제18회 한국정보처리학회 추계학술발표대회 논문집 제9권 제2호 (2002.11)

베이지안 학습법에 기초한 전자상거래에서의 고객 성향 분류 연구

전진호*, 이계성**
* **단국대학교 전자컴퓨터학부
e-mail : jhjy@cs.dankook.ac.kr

A Study on The Customer Classification of the EC based on Bayesian Learning Model

Jin-Ho Jeon*, Gye-Sung Lee**
*Dept of Computer Science and Electronics, Dan-Kook University

요 약

활성화되고 있는 전자상거래에 있어서 단순히 정해진 정보를 고객에게 제공하는 범위를 벗어나 고객의 특성에 따라 고객에 맞는 정보를 제공함으로서 매출 신장을 통하여 이윤확대를 꾀할 수 있다. 그러므로 본 연구에서는 베이지안 학습법을 이용하여 회원고객의 특성에 따른 분류화를 통하여 잠재적 구매 고객에 대한 구매 스타일을 예측하여 타겟광고가 가능한 기법에 대해 연구하였다.

1. 서론

전자상거래의 폭발적 확산에 따라 전체 시장규모는 날로 그 비중이 증가하고 있다. 국내의 경우에서도 인터넷 쇼핑몰은 99년에 급격히 증가하여 갈수록 경쟁이 심화되고 있는 실정이며, 점차 소비자의 욕구가 더욱 다양해짐에 따라 시장상황이 더욱 불투명하게 되었다.

이에 기업은 끊임없이 소비자의 욕구를 분석하여 미래의 수요를 파악함으로서 결국 누가 먼저 변화에 대응하고, 시장을 선점하느냐에 따라 기업의 명암이 달라질 수 밖에 없다.

그러나 전자상거래의 정보제공에 있어서 불특정 다수에게 고정된 광고를 제공하는 정적인 형태의 단방향성 무작위 정보 제공은 그 효과에 있어서 한계가 있다고 할수 있다.

그러므로 본 연구에서는 인터넷의 몇 가지 기본 특징인 첫째, 상호작용이 가능한 매체라는 점 즉, 인터넷은 근본적으로 정보의 제공자와 수혜자가 상호작용(Interaction)할 수 있도록 고안된 매체이다. 둘째, 고객간의 의사소통(Communication)이 용이하다는 것이고, 셋째, 개인화(Personalization)가 가능하다는 점에 기초로 하여 개개의 고객에 따라 타겟광고가 가능할 수 있도록 고객의 성향을 분류함을 목적으로 한다.

따라서 본 연구의 목적은 고객의 여러 특성들 즉, 성별, 나이, 관심분야 등을 파악하여 성격적인 성향에 따라 클래스를 나누고 기존의 데이터를 기반으로 새로운 사용자들의 성격적인 성향을 예측할 수 있으며, 또한 성격적인 성향에 따라 구매성향이 다르므로 이 고객들을 Bayesian Learning Model을 적용하여 분류 예측해 보고자 한다.

2.고객분류를 위한 Bayesian Learning Model 연구

분류는 지식발견 시스템에 있어서 클래스화 되지 않은 케이스의 집합들에 대해 클래스를 할당하는 작업이다.

본 연구에서는 Bayesian Learning Model을 적용하였다. 다음 식으로 표현이 가능하다.

$$P(A_k|B) = \frac{P(A_k \cap B)}{P(B)} = \frac{P(A_k)P(B|A_k)}{\sum_{i=1}^{n} P(A_i)P(B|A_i)} \quad (1)$$

위의 baye's therom 식 (1)에서 $P(A_i)$를 가설 A_i가 성립할 사전확률이라 한다. 즉, B가 조사되기

전의 확률이라 하고 $P(A_i|B)$를 B가 조사된 후에 수정된 확률 즉, 사후확률이라 한다.

Bayesian Learning Model은 감독 분류를 수행하는 방법으로 분류방법은 교육단계에서 Label된 클래스의 각각의 속성들에 대해 조건부 확률값을 계산함으로서 이루어지며, 새로운 데이터를 분류하는 과정은 교육단계에서 계산된 조건부 확률값들을 이용해 사후확률을 계산함으로서 진행된다.

Bayesian분류기는 클래스들의 집합 $c_j = C$ 과 속성들의 집합 $a_i = A$로 정의된다.

케이스에 대한 레이블이 주어져 있고 속성 값들이 있는 데이터베이스 D가 주어졌을 때, 교육은 D로부터 주어진 클래스에 대한 속성 값들의 조건부 확률을 계산함으로서 이루어진다. 클래스 c_j에 대해서 속성 a_{ik}가 나타날 확률값 $p(a_{ik}|c_j)$은 다음 식(2)으로 계산 가능하다.

$$p(a_{ik}|c_j) = \frac{n(a_{ik}|c_j)}{\sum_k n(a_{ik}|c_j)} \quad (2)$$

위의 식(2)을 이용하여 조건부 확률을 계산함으로서 교육을 마친 데이터베이스에 대해 두 번째 단계인 새로운 데이터들을 조건에 맞는 클래스로 분류하는 과정은 다음과 같다. 새로운 케이스에 대한 속성 값들의 집합 $(e_k = [A_1 = a_{1k}, \ldots, A_m = a_{mk}])$가 주어졌을 때 사후확률은 다음 식(3)과 같다.

$$p(c_i|e_k) = \frac{\Pi\, p(a_{ik}|c_j)p(c_j)}{\sum_{h=1}^{n} \Pi\, p(a_{ik}|c_h)p(c_h)} \quad (3)$$

Bayesian 분류기에서는 주어진 클래스들에 대한 속성 값들이 상호 독립적이라는 가정에 의해 다음과 같은 식(4)이 성립한다.

$$p(B|S_1, S_2) = \frac{p(S_1, S_2|B)\cdot p(B)}{p(S_1, S_2)} \quad (4)$$

그러므로, 위의 사후 확률값 $p(c_j|a_{1k})$를 구하려면 우선 다음 식(5)를

$$p(c_j|a_{1k}) = \frac{p(a_{1k}|c_j)p(c_j)}{\sum_{h=1}^{n} p(a_{1k}|c_h)p(c_h)} \quad (5)$$

계산한 후 계산 값을 이용해 그 다음 단계인 식(6)의 $p(c_j|a_{1k}, a_{2k})$ 값을 계산한다.

$$p(c_j|a_{1k}, a_{2k}) = \frac{p(a_{2k}|c_j)p(c_j|a_{1k})}{\sum_{h=1}^{n} p(a_{2k}|c_h)p(c_h|a_{1k})} \quad (6)$$

이러한 과정을 반복함으로서 사후 확률값 $p(c_j|e_k)$를 계산할 수 있다.

이와 같이 모든 클래스들에 대해 사후 확률값을 계산한 후 가장 큰 사후 확률값을 가지는 클래스로 주어진 데이터를 분류한다.[1]

위에서 언급한 베이지안 분류방법은 훈련과 분류의 대상이 되는 데이터가 완전하게 이루어져야 한다. 이것은 모든 데이터가 명확하게 알려져 있어야 함을 의미하며 데이터속에 모르는 데이터가 있을 경우에는 분류방법의 효율성이 감소하기 때문이다.[1)]

3. 고객자료의 수집과 분류

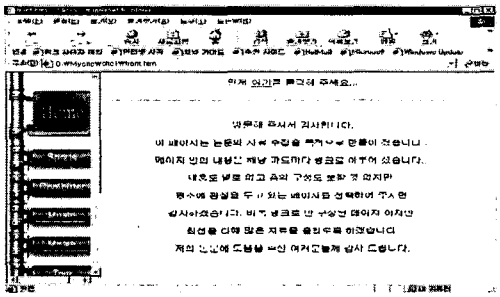
본 연구에서의 실험대상 자료수집은 내용이 스포츠, 컴퓨터, 음악, 영화, 여행, 건강, 종교에 관한 자료를 수록한 사이트의 접속을 통한 로그파일의 수집과 고객정보(사용자의 성별, 나이, 성격 등)의 입력을 병행하였다.[2][3] 대상으로는 정보수집의 용이함을 위하여 19세에서 28세까지의 A대 학생중 전산실에서 수업을 하는 학생들을 상대로 조사를 실시하였다.

위의 수집된 자료를 토대로 나이, 성별, 관심분야에 따른 사용자 부류(성격별)를 분류하고자 하였는데 성격은 사용자의 정보를 수집시 활동적, 차분함, 적극적, 소극적 중 2개 이내로 선택하게 하여 이를 통해 나뉘어진 8개의 성격으로 class를 분류하도록 하였다.

다음의 <그림1>은 고객정보와 로그파일을 얻기 위한 사이트의 메인화면이다.

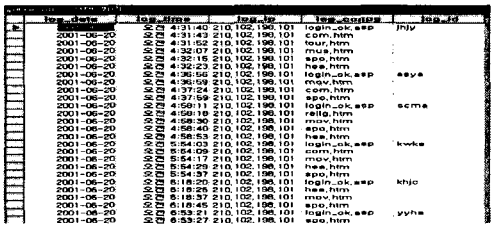
---

1) missing data가 있는 경우 정확한 조건부 확률값을 구하기 위해 각각의 missing entry에 대한 가능한 값들을 모두 고려하여 계산하여야 하며 그럴 경우 missing data의 수에 대한 지수승적으로 계산비용이 증가한다. 이러한 단점을 보완하는 방법으로 제시된 것은 특정한 가정을 통해 missing data값을 선택하거나, 훈련에 있어서는 missing data가 포함되는 것을 무시하는 것 등의 방법이 제시되었다.



<그림 1> 자료 수집을 위한 메인 사이트

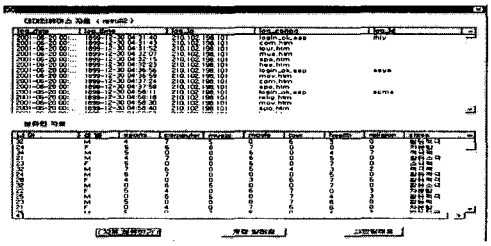
다음 <그림 2>는 사용자가 접속한 내용을 로그파일을 통해서 자료를 얻은 후, 사용자의 접속IP별로 정렬을 한 뒤 다시 시간별로 정렬을 하여 개개인의 접속자가 구분이 될 수 있도록 한 것이다.



<그림 2 로그파일의 정렬>

본 연구에서는 사용자가 가장 먼저 본 content에 대해서 가장 관심을 두는 분야라는 가정하에서 사이트에서의 항목 7가지에 의해 '7'에서 '1'까지 가중치를 주었다. 또한 회원정보에서의 성별, 나이 항목을 고객분류 분석을 수행하기 위한 각각의 케이스에 대한 속성값으로 사용하였고 회원정보를 통해 얻은 성격에 대한 정보를 레이블로 두어 각각의 케이스를 구성하였다.

다음 <그림3>은 로그파일에 대하여 가중치를 적용한 것을 나타내어 준다.

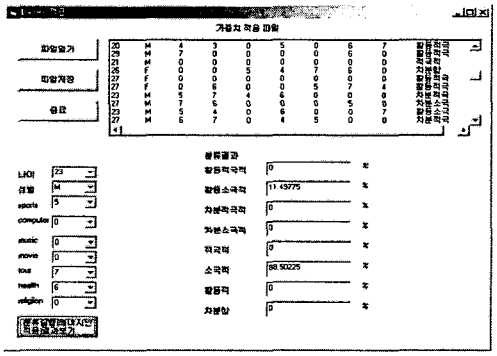


<그림 3 로그파일의 가중치 부여>

본 연구에서는 수집한 자료의 일부를 통해 학습을 한 후 새로운 데이터에 대해서 예측 분류를 하는 것이다.

여기서는 총 76개의 데이터로 구성되어 있으며 이 중 90%를 이용하여 학습을 한 후 무작위로 추출된 10%에 대해서 분류를 시험해 보았다. 이러한 과정을 10회에 걸쳐 반복을 통해 평균 예측율이 얼마나 되는지를 알아보았다.

다음 <그림4>는 파일열기하여 가중치가 적용된 파일을 통해 학습을 수행하여 학습한 기본적인 통계를 확인한 후 새로운 케이스에 대한 각각의 속성값들을 입력받아 창의 우하향에 분류결과를 제시하여 주는 것을 보여주고 있다.



<그림 4 베이지안에 의한 분류결과>

첫 번째 새로운 케이스들의 집합에서는 즉, 76개의 케이스에 대한 10%인 8개의 케이스에 대한 예측분류를 한 결과 8개중 2개는 옳은 예측 결과를 가져왔고 6개는 실패하여 25%의 적중률을 나타내었다.

아래의 [표1]은 10회에 걸친 classifying 예측결과를 나타내고 있다.

[표 1] classifying 예측 결과

1회	2회	3회	4회	5회	평균
25.0	12.5	37.5	37.5	25.0	
6회	7회	8회	9회	10회	23.75
12.5	25.0	25.0	12.5	25.0	

결과적으로 23.75%의 정확한 예측율을 가지고 있는 것을 알 수 있었다. 이것은 클래스를 8개로 나

누었을 때의 임의의 확률인 1/8의 확률보다 약 2배 많은 것이며, 광고효과를 200% 높인 결과를 가져온다고 할 수 있다.

4. 고객과 광고의 연관성 분석 및 결과

본 절에서는 사이트의 접속을 통해서 얻어진 자료를 토대로 각 class로 분류된 사용자들의 상품에 대한 지지도를 조사하여 독립성 검정을 통해서 각 class 마다 특정 상품에 대해 선호하고 있는지를 파악하였다. 광고에 대한 독립성 검정에서는 범주형 변수를 8개의 각 class와 각 class마다 선택한 11개의 제품군으로 하였고, 8×11의 제품 지지도 분할표는 [표 2]와 같다.

[표 2] 제품 지지도 분할표

	컴퓨터	영상/음향/통신	패션의류	패션잡화	코스메틱	스포츠/레저	티켓	게임	종반	도서	여행	계
활동적	6	4	4	0	0	4	3	0	1	0	0	22
차분한	6	3	1	2	0	3	4	2	7	1	1	30
적극적	2	1	0	0	0	0	0	0	1	0	0	4
소극적	3	0	2	0	1	0	2	1	0	0	1	10
활동적이고적극적	6	2	1	0	1	8	4	2	2	2	0	28
활동적이고소극적	3	3	0	0	0	1	0	0	0	0	0	7
차분하고적극적	3	0	0	0	0	0	2	0	2	0	0	7
차분하고소극적	3	0	0	0	0	0	1	3	3	0	3	13
계	32	13	8	2	2	16	16	8	16	3	5	121

[표 2]에서 상대적으로 적은 선택 율을 보여준 패션잡화와 코스메틱, 도서상품권을 소거하여 실험한 결과 5%의 유의 수준에서 각 클래스와 클래스가 선호하는 상품 군과는 서로 연관성이 있음이 증명되었으며, 모든 클래스가 컴퓨터 제품군에 대한 지지도가 높은 것을 알 수 있는데 이것은 표본으로 삼은 대상들이 대학생이라는 특수한 상황 때문이라고 사료된다. 컴퓨터 제품군에 대한 지지도를 제외하면 각 클래스마다 제품군에 대한 지지도가 특징적이라는 것을 알 수가 있다.

지금까지의 실험을 통해서 사용자의 나이, 성별, 관심분야에 따른 데이터를 가지고 사용자들의 성격적인 성향을 예측할 수 있으며, 또한 성격적인 성향에 따라 분류된 클래스들과 각 클래스들이 선호하는 제품군 사이에는 연관성이 있음을 밝혔다.

5. 결론

본 연구에서는 사용자마다 특성을 찾아 class화 할 수 있는 방법에 대해서 연구하고 각 class 마다 구매력을 갖는 제품에는 어떠한 것이 있는지를 파악하고자 하였으며 실험을 통하여 사용자의 나이, 성별, 관심분야에 따라 사용자를 성격적인 성향으로 나누고 이를 토대로 새로운 사용자에 대해서도 성공적인 분류를 수행할 수 있음을 밝혀 사용자들을 성격적이 성향에 따라 class화하는 것은 의미가 있음을 알아내었으며 그렇게 분류된 사용자 class들은 나름대로의 제품에 대한 구매성향이 있음을 파악하였다. 하지만 본 연구의 표본으로 한 대상이 대학생으로 한정적이었으며 또한 학습과 분류에 사용된 케이스들이 너무 적었기 때문에 좀 더 정확한 분석은 실시할 수 없었던 것이 아쉬움으로 남는다.

향 후 좀 더 폭 넓은 대상을 표본으로 한 연구가 이루어져야 할 것이며 지금까지의 실험결과를 기초로 인터넷 상에서 광고효과를 높이기 위한 방법으로 동적광고 process에 대한 연구가 좀 더 이루어진다면 전자상거래 분야에 있어서 개개인의 고객에게 차별화된 맞춤 광고를 제공하는데 도움이 될 것이다.

참고문헌

[1] 김경돈,"전자상거래 상에서 광고효과 제고을 위한 동적 Interface 방법에 관한 연구"., 단국대학교 석사학위논문,2000.

[2] 이화영.,"표준 로그파일을 이용한 웹마이닝에 관한 연구".,한국과학기술원.,2000.

[3] Cooley, R, Mobasher, B. and Srivastava, J.(1999), "Data preparation for mining world wide web browsing pattern", journal of knowledge and Information Systems, Vol. 1, No.1.