.

The superiority of the proposed technique can be explained as follows. Whereas the nearest neighborhood method uses only data of the link targeted for prediction, the proposed technique considers the inflow and outflow of cars to and from the link targeted for prediction based on the CA model. The experimental results presented above show that this feature becomes especially important when change in car speed is large such as in congestion outbreak and alleviation and heavy congestion.
VI. CONCLUSION

We proposed a short-term traffic prediction technique that incorporates actual VICS data in a traffic simulator targeting a wide-area road network. This technique can predict time-series traffic data for all roads in the targeted road network. On application to an actual road network with 3,405 links, the proposed technique was found to be superior to the nearest neighborhood method for traffic prediction at times of congestion outbreak and alleviation and heavy congestion. In future research, this technique must be further evaluated against a variety of maps, dates, days of the week, and time slots. We plan to combine this technique with a long-term prediction technique for use in route guidance.

REFERENCES