

# 데이터 마이닝을 이용한 교통 정보 분석 알고리즘 개발

## Traffic Information Processing & Decision Making using Data Mining Technique

강성규<sup>1</sup>, 정희석<sup>1</sup>, 이종수<sup>1</sup>, 김병성<sup>2</sup>

연세대학교 기계공학과, 연세대학교 기계공학과, 연세대학교 기계공학부,

현대-기아 연구 개발 본부 자동차 전자 개발센터

Seonggyu Kang<sup>1</sup>, Heeseok Jeong<sup>1</sup>, Jongsoo Lee<sup>1</sup>, Byungung Kim<sup>2</sup>

Dept. of Mechanical Engineering, Yonsei University<sup>1</sup>

Automotive Electronics R&D Center, Hyundai Motor Company&Kia Motors Corporation<sup>2</sup>

E-mail : mecha\_dog@yonsei.ac.kr

### 요 약

본 논문에서는 기존의 교통 상황 검지 장비들이 가지고 있는 문제점들을 해결 하기 위해 실제 통행속도 데이터의 해당 도로 속성들을 이용하여 데이터 마이닝을 통한 합리적인 오차 범위 내에서의 실시간 교통 정보 예측 알고리즘을 제안한다. 본 논문에서 제안하는 알고리즘은 데이터 마이닝의 분석 기법중 하나인 신경망(Neural Network)분석을 통하여 통행 속도 예측 근사 모델을 개발하는 것이며, 기존의 교통 상황 판단 알고리즘과의 결과 비교를 통해 비용 절감 효과와 속도 정보가 없는 도로까지의 합리적인 통행 속도 예측, 그리고 Off line상에서의 시간대별 교통 정보 제공이 가능함을 보인다.

Key words : 지능형 교통 시스템(ITS), 데이터 마이닝(Data Mining), 신경망 분석(Neural Network), 데이터 일반화(Generalized Working Relation)

### 1. 서론

오늘날 교통 혼잡 비용 증가에 따른 사회적 낭비가 지능형 교통 시스템에 대한 관심을 촉구시키면서 교통 상황 판단 알고리즘으로의 활발한 연구를 진행 시키게 되었다. 이러한 교통 혼잡 비용을 해결하고자 도로의 건설 확대 및 지하철의 건설등 대규모 시설투자를 추진 하고 있으나 새로운 교통 시설을 만드는데는 많은 시간과 막대한 건설 비용이 필요하기 때문에 현실적으로 어려움을 겪고 있는 실정이다. 이렇듯 교통난의 해소는 시설투자만으로는 근본적인 해결책이 될 수 없다는 판단하에 교통 문제를 해결하기 위해 기존의 교통시스템에 정보/통신/제어 등 첨단 기술을 접목시킨 지능형 교통 체계(ITS: Intelligent

Transport Systems)[6]도입과 적용을 적극 추진하여 교통의 안전과 효율을 높이는데 역량을 집중하고 있다. 이러한 ITS의 관심하에 현재까지 비컨, 패쇄TV, 검지기, 통신원등 다수의 검지장비를 이용한 교통 상황 판단 알고리즘이 개발되었음에도 불구하고 초기 투자 및 유지 보수 비용과다 소요 문제와 전 지역의 도로에 대한 정보 수집 불가능 그리고 On line상의 실시간 접속을 통한 이용료 부담이라는 큰 문제를 가지고 있다. 이에 본 논문에서는 데이터 마이닝 기법[1]을 이용, 실제 교통량 데이터의 도로 속성들을 이용하여 통행 예측 속도 근사 모델을 개발 함으로써 현재까지 개발된 교통 판단 알고리즘의 문제점들에 대한 해결 방안을 충족시키는 알고리즘을 제

시하였다.

## 2. 본론

### 2.1 데이터 마이닝

데이터 마이닝의 최종적인 목표는 기존의 정보 검색이나 보고서가 밝혀낼 수 없었던 정보를 밝혀 낸다는 것이며[5], 기업 경영, 과학 및 의료등 다양한 분야에서 단순한 데이터 수준이 아닌 보다 고급 수준의 정보 획득, 활용 및 효율적인 분배의 측면에서 최근 그 중요성이 급격히 부각되고 있다. 본 속도 예측 알고리즘 개발에 사용된 데이터 마이닝 제품은 분석흐름도와 노드를 이용하여 데이터 마이닝의 전 과정을 GUI환경에서 수행 할 수 있는 SAS(Enterprise Miner)이다.

### 2.2 실제 교통 데이터 수집

실제 교통데이터의 수집은 현대/기아 자동차 모젠(Mozen) 센터에서 제공한 서울/수도권 지역의 실제 교통 데이터를 이용하였으며, 본 연구의 속도 예측 개발 알고리즘을 위해 이용된 분석 범위에 대한 링크ID별 도로 속성들로는 시설물 종류, 도로 종류, 링크 종류, 차선 수, 링크 길이, 시간, 속도가 있다. 또한 2004년 1월13일부터 2004년 3월17일까지의 총 52일치 데이터 중 연휴해당 자료인 2004년 1월20일부터 2004년 1월24일까지의 5일치 자료를 제외한 47일치 데이터에 대한 분석을 수행하였다.

### 2.3 교통 상황 판단 알고리즘

운전자에게 제공되는 교통 정보의 형태는 과거 “정상”, “혼잡”, “지체” 등의 정성적인 표현에서 벗어나, 좀더 계량화되고 수치화된 구간 통행 속도의 형태로 제공되고 있다[6]. 그러나 이러한 예측된 통행 속도는 검지기 등을 통해 직접적으로 얻어지는 자료가 아니므로 구간 통행 예측 속도를 만들어 내는 작업을 필요로하게 된다. 본 연구에

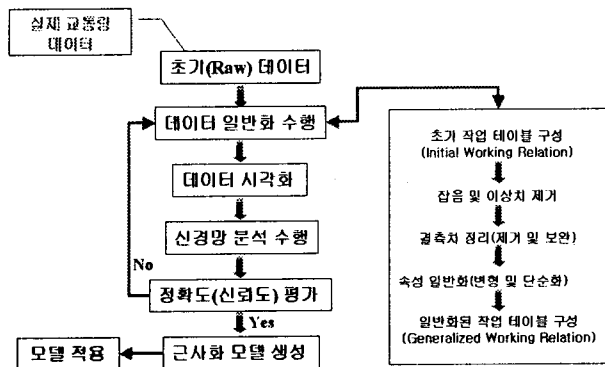


그림 1. 속도 예측 개발 알고리즘 흐름도

서 개발된 알고리즘은 실제 교통데이터로부터 데이터를 가공, 처리하여 구간 통행 속도 산출을 위한 변환 작업을 지속적으로 수행하였고, 이처럼 다양한 정보 수집 단위로부터 수집된 자료를 구간 통행 예측 속도로 변환하는 과정을 통해 실제 수정형 덩굴망 알고리즘에 의해 개발된 경로 탐색 알고리즘에 적용 할 수 있는 통행 속도 예측 모델을 개발하게 된다. 속도 예측 모델의 개발 과정에 대한 전체 흐름도인 그림 1은 본 연구에서 사용된 여러 형태의 데이터 마이닝 분석을 직접 수행한 후 가장 적합하다고 판단되는 데이터 마이닝 과정에 대한 흐름도이다.

실제 교통 데이터에 대한 데이터 마이닝 분석 과정 중 가장 중요하다고 생각되는 부분은 전처리 과정이며, 분석 기법으로서 제공되고 있는 신경망 모형[2,3]을 통해 속도 예측을 위한 근사화 모델을 생성하고 이를 실제 수정형 덩굴망 알고리즘에 의해 개발된 경로 탐색 알고리즘에 적용하였다.

### 2.3.1 데이터 전처리

일반 적으로 교통데이터에서는 모든 링크에 대한 구분을 각 링크별 해당ID를 통해 구분을 하고 있기 때문에 모든 도로에서의 동일 구성 인자별 상/하행선 구분이 자유롭다. 하지만 전국을 대상으로 하는 본 알고리즘 개발 단계에서는 예측에 대한 범위 제약으로 인하여 링크별 해당ID를 사용하지 않고 모든 도로속성에 대한 특성에 의해 동일 링크 구성조합에 대한 상/하행선을 구분할 수 있어야 했다. 이에 그림 2와같이 진/출입 링크 속성에 대해서는 본선 링크에 진/출입하는 시설물 종류, 도로 종류, 링크 종류별 모든 링크 길이 속성들의 길이에 따른 변환치를 입력 변수로 변환하여 적용하였고, 본선 링크 속성에 대해서는 시설물 종류, 도로 종류, 링크 종류, 차선 수의 본선 링크 속성이 가지고 있는 고유 속성들을 이용하게 되었다.

동일 링크 구성 조합에 대한 도로 특성별 세부 구분 방법은 표 1에서 표4와 같다.

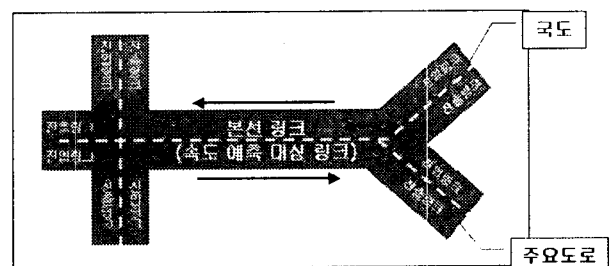


그림 2. 속도 예측을 위한 입력 속성

표 1. 도로 인자 특성별 설계변수 구성

설계 변수	Facility	In/Out ①	Road_case	In/Out ②	
설계 범위	0:일반도로 1:교량 2:터널 3:고가도로 4:지하도로	각속성별 해당 링크 길이	0:고속도로 1:도시고속도로 2:국도 3:국가지원 지방도 4:지방도 5:주요도로1 6:주요도로2 7:주요도로3 8:기타도로1 9:기타도로2 10:세도로 11:페리항로	각속성별 해당 링크 길이	
설계 변수	Link_case	In/Out ③	Num_lane	Time	Velocity
설계 범위	1:분선(비분리) 2:분선(분리) 3:연결로(JC) 4:교차로의동로 5:연결로(IC) 7:SA레이어 8:복합교차점 9:로터리내링크 10:P-Turn링크 11:U-Turn링크 6,12:reserved 13:진/출입로	각속성별 해당 링크 길이	1: 1차선 2: 2차선 3: 3차선 4: 4차선 5: 5차선 6: 6차선 7: 7차선	1[min] - 1440[min]	2[km/h] - 140[km/h]

표 2. In/Out① 해당 링크 길이

설계 변수	Facility Length (In/Out)
설계	0 : 일반 도로 1 : 교량 2 : 터널
범위	3 : 고가 도로 4 : 지하 도로

표 3. In/Out② 해당 링크 길이

설계 변수	Road_case Length (In/Out)
설계	0 : 고속 도로 1 : 도시 고속 도로 2 : 국도, 국가 지원 지방도, 지방도
범위	3 : 주요도로1, 주요도로2, 주요도로3 4 : 기타도로1, 기타도로2, 세도로, 페리항로

표 4. In/Out③ 해당 링크 길이

설계 변수	Link_case Length (In/Out)
설계	0 : 분선(비분리), 분선(분리) 1 : 연결로(JC), 교차로의 동로, 연결로(IC), SA 레이어
범위	2 : 복합 교차점 링크, 로터리내 링크, P-Turn & U-Turn 링크, reserved, 진/출입

표 1에서 표 4는 본 연구에서 통행 속도 예측을 위해 쓰인 도로 속성들에 대한 세부 결과 테이블을 이해하기 쉽게 구성해 놓은 것이다. 동일 링크 구성 조합에 대한 도로 특성별 구분을 위한 각 항목별 해당 링크 길이에 대한 적용 사항 중 표 3과 같이 도로 종류에서의 국도, 국가 지원 지방도, 지방도를 하나의 클래스로 일반화 하였고, 마찬가지로 기타도로1, 기타도로2, 세도로, 페리항로 등 링크 길이에 대한 속성별 유사점을 갖는다고 판단되는 부분에 있어 일반화를 수행하였다.

결과적으로 도로 특성별 구분을 위한 설계변수는 진/출입 링크길이에 대한 연속형 변수 24개씩과 15분 간격 시간에 대한 명목형 변수 96개를 포함하여 분선 링크 속성까지 총 152개에 대한 설계 변수가 사용되게 된다.

2.3.2 인공신경망(Neural Network)

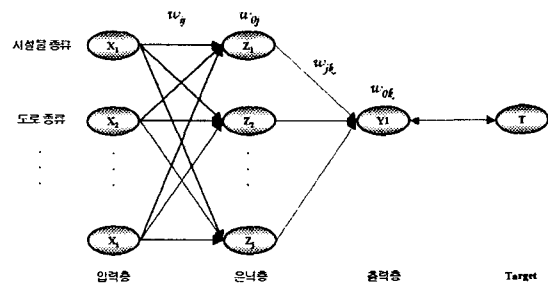


그림 3. MLP 신경망 구조

인공 신경망[2,3]에는 여러 가지 다양한 모형이 있으나 그 중 본 연구에서 사용된 모형은 그림 3과 같이 MLP신경망이다. 본 연구에서는 여러 번의 실제 테스트를 통해 1개의 은닉층과 15개의 은닉마디를 사용하였고, 훈련이라고 하는 데이터로부터의 연결강도인 계수들을 추정하기 위해 오차 제곱 합이라고 불리는 목적함수를 최소화 하는 문제로 적용하였다.

$$\sum_{i=1}^n (Y_i - P_i)^2 \quad (1)$$

식(1)의 \$Y\_i\$는 \$i\$번째 개체의 실제 속도 값이고 \$P\_i\$는 \$i\$번째 개체의 예측 속도 값이 된다.

또한 입력 변수들을 선형결합으로 변환하기 위해 결합함수로 선형 함수를 사용하였고, 입력 변수 또는 은닉 마디의 결합을 변환하기 위해 사용한 활성화 함수는 쌍곡선 탄젠트 함수를 사용하였다.

본 연구에서는 전통적인 신경망이 가지는 수치 해석적 비수렴성과 과적합 같은 단점을 보완하기 위해서 1,000,000개의 임의 표본에 대한 사전 최적화를 수행하였다.

2.3.3 데이터 평가 및 결과 분석

본 연구에서는 연속형 목표 변수의 수치적 정확도 문제 해결을 위해 훈련용 데이터에 대한 임의의 평가용 데이터의 예측 속도 값을 비교함으로써 실제 데이터의 평균 속도와 신경망 분석을 통해 예측된 속도 값의 차이가 5%이내가 되도록 연구를 진행하였으며, 임의의 불특정 표본(훈련용 데이터)에 대한 해당 구성 인자별 예측 속도 값이 평가용 데이터의 예측 속도 값에 대해 모두 동일하였다.

표 5는 실제 예측된 속도 변화 분석을 위해 속도 예측 알고리즘 결과에 대한 진/출입 해당 링크 길이를 포함한 두 가지 경우의 도로 구성 인자 조합 조건을 제시한 것이다. 특정 진/출입 링크 조건에 대한 시간대별 예측 속도 변화는 그림

표 5. 특정 도로 인자 구성 조건

	Facility	Road case	Link case	Num lane	Link length
구성	0: 일반도로	5: 주요도로1	1: 본선(비분리)	2: 2차선	200[m]
조건	0: 일반도로	1: 도시고속도로	2: 본선(분리)	3: 3차선	200[m]

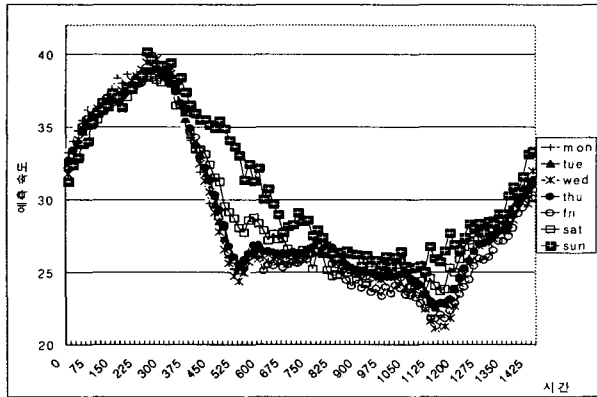


그림 4. 시간대별 예측 속도 [0 5 1 2 200]

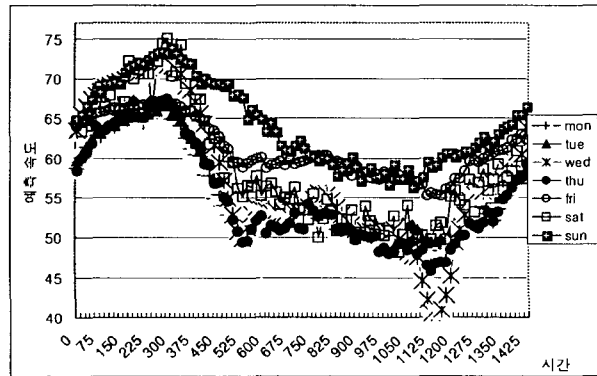


그림 5. 시간대별 예측 속도 [0 1 2 3 200]

4, 5와 같다. 예측된 시간대별 속도 변화를 보면 그림 4와 그림 5의 도로 구성 조합별 평균속도의 차이가 대략 25km/h정도인 것을 확인할 수 있으며, 이는 차선 수의 영향 보다는 주요도로와 도시 고속도로와의 평균 속도가 차이가 나고 있다는 것으로 판단된다. 또한 그림 4와 그림 5의 도로 구성 조합 조건 모두 출/퇴근 시간대의 정체 현상이 심하게 나타나고 있는 것을 확인할 수 있으며, 새벽 시간대의 평균 속도와 출/퇴근 시간대의 평균 속도가 20km/h정도 차이가 나고 있는 것을 확인할 수 있다. 본 연구에서는 실제로 그림 4, 5와 같은 도로 구성 조합 조건이 무수히 많이 존재하며 모든 경우의 도로 구성 조합 조건에 대한 속도 값이 데이터 마이닝[1,3]을 통해 예측되고 있다. 본 연구를 통해 개발된 도로 구성 조합별 예측 속도 알고리즘은 거리, 속도, 회전방향 선호도 그리고 교차로 통행비용에 의해 개발된 경로 탐색 알고리즘에서 시간대별 예측 속도로서 적용이 되며, 실제 속도의 개념만 적용된 추천 경로는 그림 6과 같다.

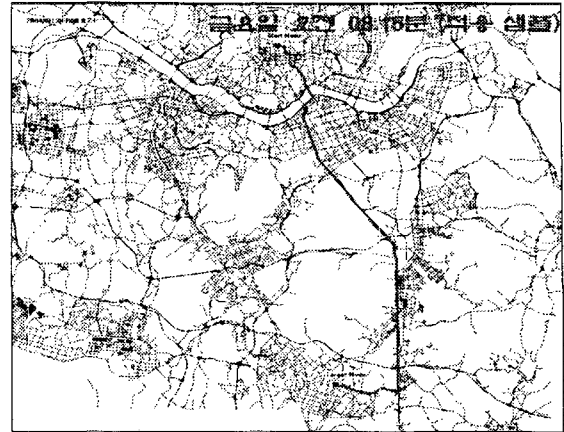


그림 6. 추천 경로 (창덕궁 ⇒ 아주대 삼거리)

### 3. 결론 및 향후 과제

본 논문에서는 실제 해당 도로 속성들을 이용하여 기존의 속도 정보가 없는 도로를 포함한 전국을 대상으로 한 통행 속도 예측과 On line상의 실시간 교통 정보 이용 요금 해결을 위한 방안으로서의 합리적인 교통 정보제공이 가능한 알고리즘을 제안 하였다. 본 논문에서 사용된 진/출입별 고유 링크 길이는 상/하행선 구분 및 특정 링크 구분을 위한 하나의 적용 사례이며, 나아가 진/출입 고유 링크 차선 수 및 본선 링크 길이, 그리고 진/출입 고유 링크와 차선 수, 본선 링크 길이를 동시에 고려하는 연구를 수행한다면 세부 링크 구간별 정체 상황 및 상/하행선 구간들의 현실적 표현에 조금 더 접근할 수 있는 결과를 도출해 낼 수 있을 것으로 본다.

### 4. 참고 문헌

- [1] Jiawei Han, "Data Mining : Concepts and Techniques", Elsevier Pte. Ltd, 2003. 9
- [2] Simon Haykin, "Neural Network a Comprehensive Foundation", Prentice-Hall Inc., 1994.
- [3] J. Bigus, "Data Mining with Neural Network" McGraw Hill, 1996
- [4] U. Fayyad, G. Piatetsky-Shapito, P. Smyth, R. Uthurusamy, "From Data Mining to Knowledge Discovery: An Overview", AAAI/MIT Press, CA, 1996
- [5] Lin, Feng Yu, McClean, Sally, "A data mining approach to the prediction of corporate failure, Knowledge-Based Systems," Volume 14, Issues 3-4, pp.189-195
- [6] 최종욱, 민중영, 남궁성, 이원하, "ITS와 첨단 정보기술", 도서출판 참말, 1997. 9