베이지안 네트워크와 신경망을 이용한 구매 패턴 분석

A Purchase Pattern Analysis Using Bayesian Network and Neural Network

황정식, 피수영, 손창식, 정환묵
대구가톨릭대학교 컴퓨터정보통신공학과

Jeong-Sik Hwang, Su-Young Pi, Chang-Sik Son, Hwan-Mook Chung
Faculty of Computer and Information Communication Engineering
Catholic University of Daegu
E-mail : icsman@cu.ac.kr

요 약

실세계에서 일어나는 문제는 매우 복잡하고 다양하기 때문에 예측하기가 어렵고 다양한 상황들이 발생한다. 특히, 소비자의 구매에 따르는 행동을 분석하고 소비자의 다양한 기호를 예측하기 위해서는 구매자의 심리적 요인과 내적 요인이 많은 영향을 미치게 된다. 이러한 요인들은 직접적인 정보 처리가 어렵기 때문에 정보의 불확실성을 취급하는 기술이 필요하다.

따라서 본 논문에서는 상품 구매에 따르는 소비자의 구매행동 패턴을 분석하기 위해 판매자의 노하우와 소비자의 구매의식을 조사하여 이 데이터를 바탕으로 베이지안 네트워크를 구성하고 구매패턴을 분류하는 방법을 제안하였다. 특히, 베이지안 네트워크를 이용하여 불필요한 속성을 가진 데이터를 제거한 후 코호넨의 SOM을 이용하여 소비자의 구매 패턴을 분류하도록 하였다.

key word : Bayesian Network, SOM

1. 서론

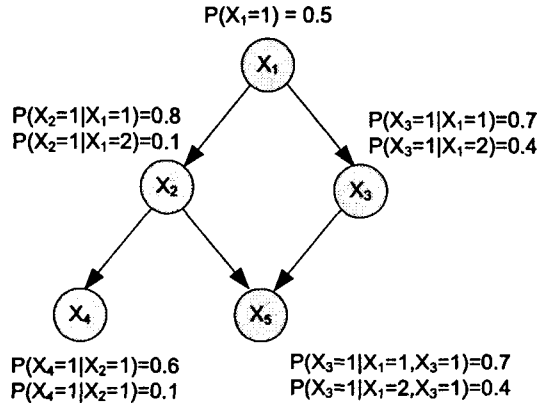
실세계에서의 문제는 매우 복잡하고 예측하기 어려우며 다양한 상황들이 발생한다. 특히, 소비자의 구매행동 예측에서는 소비자의 기호가 다양해짐에 따라 기존의 고객 프로필 데이터나 구매 경력 데이터만으로는 소비자의 구매 심리나 행동을 분석하기에는 불충분하다. 따라서 소비자의 구매에 이르는 심리나 내부 상태까지 깊이 분석하여 정확한 상품을 예측하고 기업의 이익을 최대화하는 마케팅 전략을 세우기 위해 소비자의 구매에 이르는 행동을 조사하는 설문 조사를 실시하곤 한다[1]. 그러나 설문 조사는 비교적 간단하고 대규모 조사가 실시하기 쉽지만 응답 작업의 부담이 커서 응답을 하지 않는 경우가 많고 심리적으로 대답하기 어려운 문항에 대해서는 신뢰성이 낮아 질 수 있다. 이로 인해 소비자의 구매 심리나 내적 상태를 이끌어 내기 힘들다[2].

베이지안 네트워크는 결손치(missing value) 즉, 관측이 곤란한 요소를 다루는 것이 가능하고 추측되는 가설의 확신도를 실제 데이터를 바탕으로 검증을 할 수가 있으며, 전문가의 지식을 네크워크 구조로 도입하는 것이 가능한 장점이 있다. 따라서 본 논문에서는 설문조사에서 수집되는 데이터의 성질을 고려하여 소비자의 행동 모델을 구성하기 위해 판매자의 지식과 설문조사에서 수집된 데이터를 베이지안 네트워크를 이용하여 모델링 하였다. 구축된 베이지안 네트워크에서 불필요한 속성의 데이터를 제거한 후, 코호넨 맵을 이용하여 소비자의 구매행동 패턴을 분석하였다

2. 관련연구

2.1 베이지안 네트워크

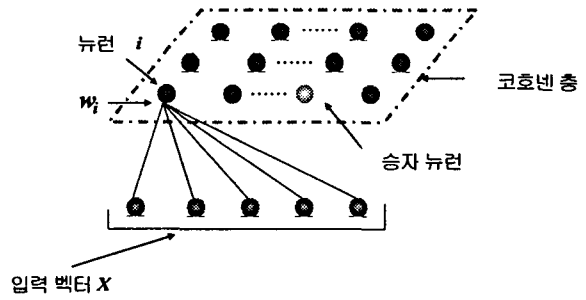
베이지안 네트워크는 확률변수들 간의 의존관계를 조건부 확률로서 기술한 네트워크 구조를 사용하여 문제의 대상을 표현하는 확률모델이다[3, 4]. 변수는 노드로, 변수간의 의존관계는 원인으로부터 결과가 되는 변수로 방향을 가지는 링크로 표현한다.



[그림 1] 베이지안 네트워크 예

2.2 Kohonen Map

코호넨 맵은 비교사(Unsupervised) 학습 방법으로 스스로 n-차원의 입력 데이터들을 클러스터링하여 2차원으로 표현시켜준다. 다음은 SOM의 구조를 나타낸다.



[그림 2] SOM의 구조

그림에서 모든 입력 벡터 X는 출력 노드인 코호넨 층과 연결되어 있고 연결 가중치(weight)를 가진다. 초기의 연결 가중치는 일반적으로 0에서 1사이에 랜덤값으로 할당한 후 유클리드 거리를 이용하여 입력 벡터 X와의 유사성을 계산한다. 계산 결과 가장 가까운 거리를 가진 뉴런이 승자 뉴런이 된다. 승자 뉴런이 결정되고 난 뒤 학습 규칙에 따라 뉴런의 연결강도를 조정한다[5]. 코호넨 맵의 학습 규칙은 다음과 같다.

[Step 1] 연결 강도 W를 초기화한다.
(0에서 1사이의 랜덤값으로 설정)
[Step 2] 연결 강도를 변경시킬 범위(radius)를 설정하고 학습율 α을 결정한다.
[Step 3] 입력 벡터 X를 입력하여 유사도 D를 계산하고, D가 가장 작은 뉴런을 승자 뉴런으로 선정한다.

$$D = (W - X)^2 \quad (1)$$

[Step 4] 승자 뉴런으로부터 반경 r 범위 내에 있는 연결강도를 변경한다. k+1 학습 단계에서 연결 강도 W_{k+1}는 다음과 같다.

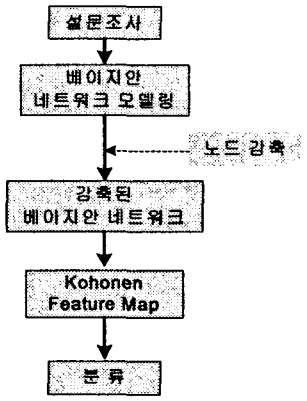
$$W_{k+1} = W_k + \alpha(X - W) \quad (2)$$

[Step 5] 규정된 반복 횟수만큼 학습이 진행되면 반경 r과 학습율 α을 감소시킨 다음 학습 과정을 반복한다.

3. 베이지안 네트워크를 이용한 노드 감축

3.1 전체 시스템 구성도

본 논문에서는 소비자의 구매 패턴 분석을 위한 전체 시스템 구성도를 [그림 3]과 같이 나타낸다.



[그림 3] 전체 시스템 구성도

3.2 노드 감축

노드 감축은 특정 변수에 대한 확률분포를 계산하려고할 때, 대상이 되는 변수를 X라 하고, 그 외의 변수는 n개 있으며 Y_i라고 한다. X와 Y는 각각 $x_1, .., x_m$, $y_{i1}, \ldots, y_{im}$의 m개의 상태값을 가지는 확률변수라고 가정하면 X의 사후

확률은 X 이외의 변수 Y_i에 대해서 있을 수 있는 모든 상태를 평균화하는 변화에 의해 식(3)과 같이 구해진다.

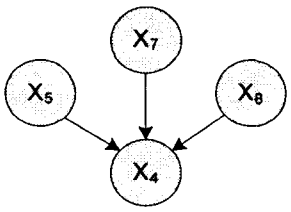
$$P(X=x_j)=\frac{P(x_j,Y_1,..,Y_n)}{\sum_{k=1}^{m}P(x_k,Y_1,..Y_n)}=\alpha P(x_j,Y_1,..Y_n) \quad (3)$$

$$P(x_j,Y_1,..Y_n)=\sum_{y^1\in Y_1}\cdots\sum_{y^n\in Y_n}P(x_j|Y_p)\Pi_{i=1}^{n}P(y^i|pa(Y_i)) \quad (4)$$

식 (3)에서 부모를 가지지 않는 최상위의 변수를 Y_T라고 하면 이 항은 다른 변수로부터 영향을 받지 않기 때문에 재귀적으로 전체 합의 바깥쪽으로 보낼 수 있으므로 식(3)은 다음과 같이 변형될 수 있다.

$$P(X=x_j)=\alpha P(Y)\sum_{y^1\in Y_1}\cdots\sum_{y^n\in Y_n}P(X|Y_p)\Pi_{Y_i\neq Y_T}P(y^i|pa(Y_i)) \quad (5)$$

식 (5)에 의해서 계산의 반복횟수를 줄일 수 있고 조건부 독립성을 이용하여 관련되지 않는 부분을 재귀적인 계산 밖으로 꺼낼 수 있다. 위와 같은 베이지안 네트워크의 성질을 이용하여 적은 계산 량으로 효율적으로 확률계산을 할 수가 있다. 구축된 베이지안 네트워크에서 불필요한 노드를 감축시켜 보면 [그림 4]와 같은 간단한 네트워크 구조로 표현할 수 있다.



[그림 4] 변수 제거법에 의한 네트워크 구조

위 그림에서 X_1, X_2, X_3은 각각 디지털 카메라의 선택기준(6가지), 원하는 가격대(5가지) 그리고 원하는 화소 수(5가지)를 나타내고, X_4는 디지털 카메라 제조 회사(6가지)를 나타낸다.

노드 감축으로 불필요한 노드를 제거한 후 구성된 베이지안 네트워크 구조를 바탕으로 비교사 학습 기법인 코호넨의 SOM을 이용하여 디지털 카메라의 구매 패턴을 분석하고자 한다.

4. 모의실험

디지털 카메라 이용자의 이용 및 구매 성향을 분석하기 위해 설문조사를 실시하였다. 12문항으로 구성한 설문조사를 바탕으로 디지털 카메라를 구입할 때 어떤 회사의 제품을 구매할 것인가를 파악하고자 한다. 설문조사의 데이터를 베이지안 네트워크를 이용하여 모델링 했으며 노드 감축을 통해 감축된 베이지안 네트워크로 모의실험을 했다. SOM의 입력 패턴 수는 3개 (선택기준, 가격대, 화소 수)이고, 클러스터 수는 6개(Cannon, Olympus, Kenox, Nikon, Sony, Etc)로 구성하였다. 그리고 뉴런 수는 2차원 상에 20*20 매트릭스로 구성하였다.

4.1 파라미터 설정

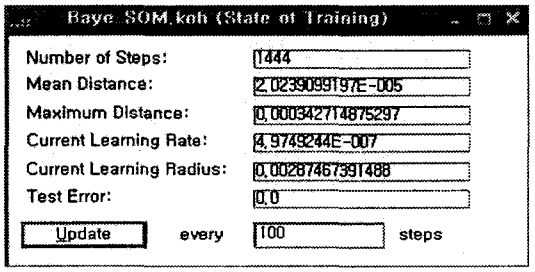
뉴런의 초기 가중치(initial weight)는 0에서 1 사이에 랜덤값으로 초기화하였고, 학습율에서 초기 학습율(initial learning rate)은 0.999, 학습율 팩타(learning rate factor)는 0.99로 설정하였다. 여기서 학습율 팩타는 네트워크의 학습율을 매 단계마다 증가시킨다.

학습 범위(learning radius)에서는 초기의 학습 범위(initial learning radius)는 4.0으로 학습 범위 팩타(learning radius factor)는 0.995로 설정하였고, presentation 순서는 random으로 하였다.

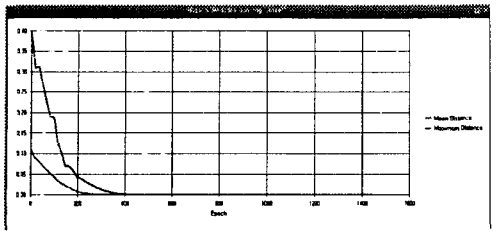
4.2 실험 결과

실험에 사용된 데이터는 150개의 데이터를 바탕으로 하였다. 실험결과 학습 횟수가 1,444회일 때 최적의 맵(map)을 얻을 수 있었다.

다음 그림들은 SOM의 최종 훈련 상태, 학습곡선, 가중치 변화량 그리고 SOM 맵을 나타낸다.



[그림 5] SOM의 최종 훈련 상태



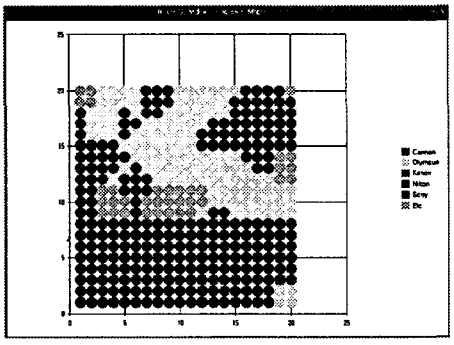
[그림 6] 학습 곡선

[그림 6]에 학습 곡선에서 'Mean Distance'와 'Maximum Distance'는 각각 모든 오브젝트들 간의 평균 거리와 최대 거리를 나타낸다.

Baye_SOM.koh (Neuron Weights)

		feature5 []	feature6 []	feature7 []
100	Olympus	2.500	3.000	2.000
101	Nikon	2.260	4.419	5.000
102	Kenox	2.017	4.021	5.000
103	Kenox	2.000	4.000	5.000
104	Kenox	1.996	4.000	4.996
105	Cannon	1.285	3.993	4.277
106	Cannon	1.000	4.000	4.000
107	Cannon	1.000	4.000	4.000
108	Cannon	1.000	3.998	4.000
109	Cannon	1.000	3.563	3.984
110	Sony	1.000	3.133	3.892
111	Sony	0.998	3.321	3.678
112	Olympus	0.904	3.435	3.390
113	Olympus	0.698	3.066	3.024
114	Olympus	0.924	3.000	3.000
115	Olympus	0.996	3.000	3.000
116	Olympus	0.996	3.000	2.990
117	Olympus	1.027	3.000	2.426
118	Olympus	1.343	3.000	2.017
119	Cannon	1.980	3.000	2.000

[그림 7] 가중치의 변화량

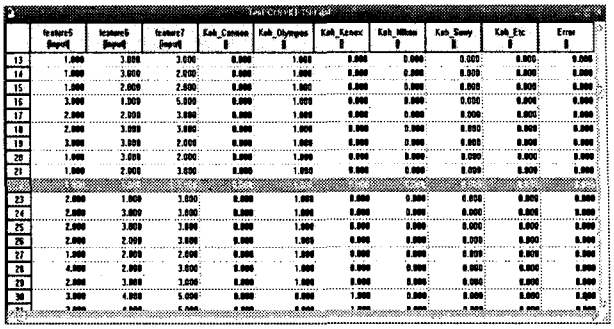


[그림 8] 최종 SOM 맵

실험 결과, 뉴런을 2차원 공간으로 20*20 매트릭스로 구성하고 '뉴런간의 학습 범위'를 4로 설정하였을 경우 6개의 클러스터가 잘 분류됨을 알 수 있었다.

만약 사용자로부터 '디지털 카메라의 선택기준은?', '원하는 가격대는?', '원하는 가격대에 화소수?'에 대한 질의에서 '기능적인 면', '30에서 40만원', '300에서 400만 화소'라는 응답을 얻었다면 'Olympus'라는 결과를 얻을 수 있다.

다음은 위 예에 대한 결과 값을 보여주는 그림이다.



[그림 9] 결과 화면

5. 결론 및 향후 연구과제

소비자의 구매 행동 패턴을 분석하기 위해서는 소비자의 구매에 따르는 심리 상태나 내부 상태까지 깊이 분석하도록 하여야 한다. 소비자의 구매 성향을 정확하게 파악하고 기업의 이익을 최대화하기 위해서는 마케팅 전략을 세우고 소비자의 구매에 따르는 행동을 조사할 필요가 있다. 설문 조사는 비교적 간단하고 대규모 조사도 실시하기가 용이하지만 문항이 많은 경우에는 응답자의 부담이 커서 응답을 회피하는 문제점과 어려운 문항에 대해서는 신뢰성이 떨어지는 문제점이 있다.

본 논문에서는 구매에 이르는 소비자의 구매 행동 패턴을 분석하기 위해 판매자의 노하우와 소비자의 구매의식 조사 데이터를 바탕으로 베이지안 네트워크를 구성한 후 불필요한 속성을 가진 데이터를 제거한 후 코호넨의 SOM을 이용하여 소비자의 구매 패턴을 분석하였다. 모의실험을 실행한 후 결과를 보면 감축 이후에도 좋은 분류 결과를 얻을 수 있음을 알 수 있었다.

6. 참고문헌

[1] 村上 外, ベイジアンネットワークによる消費者行動分析, 電子情報通信学会, 2004.

[2] 後藤秀夫, 市場調査ケーススタディ 改訂新版, 日本マーケティング教育センター, 1996.

[3] David Heckerman, A Tutorial on Learning Bayesian Networks, Technical Report MSR-TR-95-06, 1995.

[4] Thomas Dean et al, Artificial Intelligence Theory and Practice, Addison Wesley, 1995.

[5] Kohonen, T, Self-Organization Maps, Series in Information Science, vol. 30, Springer, Heidelberg. 1997.