

A Short-Term Prediction Model for Forecasting Traffic Information Using Bayesian Network

Young Jung Yu¹ and Mi-Gyung Cho²

Division of Computer Engineering, Pusan University of Foreign Studies¹ and Department of Multimedia Engineering, Tongmyong University² (corresponding author:mgcho@tu.ac.kr)

Abstract

Currently the traffic information services of Telematics have had high qualities due to easy collection of the real-time traffic information through Intelligent Transport System (ITS). In this work, a short-term prediction model is proposed for forecasting the traffic information. The Bayesian network is used for each link with some casual nodes which can affect road situations in the future. In addition, a joint probability density function of the Bayesian network is obtained by assuming Gaussian Mixture Model (GMM) which utilizes training data set. To validate the precision of our model we conducted various experiments with two measures, one is an index as root mean square error (RMSE) and the other is travel time which takes three kinds of shortest paths for given paths. Our model provides less than 8 value of RMSE and the travel time of dynamic shortest path has more than 85% correlation with the real traffic data.

Index Terms – ITS, Telematics, Traffic Information Forecasting, Short-term prediction

1. Introduction

Recently due to ITS (Intelligent Transport System) that collects and transmits some real-time traffic information such as the traffic flow or an average speed of vehicles on each road link, the traffic information services of Telematics have shown a marked trend toward the high quality [1,8]. However, in Korea, most of car navigations only broadcast real-time information for lots of roads without any process. Although real-time information can help enough to understand the situation of road links, users may want to know the traffic situation where they visit in near future and obtain the travel time of their travel path with high accuracy. In this case, we need a high prediction model for future traffic information.

If we predict the traffic information in the future, we can develop various contents for traffic services and guarantee the high quality of the traffic services. Prior to the construction of ITS, the traffic information data accumulated for over one year was used to predict the traffic information instead of real-time data. Thus, the

forecasting accuracy was limited. Many researchers have been studied on the prediction model of the traffic flow in the future with real-time traffic data as ITS has been constructed in a wide area [1]-[7].

Usually, prediction models of traffic information can be classified into two main approaches, statistical model and analytical model [1]-[7]. The statistical model can be characterized as a data-driven method that generally uses a time series of historical and current traffic variables such as travel time, speeds, and volumes as input data. Numerous statistical methods on the accurate prediction of the travel time have been proposed, such as the Bayesian network model [2], ARIMA model [4], Probability Process [6] and neural network [7]. The main idea of traffic forecasting in the statistical models is based on the fact that traffic behaviors possess both partially deterministic and partially chaotic properties. Forecasting results can be obtained by reconstructing the deterministic traffic motion and predicting the random behaviors caused by unanticipated factors. On the other hand, the analytical model predicts travel times by using microscopic or macroscopic traffic simulators [9].

In this paper, we proposed a short-term traffic prediction model based on the Bayesian Network. A short-term prediction means to forecast the traffic status in the near future, such as for 60 minutes. The Bayesian Network model is to forecast the future traffic information at given links from the historical data of themselves and their neighbor links. We use real-time traffic data, especially an average speed of the vehicle on a road as input data and will also predict an average speed of the vehicle on each road in the near future. To construct the Bayesian Network we supposed a GMM with three components as a distribution of random variables and applied the EM algorithm to extract parameters of the joint probability density function of GMM.

There are two differences between our model and the Bayesian Network model first developed by Sun et al. [2]. First, to construct the Bayesian Network we consider both flows of upstream and downstream link and its own historical data. Second, we apply the current flows of each link to calculate the final prediction result in case of incidents and accidents because the accuracy of the

Bayesian Model may be decreased in the case of the incidents and accidents.

The remainder of the paper is organized as follows. We will explain how to design and construct the Bayesian network for the road link in Sections 2 and describe our prediction system to get the final result in Section 3. And the experiment results are given in Section 4. Finally, we make our conclusions in Section 5.

2. Construction of Bayesian network

2.1. Design of Bayesian network

The Bayesian Network is known as a cause network because it can represent the causal relationship between random variables graphically. We have an intuition that the flow of each road link in the near future is highly related to the current and latest flows of the upstream and downstream links. The Bayesian network is represented as a directed acyclic model for representing conditional independence between a set of random variables. In the Bayesian network, a node x_i is conditional independent of the nodes that are not a descendant of x_i . Therefore, for the Bayesian network consisting of n nodes (x_1, x_2, \dots, x_n) , we have the representation for the joint probability distribution

$$P(x_1, x_2, \dots, x_n) = \prod_{i=1}^n P(x_i | p(x_i)) \quad (1)$$

where $P(x_i | p(x_i))$ is the conditional distribution associated with node i and $p(x_i)$ is the set of indices labeling the parents of node i .

To design the Bayesian network for the prediction of traffic information, we need to determine causal nodes for each link in the road network. In our research, three elements are considered as causal factors. First fact is the latest traffic information of the link, second is the latest traffic information of upstream links and last fact is the latest traffic information of downstream links. The latest traffic information is used for past 15 minutes.

Fig. 1 (a) shows an example of a road network. Each circle means a crossroad and an arrow of a link represents a direction of the lanes. The symbol $BC(t)$ denotes the traffic information of the link BC at a time t . Upstream links of the link BC are CD , CG , and CH and downstream links are AB , EB , and FB . Fig. 1 (b) shows an example of the Bayesian network between the link BC and its upstream and downstream links.

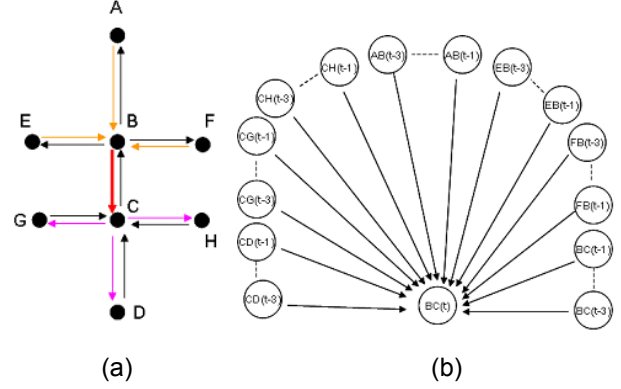


Fig. 1. (a) An example of a road network (b) The Bayesian network between the link BC and its neighbor links

The traffic information of each link is updated every 5 minutes. In Fig. 1 (b), $BC(t-1)$ and $BC(t-2)$ mean the traffic information of past 5 minutes and past 10 minute, respectively. The joint probability distribution of the Bayesian network in Fig. 1 (b) is given as follows.

$$\begin{aligned} & P(CD(t-j), CG(t-j), CH(t-j), AB(t-j), EB(t-j), FB(t-j), BC(t-j+1)) \\ &= \prod_{j=1}^3 P(BC(t) | CD(t-j)) \times \prod_{j=1}^3 P(BC(t) | CG(t-j)) \times \prod_{j=1}^3 P(BC(t) | CH(t-j)) \\ & \times \prod_{j=1}^3 P(BC(t) | AB(t-j)) \times \prod_{j=1}^3 P(BC(t) | EB(t-j)) \times \prod_{j=1}^3 P(BC(t) | FB(t-j)) \\ & \times \prod_{j=1}^3 P(BC(t) | BC(t-j+1)) \end{aligned} \quad (2)$$

2.2. Extracting parameters of GMM

In our research, we used a road network data and its traffic information of the area of Gang-Nam-Ku in Seoul. The number of upstream and downstream links used as causal nodes was about 4 or 5. The latest traffic information of upstream and downstream links for the past 15 minutes was used as causal nodes to construct the Bayesian Network. Consequently, the number of causal nodes in the Bayesian network is between 20 and 24. It is very difficult to find the joint probability density function for high dimensional data. Therefore, we assumed Gaussian Mixture Model (GMM) with three normal distribution functions to approximate the joint probability distribution in the Bayesian Network.

Fig. 4 shows a distribution of speed of vehicles between a target link and one of the downstream links. We found three groups from a distribution. Also, we obtained the same trends for the rest of other distributions. This is why we adopt the GMM with three components.

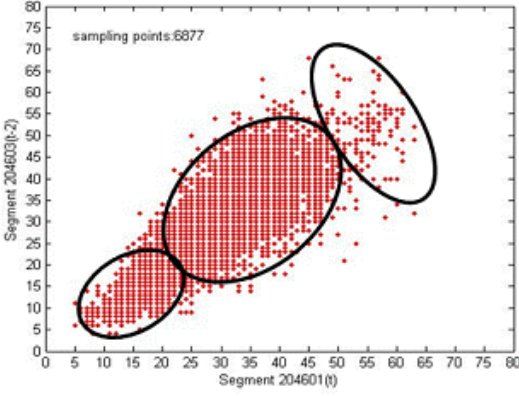


Fig. 2. A distribution of speed data between a target link (# 204601) and one of downstream links (# 204603)

The GMM with M components can be described as

$$p(x \setminus \Theta) = \sum_{l=1}^M \alpha_l p_l(x \setminus \theta_l) \quad (3)$$

where $\Theta = (\alpha_1, \dots, \alpha_M, \theta_1, \dots, \theta_M)$, $\sum_{l=1}^M \alpha_l = 1$ and $p_l(x \setminus \theta_l)$ is a Gaussian probability density function parameterized by $\theta_l = (\mu_l, \Sigma_l)$, $l = 1, \dots, M$. [2].

To estimate the GMM parameters, we used Expectation Maximization (EM) algorithm for training data. EM algorithm is an iterative method to carry out Maximum Likelihood Estimation (MLE). The below iterative equations to obtain the estimation of new parameters (i) in terms of the old parameters ($i-1$) are given as

$$\begin{aligned} \alpha_l^{(i)} &= \frac{1}{N} \sum_{k=1}^N p(l \mid x_k, \Theta^{(i-1)}) \\ \mu_l^{(i)} &= \frac{\sum_{k=1}^N x_k p(l \mid x_k, \Theta^{(i-1)})}{\sum_{k=1}^N p(l \mid x_k, \Theta^{(i-1)})} \\ \Sigma_l^{(i)} &= \frac{\sum_{k=1}^N p(l \mid x_k, \Theta^{(i-1)}) (x_k - \mu_l^{(i)}) (x_k - \mu_l^{(i)})^T}{\sum_{k=1}^N p(l \mid x_k, \Theta^{(i-1)})} \end{aligned} \quad (4)$$

where, N is the size of data set and $l = 1, \dots, M$.

2.3. Prediction by Bayesian network

In the Bayesian Network, a reasoning value of the target node is the conditional expectation of causal nodes. Equation 5 represents the conditional probability with GMM as the joint probability density function.

$$\begin{aligned} P(Y \mid X) &= \frac{P(Y, X)}{P(X)} \\ &= \frac{\sum_{i=1}^M \alpha_i G(X; \mu_{iX}, \Sigma_{iXX}) G(Y; \mu_{iY|X}, \Sigma_{iY|X})}{\sum_{j=1}^M G(X; \mu_{jX}, \Sigma_{jXX})} \\ &= \sum_{i=1}^M \beta_i G(Y; \mu_{iY|X}, \Sigma_{iY|X}) \end{aligned} \quad (5)$$

where,

$$\begin{aligned} \beta_i &= \frac{\alpha_i G(X; \mu_{iX}, \Sigma_{iXX})}{\sum_{j=1}^M \alpha_j G(X; \mu_{jX}, \Sigma_{jXX})} \\ \mu_{iY|X} &= \mu_{iY} - \Sigma_{iYX} \Sigma_{iXX}^{-1} (\mu_{iX} - X) \\ \Sigma_{iY|X} &= \Sigma_{iYY} - \Sigma_{iYX} \Sigma_{iXX}^{-1} \Sigma_{iXY} \end{aligned}$$

In equation 5, μ_{iY} is the average of the variable Y in i^{th} group and Σ_{iXX} is the covariance between the variable X and the same variable X in i^{th} group. Σ_{iXY} is the covariance between the variable X and Y in i^{th} group. The equation of a target node in the Bayesian network is as follows:

$$\begin{aligned} \hat{Y} &= E[Y \mid X] \\ &= \int Y P(Y \mid X) dY \\ &= \sum_{i=1}^M \beta_i \int Y G(Y; \mu_{iY|X}, \Sigma_{iY|X}) dY \\ &= \sum_{i=1}^M \beta_i \mu_{iY|X} \end{aligned} \quad (6)$$

The prediction result can be determined in a few seconds because the value β_i and $\mu_{iY|X}$ was previously obtained from Equation 4.

3. Prediction system using Bayesian network

As incidents and accidents happen, the prediction by the Bayesian Network becomes worse. This is caused by the fact that the Bayesian Network is a probability model using training data set. If we have lots of traffic data in case of incidents and accidents, we can estimate the GMM parameters in these cases. However, as mentioned by Sun [2], it is usually difficult to collect these data.

To overcome the weak point of the Bayesian Network for traffic information, we utilize the current information of each link to calculate the final prediction result in case of abnormal traffic flows such as the incidents and accidents. We will explain how to determine whether current traffic flow is normal or abnormal and how to get the final prediction result.

3.1. Determination of current traffic situation

As the traffic flow of a specific link has a similar pattern everyday in the case of which the road condition is not abnormal we use the result by the Bayesian Network as the final forecasting value, while we reflect the current traffic flow in case of traffic jams. Therefore, we mixed both models by applying an adapted weight to the two models. The weight is computed by Equation 7 and its graph is given in Fig. 3.

$$\alpha = \frac{\log \sqrt{\sigma_t}}{\log \sqrt{\sigma_t} + \log |PS_t - RS_t|} \quad (7)$$

where PS_t indicates the value predicted by the Bayesian Network and RS_t is the vehicle speed of the real time for past 5 minutes, respectively. The standard error (σ_t) can obtain from an accumulated database corresponding to the time sequence (t) of the each link. The value α means the ratio reflecting the forecasting value by the Bayesian Network. As shown in Fig. 3, if the difference between PS_t and RS_t is very small, the weight value α approaches one. In other words, the current traffic flow is normal, so that the final prediction result depends on PS_t as shown in Fig. 4

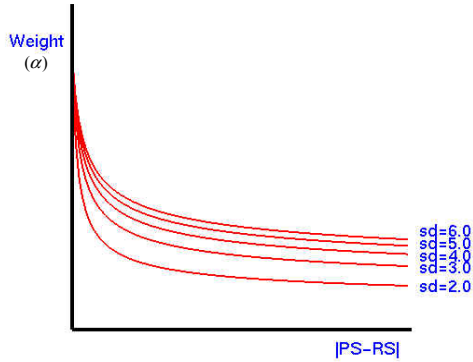


Fig. 3. Weight for reflecting the current traffic situation

3.2. Overview of our prediction system

Fig. 4 shows the overview of our prediction model to get the final result. Our system has the preprocessing step for the real-time traffic information updated every 5 minutes to estimate the traffic information of missing links and to calculate the weight value α .

Sometimes, the traffic information of some links was missing due to weather conditions and network problems of transmitted devices. We calculated each speed of missing links using the graph topology nearby these links. In order

to compute each speed of the missing links we found the incoming and outgoing edges of each node adjacent to all missing links. In addition, the speed of each missing link is replaced by the average speed of its incoming and outgoing edges.

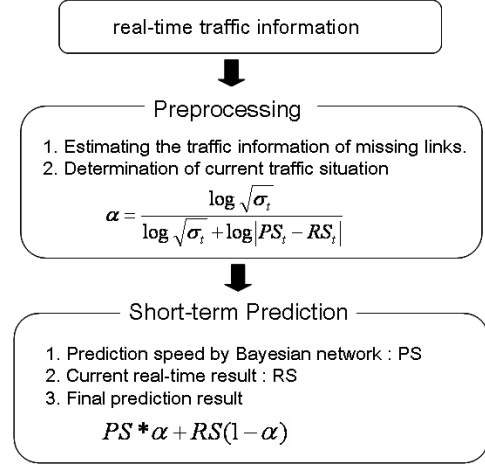


Fig. 4. Overview of our prediction system

4. Experimental results

4.1. Experimental method

To construct the Bayesian Networks and to verify the accuracy of our model, we used real-time traffic information for road links of an area of Kang-Nam-Ku in Seoul city for one month, from September 1 to September 30, 2004. The traffic data is composed of the discrete vehicle speeds of the road links updated every 5 minutes and their units are kilometer per hour (km/h). We construct the Bayesian Network for all road links of Kang-Nam-Ku. To do experiment we divided the traffic information into two parts: the training data set and the test data set. The former is extracted from Sep. 1 through Sep. 20 and is used to estimate the parameters of the GMM. The latter is from Sep. 21 through Sep. 30 and used for validation of forecasting performance. We employed common two measures such as root mean square error (RMSE) and travel time for the validation. We calculated RMSE for 2880 sampling points (10days*24hour *12unit) of all road links.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (8)$$

The travel time is obtained for a given path. We defined three kinds of shortest paths to minimize travel times: static shortest path which is the shortest path by utilizing the current traffic data obtained from the current time for a

given direction, accumulated shortest path which is the shortest path by using the accumulated averaged data obtained from the specific starting time for a given direction, and dynamic shortest path which is the shortest path applying our proposed model for a given direction. Note that our proposed model and the model using the accumulated data can have different shortest path time as the starting departure time varies, while static shortest model provides always the same shortest path time. For example, when we want to depart from 30 min. later after the current time, our model and the dynamic shortest path model can consider the delay time. And then we compared those with the known real shortest path for given starting time.

4.2. Prediction Error

Fig. 5 shows a comparison between a real speed and a prediction speed of our model after 20 minutes for a link from a certain time. The prediction speed (y-axis) is obtained from 5 minute streaming data for 24 hours (x-axis). As shown in Fig. 5, the prediction speed has a good agreement with real traffic speed for 24 hours, while our model is not a good matched with real speed in the case of which the vehicle speed changes sharply at time $t+1$ comparing time t . In general, unless vehicle speed of the same road link has sudden increase or decrease after 5 minutes we conclude that our model well follows the prediction speed at a normal load condition.

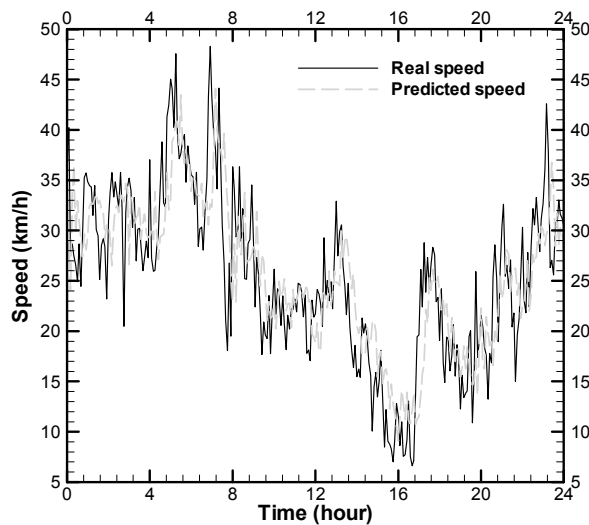


Fig. 5. Forecasting vehicle speed of a specific link

Table 1 gives the average of RMSE value for selected fifteen links randomly. As it is far from the current time, the root mean square error is a little increased. But although we forecast after 60 minutes, RMSE values for selected links keep below 8. As shown in Table 1, the prediction

error is not high and our approach gives reasonable forecasting results.

Table 1. RAME value of fifteen links

Link \ Time	5 min.	10 min.	15 min.	20 min.	25 min.	30 min.	60 min.
202322	4.48	5.24	5.67	5.99	6.1	6.29	6.71
203672	4.1	5	5.56	5.84	6.11	6.39	6.96
202420	5.52	6.45	7.01	7.36	7.57	7.8	8.32
205774	4.82	5.76	6.33	6.58	6.72	6.86	7.12
208622	3.6	4.47	4.99	5.42	5.63	5.81	6.51
204638	3.46	4.14	4.62	4.77	4.86	5.07	5.98
207899	4.33	5.4	6.26	6.84	7.15	7.42	8.2
204653	5.16	6.04	6.58	6.93	7.18	7.46	8.67
204623	4.11	4.8	5.16	5.4	5.56	5.73	6.45
205675	4.49	5.72	6.27	7.07	7.56	7.77	8.85
208209	4.8	6.04	6.92	7.49	7.87	8.2	11.31
208448	4.21	5.24	6.03	6.68	7.35	8.03	12.02
202764	4.59	5.22	5.67	5.97	6.22	6.42	7.44
204571	4.75	5.45	5.88	6.27	6.55	6.68	7.85
205650	3.63	4.22	4.5	4.82	5.05	5.3	6.9
Average	4.4	5.27	5.83	6.22	6.49	6.74	7.92

4.3. Travel Time

We compare the travel time of four kinds of shortest path for given ten different paths. In Table 2, RSP, ASP, SSP and DSP stand for real shortest path, accumulated shortest path, static shortest path and dynamic shortest path, respectively. The value in parenthesis means the difference between the real shortest path and other three kinds of shortest paths. The experiment shows that the travel time of the dynamic shortest path is the nearest to one of the real shortest path. This proves the accuracy of our model. In addition, in the case of delayed departure, the result is better and the travel time of dynamic shortest path has more than 85% correlation with the real traffic data.

Table 2. Comparison the travel time of a static, an accumulated and a dynamic shortest path

Path \ Method	RSP	ASP	SSP	DSP
245↔2443	1680	1968 (288)	1752 (72)	1728 (48)
191↔865	2251	2010 (241)	2348(97)	2324 (73)
210↔2286	2227	2945 (719)	2408 (181)	2289 (62)
224↔682	1215	1783 (568)	1098 (117)	1158 (57)
191↔797	2267	1949 (318)	2397 (130)	2321 (54)
913↔200	1212	1355 (143)	1162 (50)	1187 (25)
314↔874	1231	1515 (284)	1366 (135)	1353 (122)
2443↔224	3776	4569 (793)	3195 (581)	3232 (544)
387↔5960	1486	1430 (56)	1456 (30)	1537 (51)
192↔2443	2089	2238 (149)	2031 (58)	2011 (78)

5. Conclusion

In this paper, we proposed a short-term prediction model for forecasting traffic information in the future using the Bayesian Network. We assumed GMM as a joint probability density function of the Bayesian Network and used EM algorithm to extract parameters of the GMM from

sampling data set. For experiment we used road links and their real-time traffic data of an area of Kang-Nam-Ku in Seoul city for one month. To evaluate the accuracy of our model, we conducted various experiments with two kinds of measures: root mean square error and a travel time of dynamic shortest path. As a result, the average of root mean square error of all links is a little increased as the prediction time is going to the future. In addition, although the prediction is conducted for the future 60 min. RMSE maintains less than 8. The predicted travel time obtained from the dynamic shortest path comes close to the travel time of real shortest path than the other two methods: accumulate shortest path and static shortest path.

Acknowledgment

This work was supported by the SKTU Institute of Next Generation Wireless Communications funded by SK telecom (SKTU-07-006).

References

- [1] Hironobu Kitaoka, Takahiro Shiga, Hiroko Mori etc., "Development of a travel Time Prediction Method for the TOYATA G-BOOK Telematics Service," R&D Review of Toyota CRDL Vol. 41 No. 4, 2007.
- [2] Shiliang Sun, Changshui Zhang, Guoqiang Yu, "A Bayesian Network Approach to Traffic Flow," IEEE Transaction on Intelligent Transportation Systems, Vol. 7, No. 1, 2006.
- [3] M.G.Cho, Y.Yu, and S. Kim, "The System for Predicting the Traffic Flow with the Real-Time Traffic Information," Springer-Verlag, Lecture Note in Computer Science, Vol. 3980, 2006
- [4] E. Frascini and K. Ashausen, Day on Day Dependencies in Travel Time:First Result Using ARIMA Modeling:ETH, IVT institute for Transport, Feb. 2001.
- [5] Road E. Turochy, "Enhancing Short-Term Traffic Forecasting With Traffic Condition Information," ASCE Journal of Transportation Engineering. 2005
- [6] G. Q. Yu, J. M. Hu., C. S. Zhang, etc. "Short-term traffic flow forecasting based on Markov chain model," Proc. IEEE Intelligent Vehicles Symp., Columbus, OH, 2003.
- [7] J.W.C. van Lint, S.P. Hoogendoorn, and H.J. van Zuylen, "Robust and adaptable travel time prediction with neural networks," Proc. 6th Annual transport, 2000.
- [8] Joonwan Kim, Trends of Services and Technology of Telematics, , IITA IT Korea, 2004.
- [9] J. Maroto, E. Delso and Jose Ma, "Real-time Traffic Simulation with a Microscopic model," IEEE Transaction on transportation system, Vol. 7, No. 4, 2006.