# 다중추정방법에 의한 전자상거래 에이전트

김우정, 이수원
숭실대학교 컴퓨터학과

# Electronic Commerce Agent using Multi-Estimation Method

Woo-Jung Kim, Soo-Won Lee
School of Computing, Soongsil University

## 요 약

추정을 위한 방법으로는 K-NN과 회귀분석, 신경망 등의 다양한 방법을 적용할 수 있다. 그러나 K-NN의 경우 거리에 의해서만 결과를 추정하므로 각 속성에 대한 가중치가 속성 값들의 간격에 의해 결정되고, 회귀분석은 하나의 선으로 데이터의 경향을 표현하므로 속성의 가중치는 고려되지만, 데이터의 분포가 넓을 경우에는 많은 오차를 포함하게 되는 데이터에 의존적인 문제가 존재한다. 따라서 본 연구에서는 이러한 방법들을 혼합하여 데이터에 의존적인 문제를 보안할 수 있는 다중분석방법을 제안한다.

## 1. 서론

정보화 사회에서 효과적인 지식 추출의 도구로서 전 산업분야에 적용되고 있는 데이터마이닝(Data Mining) 기술은 자동화되고 지능을 갖춘 데이터베이스 분석기법으로 90년대 초반부터 지식발견(KDD : Knowledge Discovery in Databases)[1], 정보발견(Information Discovery), 정보수확(Information Harvesting) 등의 이름으로 소개되어 왔다. 그러나 현재 데이터마이닝은 일반적으로 지식발견의 과정 중 대용량의 데이터베이스의 데이터로부터 패턴인식, 통계적 기법, 인공지능 기법 등을 이용하여 숨겨져 있는 데이터간의 상호 관련성, 패턴, 경향 등을 추출하는 것으로 정의한다. 또한 상거래 분석, 판매전략 수립, 수요예측, 고객관리 등의 의사결정에 활용하는 도구로서 정의하며, 다양한 종류의 새로운 기법들에 대한 많은 연구가 진행되고 있다. 데이터마이닝의 종류로는 연관규칙 발견(Association Rule Discovery), 분류(Classification), 추정(Estimation), 예측(Prediction), 연관성 분석(Link Analysis), 군집화(Clustering) 등이 존재하며[1][2], 본 연구에서는 입력 데이터로부터 연속적인 변수를 산정하는 추정을 웹 상의 중고 자동차 시장 데이터 적용하여 실세 데이터에 적용할 수 있는 효과적인 추정 방법을 연구하여 에이전트의 지식으로 활용한다.

## 2. 관련 연구

추정 방법중 K-NN(K Nearest Neighbor)[3]은 기본적인 분류 알고리즘으로 임의의 점과 가장 가까운 K개의 점을 이용하여 임의의 점의 이산적인 값 또는 연속적인 변수를 산정하는 방법이다. K-NN의 연속변수 산정 방법은 임의의 점과 가까운 K개의 점을 평균하여 연속변수를 산정하는 방법과 거리에 따라 가중치를 다르게 하여 연속변수를 산정하는 방법 등이 존재한다. 그러나 KNN은 단순히 점과 점 사이의 거리를 이용하여 연속변수를 산정하기 때문에 속성(Attribute)의 중요도가 속성 값들의 간격에 의해 결정된다는 문제점이 있다.

$$\langle a_1(x), a_2(x), \ldots, a_n(x) \rangle \qquad [식1]$$

$$d(x_i, x_j) \equiv \sqrt{\sum_{r=1}^{n} (a_r(x_i) - a_r(x_j))^2} \qquad [식2]$$

회귀분석(Regression Analysis)[3]은 통계적 기법의 하나로 속성들 간의 관련성을 통하여 특정 변수의 변화를 회귀선으로 표현하여 임의의 점에 대한 값을 회귀선에 적용하여 값을 산정하는 방법으로 연속변수의 값을 추정하는데 좋은 방법이다. 그러나 회귀분석은 분석 데이터를 하나의 선으로 표현하기 때문에 데이터가 넓게 분포되어 있을 경우에는 많은 오차를 포함하게 된다는 문제점을 가지고 있다.

$$f(x) = w_0 + w_1 a_1(x) + \ldots + w_n a_n(x) \qquad [식3]$$

$$E \equiv \frac{1}{2} \sum (f(x) - f(x))^2 \qquad [식4]$$

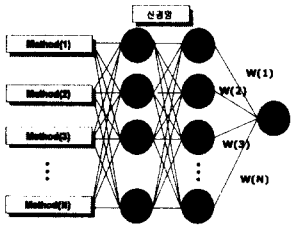
$$\Delta w_j = \eta \sum (f(x) - f(x)) a_j(x) \qquad [식5]$$

LWPR(Locally Weighted Polynomial Regression)[4]은 회귀분석의 한 방법으로 최소자승법을 사용하나 각 point들에 같은 가중치를 적용하는 것이 아니라 Query point와의 거리에 따라 각 point에 다른 가중치를 적용하여 회귀선을 구하는 방법으로 일반 회귀분석보다 좋은 성능을 보이나 일반 회귀분석의 문제점을 해결하지는 못한다.

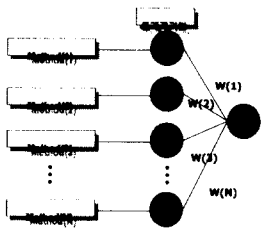
신경망(Neural Network)은 인공지능 분야의 하나로 인간의 신경 세포인 뉴런(Neuron)을 모방하여 만들어진 네트워크로 노드들간의 연결강도(Weight)를 조정함으로써 학습을 하는 시스템으로 함수근사, 분류, 군집화, 예측등의 복잡한 문제를 해결할 수 있으나 학습한 내용을 알기 어렵고 학습시간이 길다는 단점을 가지고 있다. 앞에서 제시한 추정 방법들은 데이터의 구성에 따라 성능의 차이를 보이는 데이터에 의존적인 성격을 띠고 있다. 따라서 본 연구에서는 추정 방법들을 혼합 적용하여 데이터의 의존성을 줄여 효과적으로 추정을 할 수 있는 추정 방법을 연구한다.

### 3. 다중추정방법

본 논문에서 제안하는 추정방법은 각각의 방법들에 의해 생성된 결과들을 입력 데이터로 사용하여 신경망 또는 여러 가지 통계적 기법들을 적용하여 개선된 결과 획득하는 것으로서 각 방법들에 가중치(Weight)를 조정함으로써 종합적인 새로운 결과를 얻는 다중추정방법을 연구한다. 다중추정방법의 구조는 [그림1]과 [그림2]에 나타나 있다.




[그림1] 신경망을 이용한 다중추정방법




[그림2] 통계적기법을 이용한 다중추정방법

신경망을 이용한 다중추정방법은 RBFN(Radial Basis Function Network)[5] 또는 MLP(MultiLayer Perceptron)[6] 등을 이용하여 추정하며, 통계적 기법을 이용한 다중추정방법을 회귀분석과 LWPR 등을 이용하여, 추정한다.

### 4. 실험 내용 및 결과

본 연구는 벼룩시장 사이트(www.findall.co.kr)의 사용자들이 중고 자동차를 판매하기 위해 웹 상에 올린 데이터를 사용하여 임의의 사양에 해당하는 중고 자동차의 가격을 에이전트가 추정하기 위한 효과적 추정방법을 연구하기 위한 실험이다.

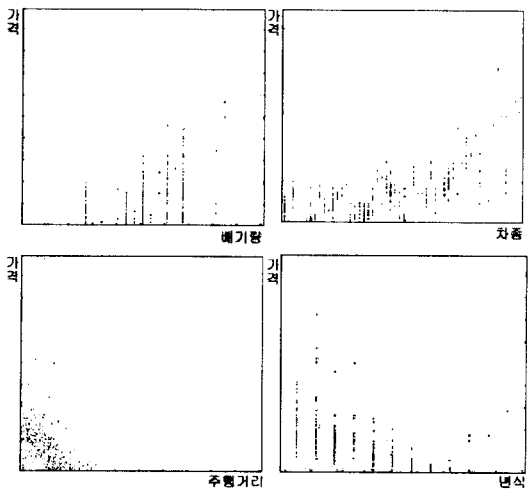
실험 데이터는 [표1]과 같으며, 실험에서는 [그림3]와 같이 실제 자동차 가격에 영향을 줄 수 있는 요소들만을 추출하여 6개의 속성으로 구성하였으며, 총 데이터 수는 420개, 데이터의 각 속성별 분포는 [그림4]와 같이 분포되어 있다. 데이터의 내용이 사용자의 주관에 의해 구성된 데이터이기 때문에 데이터의 분포가 넓게 분포되어 있다.

[표1] 데이터

| 제조회사 | 현대 |
|---|---|
| 차종 | 아토스 |
| 등급 | CAMY |
| 년식 | 1999-1 |
| 기어타입 | 수동 |
| 배기량 | 1300 |
| 색상 | 분홍 |
| 주행거리 | 20000km |
| 연료형태 | |
| 옵션 | 에어콘/ 파워핸들 |
| 사용구분 | 중고 |
| 지역 | 서울 |
| 희망판매가 | 400만원 |
| 판매주체 | 개인직거래 |

| 차종 | 년식 | 주행거리 | 배기량 | 기어타입 | 가격 |
|---|---|---|---|---|---|

[그림3] 실험 데이터 포맷




[그림4] 데이터 분포

실험 방법으로는 K-NN과 회귀분석, LWPR, 통계적 기법을 이용한 다중추정방법에 대해 실험하였다.

K-NN 실험결과는 [표2]과 [표3]와 같으며, 각 수치들은 평균오차(오차의 표준편차)로 표현되어 있다. K개의 변수를 평균하여 연속변수를 산정한 방법보다 거리의 가중치를 이용하여 연속변수를 산정한 방법이 오차가 적었다.

[표2] K-NN (평균)

| K | 차종,년식,주행거리 | 차종,년식,주행거리, 배기량 | 차종,년식,주행거리, 배기량,기어 |
|---|---|---|---|
| 3 | 68.20 (133.07) | 73.82 (139.80) | 81.44 (150.86) |
| 4 | 68.79 (133.03) | 71.54 (136.20) | 77.36 (143.33) |
| 5 | 67.48 (128.94) | 74.12 (141.20) | 77.37 (143.21) |
| 10 | 71.25 (134.77) | 76.65 (147.74) | 80.53 (151.86) |
| 20 | 75.27 (137.28) | 78.19 (146.09) | 90.09 (157.25) |
| 30 | 76.37 (144.68) | 82.92 (151.06) | 97.88 (165.66) |

[표3] K-NN (거리 가중치)

| K | 차종,년식,주행거리 | 차종,년식,주행거리, 배기량 | 차종,년식,주행거리, 배기량,기어 |
|---|---|---|---|
| 3 | 64.90 (137.59) | 69.77 (144.69) | 77.03 (149.60) |
| 4 | 64.69 (136.12) | 68.23 (141.34) | 74.30 (145.11) |
| 5 | 63.18 (131.56) | 67.35 (140.34) | 73.29 (143.49) |
| 10 | 64.21 (130.83) | 69.00 (142.86) | 73.06 (144.34) |
| 20 | 66.18 (130.62) | 69.38 (140.05) | 75.04 (144.68) |
| 30 | 66.44 (133.33) | 70.57 (140.90) | 78.44 (150.18) |

회귀분석 실험결과는 [표4]의 내용과 같으며, 회귀분석의 경우는 K-NN보다 큰 오차를 가지고 있다.

[표4] 회귀분석

| 차종,년식,주행거리, 배기량,기어타입 | 차종,년식,주행거리, 배기량 | 차종,년식,주행거리 |
|---|---|---|
| 100.59 (160.91) | 101.65 (160.54) | 101.19 (160.96) |

LWPR의 결과는 [표5]의 내용과 같으며, 실험결과는 KNN과 비슷한 결과를 얻었으며, 회귀분석보다는 좋은 결과를 보여주고 있다.

[표5] LWPR

| 차종,년식,주행거리, 배기량,기어타입 | 차종,년식,주행거리, 배기량 | 차종,년식,주행거리 |
|---|---|---|
| 70.915(132.318) | 69.047(127.841) | 69.086(127.704) |

다중추정방법 실험결과는 [표6]의 내용과 같으며, K-NN과 회귀분석, LWPR 보다 조금의 성능향상은 보이지만 큰 차이를 보이지는 않는다. 이것은 실험 데이터가 너무 넓게 분포되어 있으며, 다중추정방법에 사용된 방법들의 종류가 K-NN과 회귀분석, LWPR로 고정되어 다양하지 않고, 통계적 기법인 회귀분석만을 사용하여 실험하였기 때문이다. 따라서 신경망과 다른 통계적 기법을 사용하여 실험을 하면 보다 개선된 결과를 보일 것으로 예상된다.

[표6] 다중추정방법

| 다중추정방법 (통계적방법) | 평균오차 | 오차의 표준편차 |
|---|---|---|
| | 61.112 | 123.601 |

## 5. 결론 및 향후과제

본 논문에서 연구하고자 하는 다중추정방법은 실험 데이터의 차종별 분포가 균일하지 않고 너무 넓은 범위로 분포되어 있으며, 다양한 방법으로 실험을 하지 못하여 기대만큼의 결과를 얻지 못하였다. 따라서 향후 연구 계획은 차종별 분포가 균일하도록 데이터를 보강하여 신경망 또는 효과적인 통계적 기법을 사용하여 실험의 폭을 넓히는 것이며, 본 실험에서는 실험자가 임의로 데이터의 속성 중 실험에 영향을 줄 속성들을 선택하였지만, 속성 선택을 자동화할 수 있는 기법에 대한 연구도 필요할 것이다. 또한 관련 데이터 끼리의 군집화를 통해 군집화된 데이터를 기준으로 추정을 하는 방법에 대한 연구도 필요할 것이다.

## 참고문헌

[1] S. Chen, J. Han and P. Yu, "Data Mining : An overview form Database Perspective", IEEE Trans. on Knowledge and Data Engineering, 1997

[2] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami, "Data Mining : A Performance Perspective", IEEE Transactions on Knowledge and Data Engineering, vol. 5, No.6, December 1993, pp. 914-925.

[3] T. M. Mitchell, Machine Learning, The McGraw-Hill Co., pp. 230-248

[4] A. W. Moore and J. Schneider and K. Deng, "Efficient Locally Weighted Polynomial Regression Predictions", Proceedings of International Machine Learning Conference, 1997

[5] S. Chen, C.F.N Cowan, and P.M. Grant, "Orthogoanl Least Squares Learning for Radial Basis Function Networks", IEEE Transactions on Neural Networks, 1991

[6] Richard P. Lippmann, "An Introduction to Computing with Neural Nets", IEEE ASSP MAGAZINE APRIL 1987