베이지안 네트워크를 이용한 단기 교통정보 예측 모델

유영중* · 조미경**

A Short-Term Traffic Information Prediction Model Using Bayesian Network

Yu Young Jung* · Mi-Gyung Cho**

본 논문은 2007도 SK Telecom 재원으로 설립된 동명대학교 SKTU 차세대통신기술연구소 학술연구비 지원에 의하여 이루어진 것임 (SKTU-07-OO6)

요 약

최근의 텔레매틱스 교통정보제공서비스는 지능형교통시스템의 구축을 통한 실시간 교통정보 수집이 가능해짐에 따라 다양해지고 있다. 본 논문에서는 고품질의 다양한 교통정보제공을 위해 필요한 미래시간에 대한 단기 교통정보 예측 모델을 제안하고 개발하였다. 단기 예측 모델은 현재로부터 가까운 미래의 교통 상황을 예측하기 위한 교통 모델로 본 연구에서 제안한 예측 모델은 각 도로에 대하여 5분 이후부터 1시간 이전까지의 미래 시간에 대한 차량 평균 속도를 예측 결과로 준다. 본 연구에서 제안한 예측 모델은 베이지안 네트워크에 기반을 두고 있으며 각 도로의 미래 시간 교통상황에 영향을 줄 수 있는 요인들을 분석하여 베이지안 네트워크의 원인노드로 설정하였다. 설계된 베이지안 네트워크에 대하여 실시간 교통 정보 데이터를 이용하여 가우시안 혼합 분포를 가정한 베이지안 네트워크의 결합 확률 밀도 함수를 EM(Expectation Maximization) 알고리즘으로 구하여 미래 시간의 교통정보를 예측하였다. 예측 모델의 정확도 검증을 위해 실시간 교통데이터로 다양한 실험을 수행하였다. 실험결과 제안된 모델은 현재 시간으로부터 10분 이후, 30분 이후, 60분 이후 예측 오차로 각각 4.5, 4.8, 5.2의 RMSE(Root Mean Square Error) 값을 주었다.

ABSTRACT

Currently Telematics traffic information services have been various because we can collect real-time traffic information through Intelligent Transport System. In this paper, we proposed and implemented a short-term traffic information prediction model for giving to guarantee the traffic information with high quality in the near future. A Short-term prediction model is for forecasting traffic flows of each segment in the near future. Our prediction model gives an average speed on the each segment from 5 minutes later to 60 minutes later. We designed a Bayesian network for each segment with some casual nodes which makes an impact to the road situation in the future and found out its joint probability density function on the supposition of GMM(Gaussian Mixture Model) using EM(Expectation Maximization) algorithm with training real-time traffic data. To validate the precision of our prediction model we had conducted various experiments with real-time traffic data and computed RMSE(Root Mean Square Error) between a real speed and its prediction speed. As the result, our model gave 4.5, 4.8, 5.2 as an average value of RMSE about 10, 30, 60 minutes later, respectively.

키워드

지능형교통정보 시스템, 텔레매틱스, 교통정보 예측, 베이지안 네트워크, 예측모델

* 부산외국어대학교 컴퓨터공학과

** 동명대학교 멀티미디어공학과

접수일자 2009. 01. 14

Ⅰ. 서 론

최근의 국내외 교통정보제공서비스는 지능형교통시스템(ITS: Intelligent Transport System)의 구축으로 실시간 교통정보 수집이 가능해짐에 따라 고품질화 되는 추세이다. 우리나라의 경우 1990년대 중반부터 시작된 국가 ITS 구축사업은 2005년 국가 ITS 기본 계획 1단계가 마무리되면서 수도권, 광역시, 고속 도로, 주요 도로 및 여러 지자체 등을 중심으로 실시간 교통 정보 수집을 위한 시스템이 이미 구축되었고 수도권 등 일부 지역들에 한하여 실시간 교통 정보 서비스를 제공하고 있다[1-2].

우리보다 앞서 ITS를 구축한 선진국에서는 고품질의 다양한 교통정보서비스 제공을 위해 실시간교통 데이터를 가공하는 기술, 특별히 미래시간에 대한 교통 정보 예측 기술에 대한 연구가 활발히 진행되고 있다[3-10]. 이는 현재 시간으로부터 짧게는 몇 분후부터 길게는 몇 시간 후 교통 상황을 예측하는 것이다. 일본의 경우 도요타의 텔레매틱스 서비스 브랜드인 G-BOOK은 국가 통합 교통정보 시스템의 실시간 교통 정보와 연계하여 대도시를 중심으로 미래시간에 대한 통행량, 동적 최단 경로, 혼잡 예상, 통행 소요 시간 예측 등 다양한 예측 교통정보 서비스를 제공하고 있다[4].

우리나라의 경우 몇 개의 텔레매틱스 및 ITS 서비스에서 통행량이나 차량 평균 속도 등 각 도로의 실시간 교통 정보를 단순히 전달해 제공해 주거나 몇 명의 연구자들에 의해 통행시간과 여행시간을 예측하기 위한 몇 가지 방법들이 제안되고 있지만 제안된 방법들은 실시간 교통 정보가 아닌 누적된 교통 정보를 사용하고 있다[11-13]. 다양한 교통 콘텐츠의 개발과 함께 고정밀의 동적 최단 경로나 통행 소요 시간 서비스 등을 제공하기 위해서는 몇 분 후 혹은 몇 시간 이후의 교통 정보에 대한 예측 기술이 필수적이다.

본 논문에서는 현재로부터 비교적 가까운 미래시간의 교통정보를 예측하기 위해 베이지안 네트워크를 이용한 단기 교통정보예측 모델을 제안하고 개발하였다. 베이지안 네트워크의 원인 노드로 각 링크의 상·하류 링크들의 최근 교통정보와 해당 링크의 최근 교통정보를 사용하였다. 도로 네트워크에 대한 베이지안 네트워크를 설계한 후 결합 확률 밀도 함수를 추정하기 위해 가우시안 혼합 분포를 가정하고 EM 알고리즘을 이용하여 가우시안 혼합 분포의 파라메타 값들을 계산하였다. 제안한 모델의 정확도를 검정하기위해 RMSE을 평가도구로 하여 예측 오차를 구하기 위한 다양한 실험을 수행 하였다.

본 논문의 구성은 다음과 같다. 2장에서는 교통정보 예측 문제에 대한 정의와 이제까지 제안된 기존의 예측 모델들에 대해 소개할 것이다. 3장에서는 베이지안 네트워크에 대한 소개와 베이지안 네트워크설계 방법에 대해 설명할 것이다. 또한 실시간 교통 데이터를 이용하여 베이지안 네트워크의 결합 확률 밀도 함수를 구하는 방법과 예측 방법에 대해 설명할 것이다. 4장에서는 현재 교통상황을 분석하는 방법에 대해 설명할 것이다. 그리고 5장에서는 예측 모델의 정확도 검증을 위한 실험 방법과 결과에 대해 언급하고, 6장에서 결론을 맺는다.

Ⅱ. 교통정보 예측 문제와 기존 모델들

교통 데이터는 이력 데이터, 현재 데이터, 예측 데이터로 구분할 수 있다. 이력 데이터는 몇 년 혹은 몇 달 동안 누적된 교통 정보 데이터로부터 유용한 정보들을 추출, 가공한 것을 말하며 현재 데이터는 지능형교통시스템으로부터 실시간으로 들어오는 현재 교통 정보를, 예측 데이터는 교통정보 예측 시스템에 의해 계산된 정보를 의미한다. 현재로부터 가까운 미래에 대한 교통정보 예측을 단기 예측, 상대적으로 먼 미래에 대한 교통 정보 예측을 장기 예측이라고 한다.

최근까지 제안된 교통정보 예측 모델들을 살펴보면 확률과정 모형, ARIMA (Autoregressive Integrated Moving Average) 모형, 칼만 필터링(Kalman Filtering) 모형, 인공신경망(Artificial Neural Network) 모형, 베이지안 네트워크(Bayesian Network) 모델 등이 있다[3-15]. 확률과정을 이용한 예측 모형은 마코프 과정(Markov Process)을 이용한 것으로 현재 시점에서의 상태는 과거 시점 t-1에서의 상태에 의해 영향을 받는다는 것에 착안한 모델이다[6]. ARIMA를 이용한 예측 모형은 고전적인 시계열 모형중 하나인 ARIMA를 이용한 것으로 시간의 경과에 따라 일정한 특징을 가지는 교통정보 데이터에 시계열 모형인 ARIMA 모형을 적용한 것이다[7]. 인공신경망 모델은 시냅스의 결합으로 네트워크를 형성

한 인공 뉴런이 학습을 통해 시냅스의 결합 세기를 변화시킴으로 최적의 값을 찾아가는 인공 신경망을 교통 예측에 적용한 방법이다[8]. 칼만 필터링 모형은 오차에 의해 간섭받는 선형 동적 시스템에서 상태 벡터의 에러를 최소화한 최적의 추정치를 구하기 위한 순환적인 방정식을 사용하는 것으로 단계를 반복해 갈수록 오차를 줄여주는 방향으로 예측 값을 구해주기 때문에 다른 모델에 비해 상대적으로 좋은 결과를 준다고 보고되어 있지만 방정식에 사용해야 되는 행렬의 크기에 따라 계산양이 많아지므로 도로 네트워크의 크기가 커질 경우 실시간 예측에 어려움이 있다[10].

기존의 예측 모형을 가지고 실험한 결과들을 살펴보면 고속도로를 대상으로 예측한 결과 단기 예측에는 칼만 필터링 모형이 장기 예측에는 인공 신경망 모형이 우수한 결과를 보인다고 보고하였다[13]. 또 다른 연구에서는 서울시 28개축 주요 간선도로 축을 대상으로 하여 예측 모형을 적용한 결과 단기예측에는 칼만 필터링 모형이 장기예측에는 확률과정 모형이 상대적으로 우수하다고 보고하였다[11]. 또한 예측 시간대가 현재 시각에서 매우 멀어질수록, 즉 장기 예측의 경우 각종 예측 모델을 이용하여 예측된 결과의 예측 오차가 증가하는 반면 누적된 이력 자료를 이용하여 예측한 결과의 오차는 일정하게 나타난다고 보고된다.

최근에 제안된 베이지안 네트워크 모형은 앞서 소개한 다른 예측 모델들과는 달리 도로 네트워크의 위상정보를 명시적으로 적용하는 모형이다[3]. 베이지안 네트워크의 원인 노드를 미래의 교통 상황에 영향을 미칠 수 있는 인근의 상・하류 링크들로 설계함으로 특정 노드에 대한 교통정보 예측이 인근 링크들의 최근 교통정보에 의해 결정되도록 하는 모델이다. 본 연구에서는 베이지안 네트워크를 기반으로 하는 단기 교통정보 예측 모델을 설계하고 개발하였다. 베이지안 네트워크를 이용한 예측 모델은 2006년 Sun 등에 의해 처음 제안되었다[3].

본 연구팀에서는 Sun 등이 제안한 예측 모델의 단점을 보완하기 위해 두 가지 사항을 개선하였다[5]. 첫째는 Sun 등은 베이지안 네트워크 설계에서 특징 링크에 대한 원인노드로 교통흐름에서 특정 링크의 상류 링크만을 사용하였는데 본 연구에서 링크의 상・하류 링크들을 모두 원인 노드들로 설계하였다. 두 번째는 Sun 등이 제안한 베이지안 네트워크 모델의 가장 큰 단점은 돌발 상황이 발생했을 경우 예측 오차가 커질 수 있다는 것이다. 제시한 모델에서는 돌발 상황이 발생했을 경우 현재의 실시간 교통 정보를 모델에 반영하는 방법으로 Sun이 제안한 모델의 단점을 극복하였다. 본 논문은 앞서 발표한 연구 결과[5]에서 원인 노드의 개수와 실시간 교통정보를 반영하는 방법을 개선함으로 예측 오차를 줄였다.

Ⅲ. 베이지안 네트워크 설계와 구축

(1) 교통정보예측을 위한 베이지안 네트워크 설계

인과 모델이라고도 알려진 베이지안 네트워크는 원인노드들과 결과 노드로 구성된 변수들 사이의 조건부 확률을 표현하기 위한 모델로 변수들 간의 의존성을 방향성 에지로 표현한 비순환 그래프(DAG: directed acyclic graph)이다[14]. 방향성 그래프에서 머리가 되는 노드들은 꼬리가 되는 노드가 발생한 후 일어나게 되는 조건부 사건들로 정의할 수 있다. 조건부 확률은 사건 X가 발생했다는 가정 하에 사건 Y가 일어날 확률을 의미하며 이는 사건 X와 Y가 동시에 발생하였는데 X가 먼저 발생한 후 다음 Y가 발생했다는 것이다. 이때 사건 X와 Y가 동시에 발생할 결합 확률 함수는 식 1과 같이 표현된다.

$$P(X, Y) = P(Y \backslash X)P(X) \quad \text{(식1)}$$

식 1을 사건 X가 발생한 후 Y가 발생할 조건부 확률을 구하는 식으로 바꾸면 베이즈 규칙으로 잘 알려진 식 2가 된다.

$$P(Y \backslash X) = \frac{P(X, Y)}{P(X)} \quad \text{(식2)}$$

방향성 그래프로 표현되는 베이지안 네트워크는 원인이 되는 부모 노드와 결과 노드인 자식 노드들에 대한 조건부 사건들로 정의되며 그래프상의 모든 노드들에 대한 조건부 확률 관계를 결합 확률밀도 함수로 표현할 수 있다. 베이지안 네트워크를 구성하는 각 노드를 독립변수로 표현하면 n개의 노드를 가진 베이지안 네트워크의 결합 확률 분포는 식 3과 같이 표현된다.

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i \backslash X_{i-1}, ..., X_1) \quad (식3)$$

식 3에서 X_i의 부모 노드는 X_{i-1}이다. 교통정보 예측을 위한 베이지안 네트워크를 설계하기 위해서는 먼저 미래시간에 각 도로 링크의 교통 상황에 영향을 미치는 것이 무엇인지 고려하여 원인 노드들을 설정해야 한다. 교통정보 데이터의 분포를 분석한 결과 각 링크의 최근 교통 정보와 해당 링크에 대한 상·하류 링크의 최근 교통정보가 미래시간 해당 링크의 교통정보와 밀접한 관계가 있음을 알 수 있었다. 그림 1은 해당 링크의 현재 시간(t) 차량의 속도와 해당 링크에 대한 상류 링크의 최근 시간(t-2, 10분전) 차량 속도 분포를 보여준다. 그리고 그림 2는 해당 링크의 현재 시간(t) 차량의 속도와 해당 링크에 대한 하류 링크의 최근 시간(t-2) 차량의 속도 분포를 보여준다.

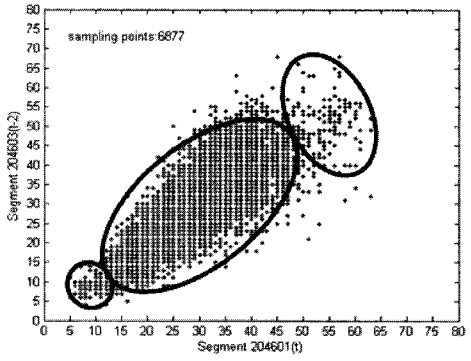


그림 1. 교통정보 데이터 분포: 특정 링크의 현재시간(t)과 상류 링크의 10분전(t-2) 차량 속도
Fig. 1 Data distribution of traffic information: Vehicles speed of a specific link at current time(t) and its upstream link(t-2) at 10 minutes ago

데이터의 분포는 그림에서 보는 것처럼 세 그룹으로 분류되고 88%이상의 샘플링 포인트들이 선형 분포를 보여주고 있는데 이는 해당 링크의 최근 속도와 상·하류 링크의 최근 속도가 차후 해당 링크의 교통 상황에 밀접한 영향을 준다는 것을 말해 준다. Sun 등이 제안한 예측 모델에서는 각 링크의 상류 링크만을 원인 노드로 설계하였다[3].

그림 2는 하류 링크도 상류 링크와 동일하게 미래 시간의 속도에 관련이 있음을 보여준다. 교통사고와 같은 문제가 발생하지 않을 경우 하류 링크의 10분 전의 교통 상황은 해당 링크의 10분 후 교통 상황이 된다. 따라서 하류 링크의 교통정보도 미래 시간의 교통정보를 예측하는 중요한 원인이 될 수 있다. 본 연구에서는 각 링크의 교통정보를 예측하기 위해 해당 링크의 상·하류 링크 모두의 최근 교통정보를 원인노드들로 설정하였다.

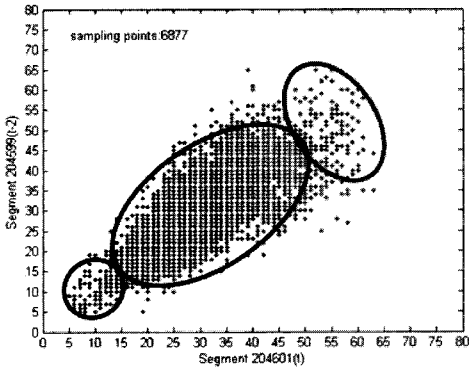
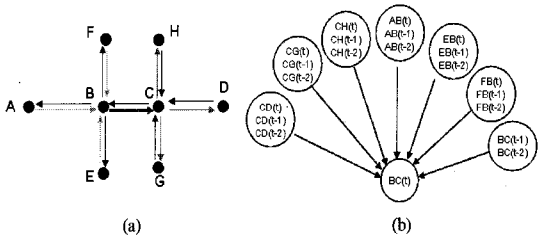


그림 2. 교통정보 데이터 분포: 특정 링크의 현재시간(t)과 하류 링크의 10분전(t-2) 차량 속도
Fig. 2 Data distribution of traffic information: Vehicles speed of a specific link at current time(t) and its downstream link(t-2) at 10 minutes ago

그림 3은 도로 네트워크와 링크 BC의 교통정보 예측을 위해 설계된 베이지안 네트워크를 보여준다. 선분 BC는 노드 B에서 C 방향의 흐름을 선분 CB는 C에서 B로의 흐름을 표현한다. 현재시간으로부터 10분후의 링크 BC의 교통 흐름을 예측하기 위한 베이지안 네트워크를 구축한다고 가정하면 그림 3의 베이지안 네트워크에서의 결과 노드는 BC(t)가 된다. 원인노드는 상류 링크인 CD, CG, CH와 하류 링크인 AB, EB, FB, 해당 링크 BC의 최근 교통정보이다.



(a) (b)

그림 3. 도로 네트워크와 설계된 베이지안 네트워크
Fig. 3 Road network and its designed Bayesian network

교통정보는 매 시간 주기마다 수집되어 갱신되는데 본 연구에서는 5분 단위로 수집되어 갱신되는 데이터를 사용하였다. 링크 BC의 10분 후의 교통 정보를 예측하기 위한 모델인 그림 3의 (b)에서 t는 현재 시간에 대한 교통정보를 의미하고 BC(t-1), BC(t-2)는 각각 5분 전, 10분 전의 링크 BC의 교통정보를 나타낸다.

베이지안 네트워크에서 결과노드를 Y라고 한다면 예측 값은 원인노드들에 대한 조건부 기댓값으로 계산할 수 있으며 식 4와 같이 표현된다. 식에서 보는 바와 같이 예측 값을 계산하기 위해서는 원인노드들에 대한 결과노드의 조건부 확률이 필요하다. 조건부 확률은 식 2에 의해 계산되며 베이지안 네트워크를 구성하는 모든 노드들의 결합 확률 밀도 함수를 통해 계산할 수 있다.

$$\hat{Y} = E[Y \setminus X] = \int YP(Y \setminus X)dY \qquad (식4)$$

(2) 결합 확률 밀도 함수의 파라메타 추정과 예측

설계한 베이지안 네트워크를 이용하여 예측 결과를 얻기 위해서는 모든 노드들에 대한 결합 확률 밀도 함수가 요구된다. 본 연구에서는 서울 강남구 일대의 도로 네트워크와 실시간 교통정보를 사용하였는데 각 링크들에 대한 상·하류 링크의 개수는 평균 4~5개로 나타났다. 원인 노드로 선택된 각 링크들의 최근 교통정보는 현재로부터 20분 이전까지의 정보를 고려하였으므로 베이지안 네트워크를 구성하는 노드의 개수는 평균적으로 25~30개가 된다. 따라서 베이지안 네트워크를 구성하는 독립변수들의 차수는 20차원 이상의 고차원 데이터가 된다. 고차원의 데이터들을 하나의 결합 확률 밀도 함수로 모델링하기는 어렵기 때문에 본 연구에서는 여러 개의 정규 분포가 서로 다른 비중으로 결합된 가우시안 혼합 분포를 따른다고 가정하였다.

가우시안 혼합 분포를 구성하는 서로 다른 정규 분포의 개수 즉, 요소의 개수는 3개로 지정하였는데 이는 그림 1, 2에서 본 것처럼 샘플데이터의 분포가 크게 세 개의 그룹으로 분류할 수 있는 형태를 보였기 때문이다. 가우시안 혼합 분포의 결합 확률 밀도 함수는 식 5와 같다 [15-16].

$$p(x \setminus \Theta) = \sum_{k=1}^{M} \alpha_k p_k(x \setminus \theta_k) \qquad (식5)$$

식에서 Θ는 가우시안 혼합 분포의 파라메타 집합인 $(\alpha_k, \theta_k, k = 1,2,3)$을 나타낸다. α_k는 각 요소의 가중치를 $\theta_k = (\mu_k, \Sigma_k)$은 각 요소의 파라메타 값을 μ_k는 각 요소의 평균을 의미하고 $\sum_l$은 공분산을 의미한다. 비지도 학습을 통해 샘플데이터의 로그-우도를 최대로 하는 클러스터링과 각 혼합성분 가우시안들의 파라메타들 $(\alpha_l, \mu_l, \sum_l)$을 추정할 수 있다[16]. 학습 데이터 집합으로 강남구 일대의 각 도로들에 대한 한 달 동안의 교통정보 데이터 중 1일에서 20일까지의 교통정보 데이터를 사용하였다. 교통정보 데이터는 각 링크들에 대한 5분 단위로 갱신된 차량의 평균 속도로 구성되어 있다. 따라서 학습을 위해 사용한 샘플링 데이터의 개수는 각 링크 당 6048(21일*24시간*12단위)이다.

본 연구에서는 가우시안 혼합 분포의 파라메타들을 추정하기 위해 샘플 데이터들의 로그-우도(log-likelihood)를 최대로 하여 국부적 최적 해에 수렴하게 해 주는 EM (Expectation Maximization) 반복 알고리즘을 이용하였다. EM 반복 알고리즘은 숨겨진 정보를 포함하고 있는 샘플 데이터에서 최적 해를 찾아내는데 유용한 방법으로 숨겨진 정보를 추정하는 E-단계 (Expectation Step)와 추정된 정보를 가지고 추정치를 개선하기 위해 로그-우도를 최대화시키는 M-단계 (Maximization Step)를 반복적으로 수행한다. 표본 데이터 집합 X, 숨겨진 정보 Y로 이루어진 확률변수 Q=(X, Y)에 대해 모든 데이터 X와 확률 변수 Z의 결합 우도는 식 6과 같다.

$$p(X, Q \setminus \theta) = \prod_{i=1}^{N} \prod_{k=1}^{M} \alpha_k^{z_{i,k}} p(x_i \setminus \mu_k, \Sigma_k)^{z_{i,k}} \qquad (식6)$$

식에서 $z_{i,k}$는 k번째 가우시안 분포에 데이터 x_i가 포함되면 1의 값을 가지고 그렇지 않을 때는 0의 값을 가지는 변수이다. N은 전체 데이터의 개수를 M은 가우시안 혼합 분포의 개수를 의미한다. 식 6에 로그를 취한 로그 우도의 기댓값 함수를 함수 A로 나타내면 식 7과 같다.

$$A(\theta, \theta^s) = E_Q[\log p(X, Q \setminus \theta) \setminus X, \theta^s] =$$
$$E_Q[\sum_{i=1}^{N} \sum_{k=1}^{M} (z_{i,k} \log \alpha_k + z_{i,k} \log p(x_i \setminus \theta) \setminus X, \theta^s)] \quad (식7)$$

M-단계에서는 식 7의 함수 A를 최대화시키는 파라메타 θ를 찾는다. 이를 위해 각 파라메타들에 대해 식 8을 편미분하여 0으로 두고 정리하면 식 9와 같은 다음 단계의 파라메타(θ^{s+1}) 추정치를 얻을 수 있다[16]. 모든 링크에 대해 구축된 베이지안 네트워크의 파라메타 추출 작업은 EM 반복 알고리즘을 이용하여 국부적 최적 해에 도달하기까지 식 8과 같이 반복적인 계산 작업을 통해 얻을 수 있다.

$$\alpha_k^{(s+1)} = \frac{1}{N}\sum_{i=1}^{N} p(k \setminus x_i, \theta^s)$$

$$\mu_k^{(s+1)} = \frac{\sum_{i=1}^{N} x_i p(k \setminus x_i, \theta^s)}{\sum_{i=1}^{N} p(k \setminus x_i, \theta^s)}$$

$$\Sigma_k^{(s+1)} = \frac{\sum_{i=1}^{N} p(k \setminus x_i, \theta^s)(x_i - \mu_k^{(s+1)})^2}{\sum_{i=1}^{N} p(k \setminus x_i, \theta^s)} \quad (k=1,2,3) \quad (식8)$$

결과 노드의 추정 값은 원인노드들에 대한 조건부 기댓값이므로 앞서 구한 결합 확률 밀도 함수의 조건부 확률을 이용하여 계산할 수 있다. 식 2에서 언급한 원인노드들에 대한 조건부 확률 P(Y\X)를 계산하는 식을 가우시안 혼합 분포를 적용하여 다시 표현하면 식 9와 같다.

$$P(Y \setminus X) = \frac{P(Y,X)}{P(X)} = \frac{\sum_{k=1}^{M} \alpha_k G(X;\mu_{kX},\Sigma_{kXX}) G(X;\mu_{kYX},\Sigma_{kYX})}{\sum_{k=1}^{M} \alpha_k G(X;\mu_{kX},\Sigma_{kXX})} \quad (식9)$$

식에서 μ_{kX}는 가우시안 혼합 분포의 k번째 요소로 분류된 샘플링 데이터들의 X값에 대한 평균을 Σ_{kXX}은 k번째 요소에서의 X 변수들 사이의 공분산을, Σ_{kXY}는 k번째 요소에서의 X변수와 Y변수들 사이의 공분산을 나타낸다. 식 9을 이용하여 결과 노드의 예측 값을 식 10과 같이 계산할 수 있다.

$$\begin{aligned}\hat{Y} &= E[Y \setminus X] \\ &= \int YP(Y \setminus X)dY \\ &= \sum_{q=1}^{M} \frac{\alpha_q G(X;\mu_{qX},\Sigma_{qXX})}{\sum_{k=1}^{M} \alpha_k G(X;\mu_{kX},\Sigma_{kXX})} \int YG(Y;\mu_{qY\setminus X},\Sigma_{qY\setminus X})dY \\ &= \sum_{q=1}^{M} \frac{\alpha_q G(X;\mu_{qX},\Sigma_{qXX})}{\sum_{k=1}^{M} \alpha_k G(X;\mu_{kX},\Sigma_{kXX})} \mu_{qY\setminus X} \quad \text{--------} \ (식10)\end{aligned}$$

식 10에서 분모 $\alpha_q G(X;\mu_{qX},\Sigma_{qXX})$와 분자 $\sum_{k=1}^{M} \alpha_k G(X;\mu_{kX},\Sigma_{kXX})$의 값은 추정된 가우시안 혼합 분포의 파라메타들로 미리 계산해 놓을 수 있다.

Ⅳ. 교통상황 분석과 반영 방법

베이지안 네트워크 모델의 가장 큰 단점은 돌발 상황이 발생했을 경우 예측 오차가 커질 수 있다는 것으로 이것은 베이지안 네트워크 모델이 훈련 데이터에서 패턴을 학습하는 것에서 기인한다. 일반적으로 돌발 상황에 대한 충분한 데이터를 확보하기 매우 어렵기 때문에 이러한 경우를 위한 학습이 힘들고 따라서 돌발 상황에 대한 예측 오차가 커질 수밖에 없다. 본 연구에서 이러한 단점을 보완하기 위해 현재의 교통 상황을 분석하여 돌발 상황이 발생할 경우 현재의 실시간 교통 정보를 예측 모델에 반영하기 위한 방법을 연구하였다.

표 1은 현재 교통상황 분석을 위해 사용된 값들을 보여 준다. 교통 상황 분석을 위해 사용된 측정 도구는 식 11과 같다. 누적 정보는 한 달 동안의 교통정보 데이터에 대하여 시간대별로 누적하여 계산한 평균과 분산을 의미한다. BS_t는 베이지안 네트워크에 의한 예측 결과를 나타내며 예측 속도는 현재 시간 t에서 예측된 속도를 실제 속도는 5분이 지난 뒤 실제로 관측된 속도를 의미한다. 실제 속도 RS_t는 미래 시간이 현재가 되었을 때 알 수 있으므로 t+1의 실제 속도는 알 수 없다.

표 1. 교통상황 분석에 필요한 정보
Table 1. Required information for analyzing the traffic status

	t-4 (20분전)	t-3 (15분전)	t-2 (10분전)	t-1 (5분전)	t (현재)	t+1 (5분 후)
누적 속도 (Accum. Speed)	μ_{t-4}, σ_{t-4}	μ_{t-3}, σ_{t-3}	μ_{t-2}, σ_{t-2}	μ_{t-1}, σ_{t-1}	μ_t, σ_t	μ_{t+1}, σ_{t+1}
예측 속도 (Predict Speed)	$BS_{(t-4)}$	$BS_{(t-3)}$	$BS_{(t-2)}$	$BS_{(t-1)}$	BS_t	$BS_{(t+1)}$
실제 속도 (Real Speed)	$RS_{(t-4)}$	$RS_{(t-3)}$	$RS_{(t-2)}$	$RS_{(t-1)}$	RS_t	?

$$w = \frac{\log\sqrt{\sigma_t}}{\log\sqrt{\sigma_t} + \log|BS_t - RS_t|} \quad \text{----} \quad (식11)$$

가중치 w는 0에서 1까지의 값을 가지며 최종 예측 값을 위한 계산에서 베이지안 네트워크에 의한 예측 값이 적용되는 비중을 의미한다. 실제속도와 예측 속도의 차이 $|BS_t - RS_t|$가 동일한 경우 표준편차가 높을수록 가중치의 값은 커져서 베이지안 네트워크에 의한 예측 결과를 선호하게 된다.

실시간 교통정보와 베이지안 네트워크에 의한 예측 결과를 이용하여 최종 예측 결과를 산출하는 식은 식 12와 같다. 실제속도와 예측 속도의 차이 $|BS_t - RS_t|$가 커질수록 가중치의 값은 낮아지므로 실시간 교통정보를 최종 예측 결과에 적용하는 비중이 높아진다. 예측속도와 실제 속도의 차이가 많이 발생했다는 것은 현재의 교통 상황이 일상적인 흐름과 다른 양상을 보여주고 있다는 것을 의미하므로 베이지안 네트워크에 의한 예측 결과보다 돌발 상황이 발생한 실시간 교통정보를 더 높은 비중으로 반영하게 된다.

$$PR_t = BS_t \cdot w + RS_t \cdot (1-w) \quad \text{----} \quad (식12)$$

V. 실험 결과

본 연구에서 사용한 실험 데이터는 서울 강남구 일대의 도로 네트워크와 도로 네트워크에 대한 한 달 동안의 교통데이터이다. 30일의 교통 데이터 중에서 1일부터 20일까지의 데이터는 4장에서 설명한 베이지안 네트워크 구축을 위한 훈련 데이터로 사용하였고 나머지 20일부터 30일까지의 데이터를 이용하여 정확도 분석을 위한 실험을 수행하였다. 실험 방법은 특정 시점(time point)을 현재 시간으로 설정한 후 5분 이후부터 1시간 이후까지의 교통정보를 예측하도록 하였다.

예측 모델의 정확도를 검증하기 위한 평가 도구로 제곱근-평균-제곱오차(RMSE)를 사용하였다. RMSE은 예측이나 추정 오차를 측정할 때 일반적으로 사용하는 평가 도구이다. 제곱근-평균-제곱오차를 구하는 식은 식 13과 같다. 식에서 y_i는 실측값을 $\hat{y}_i$는 예측 값을 의미한다.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad \text{---} \quad (식13)$$

그림 4와 5는 제안한 예측 모델의 예측 결과가 실제 속도에 어느 정도 근접하게 따라가는지를 보여 주기 위해 특정 링크(링크 202322)를 선택하여 10분, 30분 이후의 예측 속도와 실제 속도를 그래프로 그려본 결과이다. x축은 시점을, y축은 차량의 평균 속도(km/hour)를 나타낸다. 축 x에서 시점 288(24시간*12) 까지는 하루에 해당하는 기간을 나타내므로 그래프는 시작 일을 포함하여 3일 동안의 예측 속도와 실제 속도의 차이를 보여 준다. 그림에서 보는 것처럼 예측 속도는 실제 속도와 동일한 패턴으로 따라가고 있지만 실제속도가 t시간에 비해 t+1 시간에서 갑작스럽게 변화되는 곳에서는 예측결과가 실제 속도를 따라가지 못함을 볼 수 있다. 10분 이후와 30분 이후의 예측 결과를 비교해 보면 10분 이후의 예측 속도가 실제 속도에 더 근접하게 따라가는 것을 확인할 수 있는데 실제로 예측 오차 RMSE의 값도 각각 2.49, 2.81로 현재시간으로부터 멀어질수록 예측 오차가 조금씩 증가함을 보여주었다.

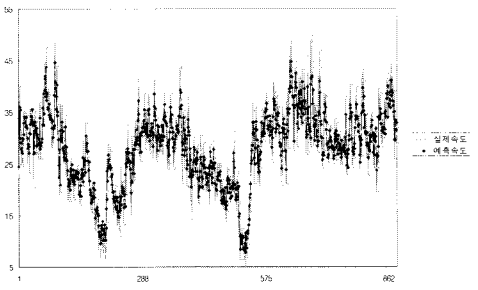


그림 4. 특정 링크에 대한 10분 이후의 예측 속도와 실제속도(RMSE: 2.49)
Fig. 4 Forecasting speed and its real speed of a specific link 10 min. later(RMSE: 2.49)

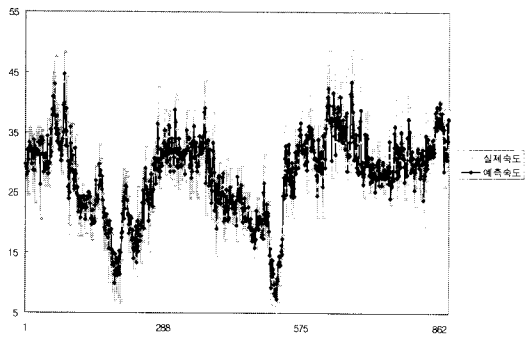


그림 5. 특정 링크에 대한 30분 이후의 예측 속도와 실제속도(RMSE: 2.81)
Fig. 5 Forecasting speed and its real speed of a specific link 30 min. later(RMSE: 2.81)

예측 오차를 분석하기 위해 열 개의 링크들을 임의로 선택하여 10분 이후부터 60분 이후까지 시간대 별로 RMSE 값을 비교해 보았다. 표 2는 제안한 예측 모델의 결과를 보여 주는데 본 연구팀이 2008년 11월에 출판한 논문 결과[5]와 비교할 때 RMSE 값이 전체적으로 25% 내외로 감소하였다. 이것은 첫째 베이지안 네트워크의 원인 노드의 개수를 증가시키고 둘째, 현재 교통 상황을 반영하기 위한 방법을 수정하여 시스템의 예측 결과를 개선시켰기 때문이다. 표에서 보는바와 같이 몇 개의 예외적인(*표시) 경우를 제외하고는 현재 시간으로부터 멀어질수록 RMSE의 평균값이 조금씩 증가하지만 1시간 후의 예측 오차도 평균 6이하를 보여 주었다.

표 2. 시간에 따른 예측 모델의 RMSE 변화
Table 2. changes of RMSE of our forecasting model with time

시간 / 링크번호	10분 후	20분 후	30분 후	40분 후	60분 후
(202322)	4.52	4.54	4.56	4.57	4.68
(203672)	4.22	4.36	5.29(*)	4.54(*)	4.70(*)
(202420)	5.53	5.58	5.66	5.83	6.03
(205774)	4.83	5.33(*)	5.23(*)	6.36	6.88
(208622)	3.87	4.07	4.25	4.27	4.30
(204638)	3.66	3.89	3.98	5.17(*)	4.69(*)
(207899)	4.48	4.82	4.98	5.27	5.42
(204653)	5.02	5.20	5.21	5.40	5.55
(204623)	4.12	4.13	4.13	4.38	4.49
(204571)	4.72	4.90	4.94	5.37	5.57
평균	4.49	4.68	4.82	5.11	5.23

VI. 결론

본 연구에서는 베이지안 네트워크를 이용한 단기 교통정보 예측 모델을 제안하고 개발하였다. 베이지안 네트워크의 결합 확률 밀도 함수는 가우시안 혼합 분포를 가정하였으며 파라메타 추출을 위해 EM 반복 알고리즘을 사용하였다. 실험을 위해 서울 강남구 일대의 도로 네트워크와 한 달 동안의 실시간 교통정보를 사용하였다. 예측 모델의 정확도 검증을 위해 예측된 속도와 실제 속도에 대한 제곱근-평균-제곱오차를 적용하였다. 실험 결과 본 논문에서 제안한 예측모델은 10분후, 30분후, 60분후 예측 결과의 제곱근-평균-제곱오차로 각 각 4.5, 4.8, 5.2의 값을 주었다. 현재 시간으로부터 60분 이후의 예측 속도의 오차가 6이내이므로 본 연구에서 사용한 도로 데이터가 서울 중심부의 교통 흐름의 변화가 많지 않은 도로라는 것을 감안하더라도 제안한 예측 모델이 실용적임을 보여주었다.

참고문헌

[1] 오철, ITS 진단 체계 구축 방안 연구, 한국교통연구원, 2005, 9월.

[2] 강연수, 문영준, 박유경, 이주일, 텔레매틱스 시대를 대비한 첨단 종합교통정보서비스 체계화 방안 연구, 교통개발연구원, 2003.

[3] Shiliang Sun, Changshui Zhang, Guoqiang Yu, "A Bayesian Network Approach to Traffic Flow," IEEE Transaction on Intelligent Transportation Systems, Vol. 7, No. 1, 2006.

[4] 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.

[5] Y. Yu and M. G. Cho, "A Short-Term Prediction Model for Forecasting Traffic Information Using Bayesian Network," Third 2008 International Conference on Convergence and Hybrid Information Technology, pp. 247-253, 2008.

[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] E. Fraschini and K. Ashausen, Day on Day Dependencies in Travel Time: First Result Using ARIMA Modeling: ETH, IVT institute for Transport, Feb. 2001.

[8] 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.

[9] Rui Wang, Hideki Nakamura, "Short Term Prediction Works in Traffic Engineering: The state-of-The-Art, ITS World Congress, 2002.

[10] 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

[11] 남궁성, 윤일수, 조범철, TCS 자료를 이용한 고속도로 통행 시간 예측, 한국도로공사 보고서, 2000.

[12] 김동호, 노정현, 박동주, "고속도로 통행시간 예측을 위한 과거 통행시간 이력 자료 구축에 관한 연구," ITS 학회 춘계 발표논문집, 2005.

[13] 이승재, 김범일, 권혁, "단기 통행시간예측 모형 개발에 관한 연구," 한국 ITS 학회 논문지 제 3권 제 1호, 2004.

[14] Finn V. Jensen, An introduction to Bayesian networks, UCL Press, 1996.

[15] 한학용 저, 패턴인식 개론, 한빛미디어, 2005

[16] Carlo Tomasi, Estimating Gaussian Mixture Densities with EM - a tutoral, Duke University

저자소개

유영중(Yu Young Jung)

1996년 2월 부산대학교 전자계산학과(이학사)
1998년 2월 부산대학교 전자계산학과(이학석사)
2002년 2월 부산대학교 전자계산학과(이학박사)
2002년 3월 ~ 현재 부산외국어대학교 컴퓨터공학부(부교수)
※관심분야 : 컴퓨터그래픽스, 애니메이션, 시뮬레이션

조미경(Mi-Gyung Cho)

1990년 2월 부산대학교 전자계산학과(이학사)
1992년 2월 부산대학교 전자계산학과(이학석사)
1998년 2월 부산대학교 전자계산학과(이학박사)
2000년 9월 ~ 2002년 8월 부산대학교 연구교수
2002년 9월 ~ 현재 부산외국어대학교 컴퓨터공학부 (조교수)
※관심분야 : 알고리즘, ITS/텔레매틱스, 바이오칩 설계