

베이저안 네트워크를 이용한 전자상거래 고객들의 성향 분석

Analysis of Web Customers Using Bayesian Belief Networks

양진산 · 장병탁

Jin-San Yang and Byoung-Tak Zhang

서울대학교 컴퓨터공학부

요 약

전자 상거래에서 고객의 성향을 이해하기 위해서는 일반적으로 판매 실무에서의 경험과 전문적인 지식을 필요로 하게 된다. 데이터 마이닝은 고객들에 대한 데이터의 분석을 통해서 이러한 성향들을 알아내는 것을 목표로 한다. 베이저안 네트워크는 DAG (Directed Acyclic Graph) 를 이용하여 데이터의 구조를 시각적으로 표현하여 주는 확률모형으로 변수사이의 종속관계를 밝히고 데이터 마이닝의 기법으로 이용할 수 있다. 본 논문에서는 베이저안 네트워크를 사용하여 전자 상거래 고객들의 성향을 분석하기 위한 방법을 제시한다. 또한 고객성향에 대한 주요 요인을 분석하기 위해 Discriminant 모형을 이용하고 그 유용성을 다른 방법들과 비교하였다.

ABSTRACT

Understanding customers tendency and preference in the web business usually requires expertise and experience of sales and marketing area in the company. Data mining can predict and reveal tendencies and properties of customers by analyzing data of customers. The Bayesian network is a graphical model using DAG (Directed Acyclic Graph) to express the structure of data. It can be used to explore the dependency relations among the properties of web customers as a data mining method. In this paper we used several feature selection methods including discriminant model to construct Bayesian networks and compared their results.

Key Words : 베이저안 네트워크, 전자 상거래, Discriminant 모형.

1. 서 론

베이저안 네트워크는 데이터에 내재하는 변수와 그들의 관계를 DAG(Directed Acyclic Graph)의 형태로 노드(node)와 호(arc)를 가지고 시각적으로 나타내어 주는 모델링 방법이다. 초기에는 데이터에 대한 경험과 전문적인 지식을 이용하여 베이저안 네트워크를 표현하였으나 컴퓨터의 발달로 점차 경험과 직관에 의존하지 않고 데이터로부터 직접 베이저안 네트워크를 구성하는 것이 가능하게 되었다. 데이터로부터 베이저안 네트워크를 구성 해주기 위해서는 적절한 모델의 선택과 국부적인 조건부 확률분포의 결정 등이 고려되어야 한다. 주어진 네트워크 구조 아래에서의 주변확률의 계산은 적절한 모델을 찾는 과정에서 특히 중요하게 된다. 특히 데이터의

일부가 숨겨져 있거나 결여되어 있는 경우, 직접적인 주변확률의 계산이 어렵기 때문에 EM (Expectation Maximization) 이나 MCMC (Markov Chain Monte Carlo) 등의 간접적인 방법을 사용하게 된다.

데이터의 분류나 예측에는 Decision Tree 나 신경망 모형 등이 있으나 이것으로부터 데이터의 특성이나 구조 등을 이해하는 것은 쉽지 않다. 베이저안 네트워크는 이와 같은 구조적인 특성을 나타내는 데에 적합한 모형으로서 데이터에 대한 이해를 통하여 데이터 마이닝에 활용될 수 있다. 베이저안 네트워크를 구성하기 위해서는 중요한 요인들을 분석하는 과정이 필요하게 된다. Discriminant모형은 범주형 결과치와 다른 변수와의 관계를 나타낸 모형으로 KDD 데이터의 전처리 방법으로 사용하였고 Decision Tree 와 Factor Analysis 에 의한 방법과 그 분석결과를 비교하였다.

2. KDD 웹데이터

KDD Cup 2000 데이터는 발에 관련된 용품들을 취급하는 회사(Gazelle.com)에서 일정기간(2000.1.30 ~ 2000.3.30)에 걸쳐서 그 회사를 인터넷상으로 접속한 방문자들의 실지기

접수일자 : 2000년 11월 18일

완료일자 : 2001년 01월 15일

본 연구는 정보통신부 대학기초과제와 BK21-IT 프로그램에 의해 지원 받았습니다.

록에 기초한 데이터들이다. KDD Cup 2000 데이터는 단순히 웹을 통하여 홈페이지를 방문한 경우(clickstream)와 웹을 이용하여 물건을 구입한 경우(purchase transactions)로 이루어져 있다. 적절한 Mining을 통한 웹 데이터의 분석결과는 앞으로의 영업방침이나 마케팅에서의 의사결정에서 비용과 시간을 절감하여 경제적이고 효율적인 경영에 이바지할 수 있다. 여기서는 웹을 이용하여 물건을 구입한 고객 중에서 일정금액(\$12) 이상을 구입한 고객들의 성향을 분석하는 것을 목표로 한다. KDD 데이터의 구성은 1782명에 대하여 나이, 성별, 수입, 자녀의 수, 웹을 방문한 시간, 횟수 등을 기록한 465 개의 항목으로 이루어져 있다. 이러한 변수들을 그대로 사용할 경우 결과의 해석에 있어서 불필요하게 복잡하고 계산상으로 많은 오차를 수반하므로 분석하기에 편하고 의미 있는 변수들로 바꾸어 주는 전처리 작업이 필요하게 된다. 여기서는 데이터의 전처리 방법으로 Discriminant 모델을 사용하여 6개의 요인을 사용하였고 그 결과를 Decision Tree에 의한 선택과 비교하였다. (표 1 참조)

3. 베이지안 네트워크

변수들간의 관계를 고려해서 데이터의 구조를 그래프로 나타내하고자 할 때는 DAG (Directed Acyclic Graph) 를 이용한 베이지안 네트워크를 이용할 수 있다. 베이지안 네트워크의 이점으로는 (1) 결측치가 많이 포함된 데이터를 자연스럽게 처리할 수 있고 (2) 성분들 간의 인과관계를 알려줌으로써 특정 조건 아래에서 결과를 예측할 수 있도록 하여준다. 또한 (3) 인과관계의 분석에서 선형적 지식(prior)을 충분히 활용할 수 있고 (4) 데이터를 나눌 필요가 없으므로 overfitting 을 줄일 수 있다. N 개의 case를 가진 데이터 D에서 n 개의 성분으로 구성된 case $X = \{X_1, \dots, X_n\}$ 에 대한 베이지안 네트워크는 X 의 각 성분들간의 종속적 조건들을 정해주는 네트워크 구조 S 와 각 성분들에 대한 주변확률 P 로 이루어져 있다. 각 노드를 변수로 하고 Pa(X)를 변수 X 의 부모 쪽 노드로 표시하면 데이터의 분포는 네트워크 구조 S 에 따른 주변확률들의 곱으로서 나타내어지게 된다.

$$P(X) = \prod_{i=1}^n P(X_i | Pa(X_i)) \quad (1)$$

먼저 X 가 단일 변수이고 r개의 범주 x^1, \dots, x^r 를 가질 경우 X 는 multinomial 분포를 따르게 되어 $X \sim MN(\theta_1, \dots, \theta_r)$ 이된다. 즉 $\theta_k = P(x^k | (\theta_1, \dots, \theta_r))$ 의 관계가 성립한다. 모수 $\theta = (\theta_1, \dots, \theta_r)$ 의 분포는 X 에 대한 conjugate prior 분포인 Dirichlet 분포로서 $\theta \sim Dir(\alpha_1, \dots, \alpha_r)$ 로 주어진다. (여기서 $E[\theta_i] = \frac{\alpha_i}{\sum_{i=1}^r \alpha_i}$

이고, α_i 는 prior 분포를 구성하기 위한 가상의 숫자로 볼수 있다.)
데이터 D에서 k 번째 범주의 숫자가 N_k 이면 새로운

case X_{N+1} 이 k 번째 범주에서 관측될 posterior 확률은

$$P(X_{N+1} = x^k | D) = \sum_{\theta} P(X = x^k | \theta) P(\theta | D) = \frac{\alpha_k + N_k}{\alpha + N} \quad (2)$$

이 된다. 또한 evidence (marginal likelihood; BDe-Score) $P(D)$ 는

$$P(D) = \frac{\Gamma(\alpha)}{\Gamma(\alpha + N)} \prod_{k=1}^r \frac{\Gamma(\alpha_k + N_k)}{\Gamma(\alpha_k)} \quad (3)$$

가 된다.[3]

case $X = \{X_1, \dots, X_n\}$ 가 여러 성분으로 이루어진 경우는 k, j 를 각각 변수 X_i 와 $Pa(X_i)$ 의 k 번째, j 번째 범주라 하고 그에 따른 모수를 $\theta_s = (\theta_1, \dots, \theta_n)$, $\theta_i = ((\theta_{ij})_{j,k}, k=1, \dots, r_i; j=1, \dots, q_i)$ 로 표시하면 각 (i-j)쌍의 모수 $\theta_{ij} = (\theta_{ij1}, \dots, \theta_{ijr_i})$ 의 성분 $(\theta_{ij})_{j,k} = \theta_{ijk}$ 는 $\theta_{ijk} = P(x_i^k | pa_i^j, \theta_i, S)$ 이 된다.

N_{ijk} 를 $Pa(X_i) = pa_i^j$ 일 때 $X_i = x^k$ 에 해당하는 (i-j)쌍의 D에서의 숫자라 할 때 $X_i \sim MN(\theta_{ij1}, \dots, \theta_{ijr_i})$ 이 된다. 각각의 case들이 서로 독립이면

$$P(\theta_s | D) = \prod_{i=1}^n \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \theta_{ijk}^{N_{ijk}} \quad (4)$$

이 되며 θ_{ij} 의 conjugate prior 에 대하여 $\theta_{ij} | D \sim Dir(\alpha_{ij1} + N_{ij1}, \dots, \alpha_{ijr_i} + N_{ijr_i})$ 이 된다. 또한 (i-j)쌍 간의 모수의 독립성을 가정하면 X_{N+1} 의 각 성분들이 특정한 범주 내에서 관측될 확률은

$$P(X_{N+1} = x_{N+1} | D, S) = E\left[\prod_{i=1}^n \theta_{ijk}\right] = \prod_{i=1}^n \frac{\alpha_{ijk} + N_{ijk}}{\alpha_{ij} + N_{ij}} \quad (5)$$

이 된다. 데이터에 결측치가 있는 경우에는 MCMC의 일종인 Gibbs Sampling 이나, EM 등의 방법을 이용하여 BDe-score가 계산되어진다. 이렇게 하여 얻어진 베이지안 네트워크는 데이터의 구조적인 특징을 보여주는 reasoning system으로 의사결정이나 예측 등을 할 수 있게 하여 준다.

4. KDD 데이터의 분석

KDD 2000 Cup 데이터는 전자상거래 고객에 관한 여러 가지 특성들을 측정한 데이터이다. Gazelle 회사에 따르면 2000.1.30 ~ 3.30 기간동안 2월 29일부터 TV 광고와 함께 Friends Promotion (친구를 소개하면 함께 혜택을 주는 것) 을 시작하였고 중간에 등록양식이 변경되어

설문의 내용이 바뀌었다고 한다. 그 회사 판매전문가에 의하면 그 동안 웹을 통해서 상품을 구입한 고객들의 특징들 중 알려진 몇 가지는 다음과 같다.

1. 할인된 용품을 구입하는 경우, 친구를 소개하여 함께 할인 받은 경우 현저하게 구매금액이 적었다. (그림 2)
2. 보통양말이나 운동용 양말을 자주 신는 경우, 친구나 친지들의 소개로 방문한 경우 비교적 적은 금액의 물건을 구매하는 경향이 있다.(그림 3)

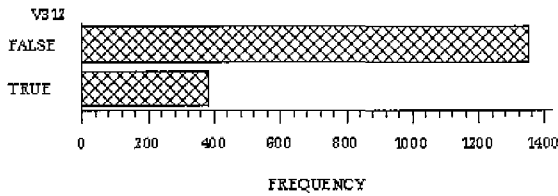


그림 1. \$12 이상 구매한 고객(True)의 비율
Fig. 1. The proportion of heavy spenders

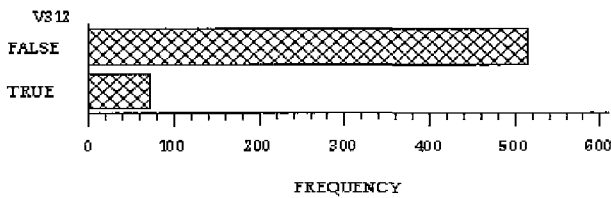


그림 2. 친구를 소개하고 할인 받아서 구매한 고객 중에서의 고액 구매고객 비율.
Fig. 2. The proportion of heavy spenders who used Friends Promotion.

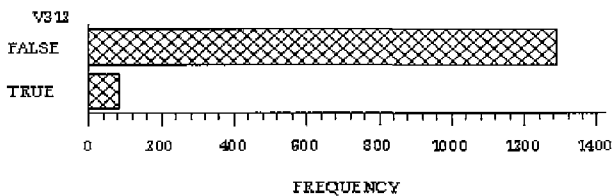


그림 3. 친척이나 친지의 소개로 회사를 알게된 고객 중에서의 고액 구매고객 비율.
Fig. 3. The proportion of heavy spenders who knows the company by friends and family.

4.1 베이저안 네트워크를 구성하는 중요 변수들의 탐색
이산형의 값을 가지는 Target 에 대한 변수들의 선택 방법으로서 Discriminant 모형을 이용하여 6개의 feature 들을 선택하여 범주형 변수들과 함께 노드로 사용하였다.

표 1. Discriminant 모형 과 Decision Tree에 의한 노드의 선택.

Table 1. Node selections by discriminant model and decision tree.

선택방법	노드
Discriminant 모형 + 범주형 변수 (Set A)	V240(Friend) V229(Order-Average) V304(Order shipping amt. min.) V368(Weight Average) V43(Home Market Value) V377(Num acount Template Views) + v11(Which DoYouWearMostFrequent) v13(SendEmail) v17(USState) v45(VehicleLifeStyle) v68(RetailActivity) v19(Date)
Decision-Tree (Set B)	V13(SendEmail) V234(OrderItemQuantitySum%HavingDiscountRange(5 . 10)) V237(OrderItemQuantitySum%HavingDiscountRange(10.)) V240(Friend) V243(OrderLineQuantitySum) V245(OrderLineQuantityMaximum) V304(OrderShippingAmtMin) V324(NumLegwearProductViews) V368(Weight Average) V374(NumMainTemplateViews) V412(NumReplenishableStockViews)
Decision Tree + Factor Analysis (Set C)	V368(Weight Average) V243(OrderLineQuantitySum) V245(OrderLineQuantityMaximum) $F1 = 0.94*V324 + 0.868*V374 + 0.898*V412$ $F2 = 0.829*V234 + 0.857*V240$ $F3 = -0.795*V237 + 0.778*V304$

그것과 비교하기 위하여 Decision Tree의 결과 중에서 가장 영향력이 있다고 보여지는 11개의 변수들을 선택하고, 여기에 Factor Analysis를 적용하여 새로운 변수 F1-F3 를 만들어서 각각에 대한 베이저안 네트워크를 찾아보았다. 각 방법들을 비교하여 보면 물건의 최소주문량 (V304) 와 친구의 소개에 의한 할인비율 (v240) 물건의 평균무게 (v368) 등이 공통으로 선택되었음을 알 수 있다. 또한 전반적인 할인비율을 나타내는 v229, v234, v237 등이 Decision Tree와 Discriminant 모형에서 나타남을 알 수 있다 (표 1). Decision Tree의 변수를

factor analysis 로 줄여준 경우 변수들이 대체적으로 View 에 관한 내용 (F1), 할인에 관한 내용 (F2), 구매력에 관한 내용 (F3) 으로 나누어진다.

4.2 베이저안 네트워크를 이용한 비교 및 분석

Set A (Discriminant 모델) 에 의한 베이저안 네트워크에서는 v240 (친구의 소개로 인한 구입률) 이 구매성향을 비롯하여 v11 (누구에게 듣고 구입하였는지), v368 (주문한 물건의 무게), v17 (거주지역), v43 (주택의 가격), v19 (방문 날짜), v13 (이메일 선호여부) 에 직접 영향을 주는 것으로 나타났다. Set B (Decision Tree) 의 경우, v304 (최소 주문량) 이 추가로 구매성향에 영향을 주고 v234 (할인률 5 ~ 10%의 비율) 이 v13 (이메일 선호여부), v237 (할인률 10%이상의 비율), v240 (친구의 소개로 인한 구입률), v243 (Order line 의 총합) 등에 영향을 준다. 위의 결과에서 v240 이나 v234 등이 v13, v19 등에 영

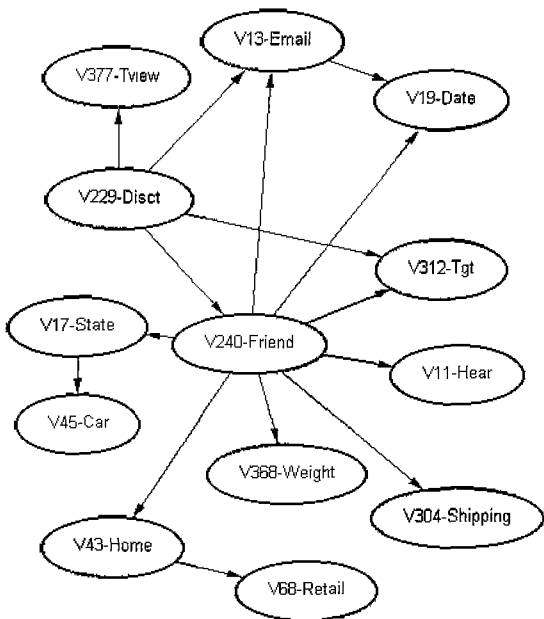


그림 4. Sset A로 찾은 네트워크
Fig. 4. Network from set A

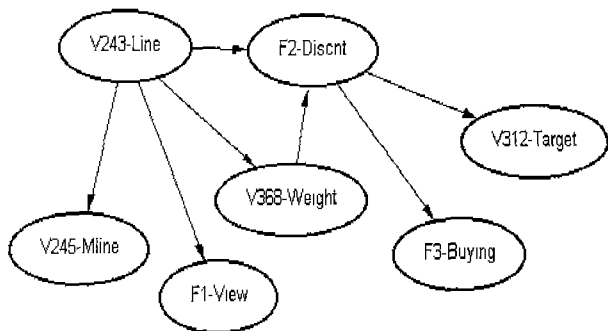


그림 5. Set C로 찾은 네트워크
Fig. 5. Network from set C

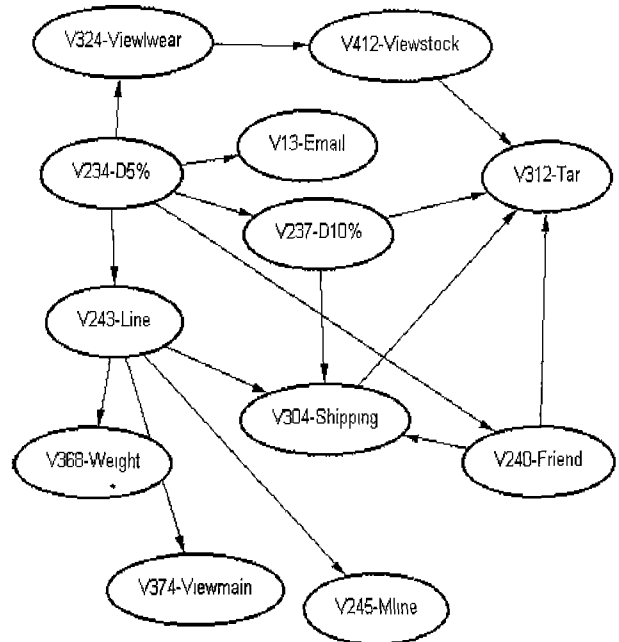


그림 6. Set B 로 찾은 네트워크.
Fig. 6. Network from set B.

향을 미치는 사실은 앞서 언급된 사실인 “Gazelle 회사에 따르면 2000.1.30 ~ 3.30 기간동안 2월 29일부터 TV 광고와 함께 Friends Promotion (친구를 소개하면 함께 혜택을 주는 것) 을 시작하였고” 에 부합한다. 세모형 모두 앞서 언급된 사항 1 (그림2참조) 을 반영하여 친구 소개에 의한 할인 (V240) 과 구매 물건의 할인을 (V229, v234, v237) 이 고객들의 구매성향 (v312)에 직접적인 영향을 미치는 요인임을 보여준다 (그림 4 - 그림 6). 데이터에 대한 네트워크의 likelihood 값은 네트워크의 구조와 데이터에 따라 다르며 비교 가능한 값들이 표 2 에 표시되었다.

표 2. 조건부 log likelihood 값의 비교
Table 2. Comparison of conditional log likelihoods

Node Set	likelihood forms	log likelihoods
Set A	$P(v312 v240)$	-516.844
Set B	$P(v312 v240)$	-516.777
Set A	$P(v312 v240, v229)$	-448.034
Set B	$P(v312 v240, v234)$	-488.953

조건부 확률분포를 이용하면 좀더 자세히 각각의 경우를 살펴볼 수 있다. Set A 에 의한 네트워크 (그림 4) 의 경우 v229 (평균 Order Discount Rate) 의 범위가 10% 미만이고 v240 (Friend Promotion Rate) 가 0% 일 경우 heavy spender 가 될 확률이 92.7% 가 된다. 반대로 v229 의 범위가 10% 이상이고 v240이 100% 인 경

우에는 4.6% 에 불과하다. 또한 v229 나 v240 중 하나의 값이 낮고 다른 쪽은 높은 경우에는 light spender 의 경우가 약간씩 높게 (0.631 , 0.589) 나왔다 (표 3). Set B 에 의한 네트워크 (그림 6) 에서는 v229 와 v240 의 값이 0% 인 경우에는 heavy spender 가 될 확률이 v304= 3.95, 14.95, 19.5 일 때 99.9% 와 87.0% 가 된다. 반면에 v229 나 v240 중의 하나의 값이 높을 경우에는 heavy spender 가 될 확률은 낮아진다 (표 4). v312 (Target) 와 다른 요인들간의 관계를 Set C 에 의한 네트워크 (그림 5) 에서 살펴보면 할인 정도를 나타내는 $F2 (= 0.829*v234 + 0.857*v240)$ 의 값이 증가하면 heavy spender 의 확률이 그에 따라 감소하는 것을 알 수 있다 (표 5). 즉 할인 요인의 값이 낮은 경우 ($F2 < 85.7$) 그에 따른 heavy spender 의 확률은 82.8%의 높은 값을 나타내는 반면 할인 요인의 값이 높아지면 4.5%까지 감소하게 된다.

표 3. Set A에서 고객에 대한 조건부 확률분포 표
Table 3. The conditional distribution of v312(Target) from networks of Set A

v312 : Target (customer)			
v229 : Order Discount rate	v240 : Friend Promotion rate	light spender	heavy spender
(0 10)	0	0.073	0.927
	50	0.980	0.020
	100	0.631	0.369
(10 50)	0	0.589	0.411
	50	0.500	0.500
	100	0.954	0.046

표 4. Set B에서 고객에 대한 조건부 확률분포 표
Table 4. The conditional distribution of v312 (Target) from networks of set B.

v312 : Target (customer)			
v237 : OrderItemQuantitySum%HaveDiscountRange (0 10).	v240: Friend Promotion rate	v304:OrderShippingAmtMin	heavy spender
0	0	14.95, 19.5	0.999
0	0	3.95	0.870
0	100	3.95	0.060
0	50	3.95	0.002
100	0	0	0

표 5. Set C 에서 고객에 대한 조건부 확률분포 표
Table 5. The conditional distribution of v312 (Target) from networks of set C

v312 : Target (customer)		
F2(Discount Score)	heavy spender	light spender
(0 85.7)	0.828	0.172
(85.7 168.6)	0.38	0.620
>168.6	0.045	0.955

베이지안 네트워크를 이용하면 v312 (customer) 의 특성 외에도 서로 다른 노드들간의 관계를 살펴볼 수가 있다. 즉 Set C 로 부터 만들어진 네트워크 (그림 5) 를 보면 F2 (할인 요인) 은 v243 (주문량) 과 v368 (무게량) 과 서로 연관되어져 있음을 나타내어 주는데, 조건부 확률 분포를 이용하면 좀더 자세히 그특성을 알 수 있다. 표 6에서 v243 (주문량) 이 많을 때 (2 이상) 낮은 할인요인 (85.7 이하) 이 될 확률은 v368 (무게) 에 비례한다. 반대로 v243 (주문량) 이 적을 경우에는 높은 할인요인에 대한 확률은 무게가 증가함에 따라 감소함을 나타내어 준다.

표 6. Set C 에서 F2에 대한 조건부 확률분포 표
Table 6. The conditional distribution of F2 from networks of Set C

		F2: 0.829*v234+ 0.857*v243	
v243	v368	Low (0 85.7)	High (85.7 168.6)
Low (1 2)	(0 1.3)	0.161	0.419
	(1.3 2)	0.055	0.117
	(2 3)	0.048	0.067
	(3 29)	0.171	0.019
High(2 53)	(0 1.3)	0.071	0.036
	(1.3 2)	0.150	0.006
	(2 3)	0.489	0.026
	(3 29)	0.615	0.032

5. 결론

베이지안 네트워크는 광범위한 데이터를 변수간의 관계에 따라 그래프로 표시함으로써 단순히 분류하거나 예측하는 데에서 간과될 수 있는 데이터의 특성을 이해할 수 있게 하여 준다. KDD 데이터의 경우 단순히 고객성향을 판단하고 예측하는 것 외에 각 요인들이 어떻게 관계되어서 서로간에 영향을 주는가에 대한 명확한 구조를 알 수 있게 하여 준다. 그렇지만 노드의 수가 큰 경우나 결측치를 가지는 경우 매우 많은 계산이 요구되므로 그에 따른 효율적인 알고리즘이 연구되어야 한다.

데이터의 분석에서 전처리 과정은 분석방법과는 별도로 분석결과에 영향을 미치는 중요한 작업이다. 본 논문에서는 Discriminant 모형을 사용함으로써 전처리과정에서의 탐색을 일반화 규칙화해서 보다 대용량의 데이터를 처리할 수 있는 방법으로 발전시켰다. Entropy나 비모수 함수 등을 이용하면 보다 효과적인 변수 탐색방법으로 개선될 것을 기대할 수 있다.

참고 문헌

- [1] Marco Ramoni and Paola Sebastiani. Learning Bayesian Networks from incomplete Databases, Technical Report, KMi-Tr-43, The Open University, 1997
- [2] G.F. Cooper and E. Herskovitz. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning* , 9:309-347, 1992
- [3] D. Heckerman. A tutorial on learning Bayesian networks, Technical Report MSR-TR-95-06, Microsoft Research, Redmond, Washington, 1995
- [4] N. Friedman, Iftach Nachman, Dana Peer. Learning Bayesian Network Structure from Massive Datasets: The Sparse Candidate Algorithm, UAI1999
- [5] M. Fayyad, G.P. Shapiro, P. Smyth and R. Uthurusamy, *Advances in Knowledge Discovery and Data Mining*, MIT Press, 1996
- [6] Michael I. Jordan, editor, *Learning in Graphical Models*, Kluwer Academic Publishers, 1999
- [7] M. Berthold , David J. Hand, *Intelligent Data Analysis - an Introduction*, Springer, 1999

저자 소개



장 병 탁 (Zhang Byoung-Tak)

1986년 서울대 컴퓨터공학부 학사
 1988년 서울대 컴퓨터공학부 석사
 1992년 독일 Bonn 대학교 컴퓨터공학 박사
 1992-1995 독일 정보기술 연구소 (GMD) 연구원

1995-1997 건국대학교 컴퓨터 공학과 조교수
 1997-현재 서울 대학교 컴퓨터 공학부 조교수
 관심분야: 인공지능, 기계학습, 신경망, 진화연산

Phone : 02 880-1833
 Fax : 02 883-3595
 E-mail : btzhang@cse.snu.ac.kr



양 진 산 (Yang Jinsan)

1981년 한양대 물리학부 학사
 1988년 미시시피주립대 통계학석사
 1995년 미시시피주립대 통계학박사
 2001년 현재 서울대학교 인공지능 연구소 (SCAI) 연구원

Phone : 02 880-7302
 E-Mail : jsyang@scai.snu.ac.kr