인공신경망과 귀납규칙기법을 이용한 제품별 예상 구매고객예측
Identifying prospective buyers for specific products using artificial neural network and induction rules

이건호*, 정수미**, 정병희***
*숭실대학교 산업정보시스템공학과 부교수, ghlee@ssu.ac.kr
**숭실대학교 일반대학원 산업정보시스템공학과, wjdtnal79@hotmail.com
***숭실대학교 산업정보시스템공학과 교수, bhchung@ssu.ac.kr

Abstract

It is effective and desirable for a proper customer relational management(CRM) to send an email of product sales' advertisement bills for the prospective customers rather than to send spam mails for non specific customers.
This study identifies the prospective customers with high probability to buy the specific products using Artificial Neural Network(ANN) and Induction Rule(IR) technique. We suggest an integrated model, IRANN of ANN and IR of decision tree program C5.0 and, also compare and analyze the accuracy of ANN, IR, and IRANN each other.
keywords : ANN, induction rule, C5.0

1. 서론

기업의 상품 광고에 있어서 단순한 스팸 형식의 대량 메일(Mass Mail) 또는 가격 할인 판촉 중심의 푸시 마케팅 전략만으로는 고객의 관심을 끌기에 역부족이며, Bain & Company의 조사결과에 따르면 크기나 수익면에서 가장 큰 시장은 푸시 마케팅을 선호하지 않는 경향이 있다[13].

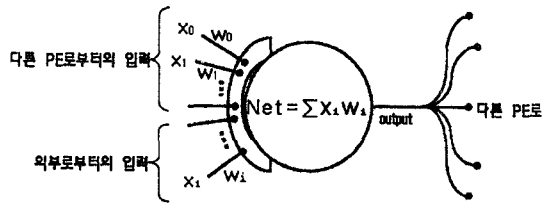
또한 구매량, 구매빈도 및 인구통계학적 변수 중심의 연구를 통한 마케팅은 고객을 단편적으로 이해하거나 잘못된 정보를 제공할 수 있다. 선별된 소수 고객을 중심으로 직접 메일(Direct Mail)을 발송하는 마케팅 전략은 무작위로 추출된 고객의 반응률에 비해 향상된 결과를 보인다[3]. 고객 선별 과정은 구매 경험이 있는 고객 DB를 데이터마이닝에 의해 분류하고 낮은 반응률이 예측되는 고객에게 발송되는 불필요한 비용을 절감하여 전체 수익을 증가시킬 수 있는 것이다.

본 연구에서는 기업의 특정 시즌 제품에 대한 고객의 특성을 예측하기 위하여 인공신경망 기법을 이용하여 예측 정확도가 우수한 모형을 선택하고, 예측에 대한 원인을 이해하기 위하여 해석이 용이한 장점을 가진 귀납 규칙 기법인 C5.0과 조합하여 분석했다. 본 논문에서는 고객특성 예측을 위한 인공신경망(ANN: Artificial Neural Network)과 귀납규칙(Induction Rule)의 조합한 모형(IRANN)을 이용하였다.

2. 관련 연구

2.1 인공신경망

신경망 시스템은 많은 양의 간단한 PE(Processing Element)를 사용하여 신경망 적으로 영향을 준 수학적 모델이다. PE는 layer로 이루어져 있고 한 layer에 있는 각 PE는 다음 layer에 있는 PE에 연결 강도(weight)를 가진다.[4] 다계층 인공신경망을 학습시키기 위한 대표적인 백프로퍼게이션 알고리즘은 $net = \sum_{i=0}^{n} x_i w_i$로 표현된다. 출력층에서 계산값이 구해질 때 사용되는 전이함수는 시그모이드 함수가 대표적이고, 구간 [0, 1]로 제약하기 위해 $f(net) = 1/(1+e^{-net})$으로 표현된다.



[그림 1] 인공적인 신경망의 PE 기본구조

인공신경망의 학습은 모형의 출력값이 목표값에 가깝게 연결강도를 조정하는 과정이며 다층신경망의 출력값이 구해지면 목표 값과의 오차를 구한다. 그 다음 이 오차가 최소화 되는 방향으로 각 층에서의 연결 강도를 조정하는 것이다. 오차 크기가 최소로 감소할 때 까지 반복함으로서 학습이 이뤄진다.[16]

2.2 귀납규칙 기법 C5.0

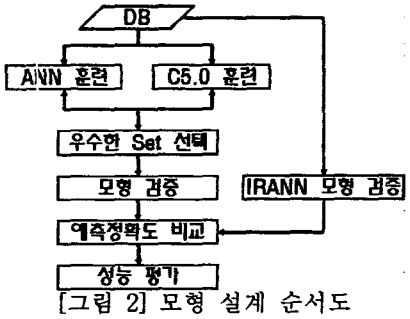
Ross Quinlan(1996)이 ID3에 이어 개발한 C4.5를 영국 ISL사의 데이터마이닝 툴인 Clementine에서 제일 먼저 채택하여 C4.5를 바탕으로 개발한 귀납적 학습방법이다. C5.0은 생성된 의사결정트리가 지나치게 많은 단계와 잎노드를 가질 경우에 학습된 의사결정트리의 일반화 능력을 제고하기 위하여 잎노드 제거 방법인 Pruning을 한다. 이 방법은 예측정확도를 향상시킬 수 있으나 오류율도 증가하므로 오류기반 Pruning으로 오류율의 증가를 통제한다. 또한 자동적으로 If-Then Rule을 생성해주는 규칙들의 집합형태인 명제형 지식(Propositional Knowledge)으로[10] 해석 및 분석이 용이하기 때문에 일반 사용자들도 쉽게 이해할 수 있다.

3. 연구 모형 및 방법

본 장에서는 특정 시즌 상품에 구매 경험이 있는 고객 데이터와 관련 데이터를 수집하고, 인공 신경망의 교사 학습 모형인 백 프로퍼게이션을 적용하여 구매 가능성이 높은 고객의 리스트를 선별하여 고객 특성을 예측한다. 찾아낸 특성에 따라서 전체 고객 중 선별된 고객에게만 메일링을 통한 추천이 이뤄지게 한다. 제품 선정은 봄 상품(Spring home-appliance : SPA) 2개와 여름상품(Summer home-appliance : SUA) 2개로 총 4개의 계절상품을 선택하였다.

3.1 구매 의도 예측을 위한 모형 설계

본 연구에서는 선정한 각 제품마다 고객들의 특성에 차이가 있고, 이 특성을 고려하여 제품을 추천을 할 경우 구매 가능성을 높일 수 있다는 가정 하에 진행되었다. 모형에 사용된 데이터는 설문조사하여 수집된 고객 데이터를 카테고리별로 작성하였다. 작성된 데이터는 인공 신경망을 학습시켜 이원 반응적 구매 예측률과 의사결정트리인 C5.0을 적용한 구매 예측률을 분석하고, 인공신경망과 C5.0을 조합한 IRANN모형을 적용시켜 분석하였다.



[그림 2] 모형 설계 순서도

3.1.1 데이터 수집

계절상품인 에어컨, 선풍기, 공기청정기, 가습기를 구매한 경험이 있는 약 1,200명을 대상으로 약27일 동안 설문조사하였고, 서울 및 경기도, 대구, 강원도 일부에서 오프라인 조사와 인터넷을 통한 온라인 조사를 하였다. 모형에 사용되는 표본의 크기는 총 1,000개의 설문 데이터이며, 각 사례에 포함되는 카테고리는 입력 요소로서 [표 1]과 같다.

모형을 구축하기 위해서는 훈련용, 테스트용, 검증용 3개의 데이터 셋을 나누어 분석하는데 전체 표본을 3개의 데이터 셋으로 나누는 비율의 대한 정확한 근거는 없다. 단, 연구 주제나 데이터의 형태 및 사용된 표본의 수를 고려하여 결정한다[7].

본 연구 인공 신경망 모형에서는 전체 표본 1,000개 가운데 각 셋을 6:2:2 비율로 랜덤하게 추출하여 각 셋을 구분하고, 4가지 제품의 구매 수를 동일하게 하여 학습률과 모멘텀은 0.1로 하였다. 한편 C5.0에서는 테스트용 Set을 이용할 필요가 없는 관계로 학습용 Set에 포함시켜 분석한다.

3.1.2 변수의 특성 및 선정

본 연구에서 입력 변수는 인구 통계학적 항목변수와 고객의식 항목변수를 사용하고, 출력 변수는 선정된 봄 상품인 공기청정기 구매, 가습기 구매와 여름 상품인 에어컨 구매, 선풍기 구매를 각각 이분형으로 출력하였다.

변수명	출력 변수
x_{spa}	구매한 봄 가전제품 구분
x_{sua}	구매한 여름 가전제품 구분

변수명	고객의식 항목 변수
XQ1 ~XQ8	8개의 의식항목변수는 정도의 범위로 척도

변수명	인구 통계학적 항목 변수
x_g	남자, 여자 구분
x_a	나이 표현
x_m	미혼, 기혼 구분
x_c	미취학자녀 양육여부
x_e	교육 수준을 구분
x_i	연간 소득 자료
x_j	9개의 직업으로 분류
x_h	자가, 전세, 월세 구분
rx_a	정수형의 연령 자료를 척도

[표 1] 변수의 선정

선정된 입력 변수들을 카이 제곱(Chi-square) 통계량을 이용하여 출력 변수에 대하여 유의한지를 검정하였다. 이를 위해 통계전문 툴인 SPSS 10.0 for window를 사용하였고, 검정 결과는 [표 2]와 같다.

검정 결과에서 p값이 0.01보다 작은 변수는 통계적 유의성이 있으므로 입력 변수로 사용한다. 그러나 p값이 0.01보다 작은 결과가 나왔음에도 불구하고 SPA의 변수Xj, XQ5와 SUA의 변수Xc, 변수Xj, 변수XQ2, 변수XQ4, 변수XQ6, 변수XQ8은 최소기대빈도 5보다 작은 빈도값을 나타냈기 때문에 유의수준이 정확하지 않을 수 있으므로 입력변수에서 제외하였다[9].

* : p < 0.01 수준에서 유의함.

변수명	SPA		
	Chi-square	df	p-value
x_g	104.636	1	0.000*
rx_a	4.035	4	0.401
x_m	4.201	1	0.040
x_c	1.594	1	0.207
x_e	23.772	3	0.000*
x_i	15.969	5	0.007*
x_j	46.007	8	0.000*
x_h	16.304	3	0.001*
XQ1	72.058	5	0.000*
XQ2	52.395	5	0.000*
XQ3	48.137	5	0.000*
XQ4	7.067	5	0.216
XQ5	19.502	5	0.002*
XQ6	38.707	5	0.000*
XQ7	54.690	5	0.000*
XQ8	30.935	5	0.000*

변수명	SUA		
	Chi-square	df	p-value
x_g	15.201	1	0.000*
rx_a	126.635	4	0.000*
x_m	114.038	1	0.000*
x_c	78.837	2	0.000*
x_e	22.169	3	0.000*
x_i	276.037	5	0.000*
x_j	86.202	8	0.000*
x_h	173.371	2	0.000*
xQ1	89.852	5	0.000*
xQ2	44.028	5	0.000*
xQ3	138.474	5	0.000*
xQ4	59.226	5	0.000*
xQ5	67.760	5	0.000*
xQ6	146.157	5	0.000*
xQ7	137.395	5	0.000*
xQ8	124.754	5	0.000

[표 2] Chi-square검정 결과

3.2 인공신경망 모형 구축

인공신경망 모형 구축 시 은닉층 수와 은닉층 노드 수의 결정은 휴리스틱한 지식에 의해 결정된다고 할 수 있다. 일반적으로 은닉층의 PE의 개수는 결과에 큰 영향을 주지 않는다고 한다[8]. 따라서 은닉층 수와 PE의 개수 결정은 어떻게 응용하는가에 따라 달라지고, 주관적일 수 있으므로 실험을 통한 확인이 타당하다. 만약 입력 자료가 특성 추출이 용이하지 않은 자료로 구성되어 있으면 고수준의 특성을 추출하기 위해서는 여러 개의 은닉층이 요구된다. 반면 어느 정도 고수준의 특성치를 나타내고 있으면 하나 또는 두개의 은닉층만 있어도 거의 모든 형태의 문제 해결 공간을 구성할 수 있다[12]. 은닉층 노드 수는 입력층 노드 수보다 두 배 이상이 되어서는 안 된다는 것이 일반적이고, 시작하기 적합한 수는 은닉층과 입력층의 노드 수를 같은 크기로 만드는 것이다[7]. 휴리스틱한 지식에 의한 방법으로 은닉층 노드 수를 다음과 같이 결정한다고 보고하고 있다[8].

$$h = 2n + 1 \quad (1)$$

$$h = 0.1S_n \quad (2)$$

S_n : 훈련 데이터 수

m : 출력층 노드 수

n : 입력층 노드 수

경험률에 의한 방법으로 은닉층 노드수를 학습 자료의 수에 비례하여 다음과 같이 사용하고 있다.[14]

$$h = \frac{S_n}{5 \times (m+n)} \quad (3)$$

또는,

$$h = \frac{2}{3}(m+n) \quad (4)$$

본 논문에서는 다층 퍼셉트론(MLP)과 백 프로퍼게이션 알고리즘으로 입력층과 출력층, 그리고 1개의 은닉층으로 이뤄진 모형을 사용하였다. 16개의 입력 필드 중에서 SPA에서는 이분형 1개와 범주형 3개, 정수형 6개로 총 10개의 입력노드를 사용하고, SUA에서는 이분형 2개, 범주형 3개, 정수형 7개로 총 12개의 입력노드를 사용한다.

은닉층 노드 수는 [표 3]과 같이 은닉층 노드 수에 따라 각 Set 모형이 구성된다. Set 1은 입력층 노드수와 같다는 제안에 따라 n개로 구성되고, Set 2는 $(n \times \frac{1}{2})$개 , Set 3은 $(n \times 2)$개, Set 4는 식(1)에 의한 구성, Set 5는 식(4)에 의해 구성된다. 휴리스틱에 의한 식과 경험률에 의해 제안된 식 가운데 각각 한 개씩 사용하기 위해 비교 실험하였고, 결과가 안 좋은 식(1)과 식(3)은 Set 구성에서 제외한다.

구분	Set 1	Set 2	Set 3	Set 4	Set 5
SPA	10	5	20	21	7
SUA	12	6	24	25	9

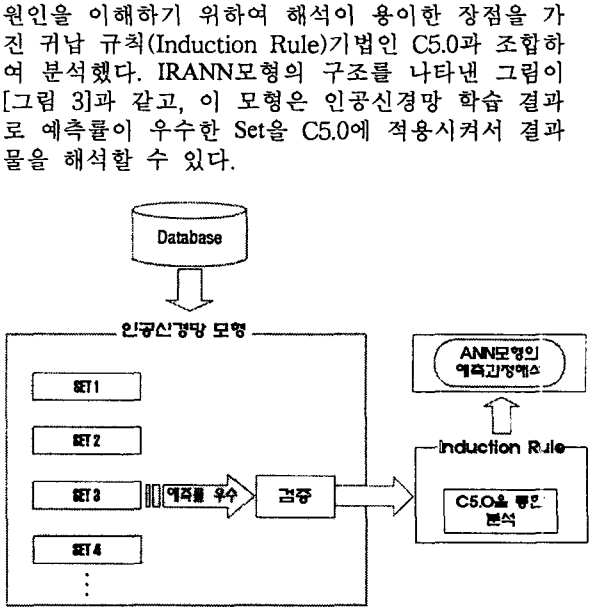
[표 3] 구성된 Set의 은닉층 노드수

3.3 C5.0 모형 구축

인공 신경망과 같은 표본 데이터를 사용하므로 동일한 변수를 사용하고, 테스트용 Set을 학습용 Set에 포함시켜 분석하고자한다. 본 연구에서는 SPSS Inc의 Clementine6.5를 이용하여, 심각도 잘라내기(Pruning Severity)는 75%, 가지 당 최소 레코드 2로 설정하였고, Use Boosting의 시도 수는 10으로 설정한다. 이 옵션의 사용은 C5.0의 예상 정확도를 보다 월등하게 제시할 수 있다.

3.4 IRANN 모형 구축

C5.0의 경우 결과를 쉽게 이해할 수 있는데 반해, 인공신경망은 예측 결과의 과정과 구체적인 중간 모형을 알기 어려운 단점을 지닌다. 쉽게 설명되지 않는 내부적인 작업을 수행하여 얻어진 결과물을 제공할 뿐 어떤 변수가 얼마나 중요한지, 어떻게 상호작용이 이뤄져서 그러한 결과물을 주게 되는지에 대한 해석이 없다[6]. 그래서 예측에 대한 원인을 이해하기 위하여 해석이 용이한 장점을 가진 귀납 규칙(Induction Rule)기법인 C5.0과 조합하여 분석했다. IRANN모형의 구조를 나타낸 그림이 [그림 3]과 같고, 이 모형은 인공신경망 학습 결과로 예측률이 우수한 Set을 C5.0에 적용시켜서 결과물을 해석할 수 있다.



[그림 3] IRANN모형 구조

4. 연구 결과 및 분석

각 모형 테스트는 제품에 따라 크게 두 부분으로 나눠 실행하였다. 봄 상품에 대한 고객 특성 예측과 여름상품에 대한 고객 특성 예측이 그것이며, 선정된 제품은 계절상품으로 다른 계절상품과의 결과 차이를 보이는지 알아보기 위한 것이다.

4.1 ANN(Artificial Neural Network)모형 분석결과

[표 4]의 ANN 모형 테스트에서 출력변수의 예측 정확도 차이는 표본 데이터 크기와 내용에 의한 것이라고 할 수 있고, 두 개의 출력변수의 예측 정확도 차이보다는 출력변수의 각 Set의 비교 분석을 위한 것이다.

(단위 %)

출력변수	Set	Set 1	Set 2	Set 3	Set 4	Set 5
X_{spa}	TRAIN	49.66	73.78	53.57	51.61	55.15
	TEST	53.19	38.78	44.26	45.10	44.44
X_{sua}	TRAIN	91.45	87.73	93.53	89.80	89.44
	TEST	84.31	85.42	84.91	97.67	83.61

[표 4] ANN 모형 간의 예측 정확도

봄 상품의 변수 Xspa에서 Set2는 훈련용에서 73.78%의 가장 우수한 예측률을 보였지만 테스트용에서 38.78%의 결과를 보이므로 훈련 셋을 외워버린 결과라고 보고 검증에서는 $(n\times 2)$개의 은닉층 노드수를 가진 Set 3이 적용되었다. 여름상품의 변수 Xsua는 가장 예측률이 좋은 Set 3이 적용되었다.

4.2 C5.0모형 분석결과

C5.0 모형 테스트는 보다 정확한 예측 결과를 위해서 Boosting 옵션을 10으로 사용하여 분석하였다. 또한 검증용 Set의 실험에서도 동일하게 사용하였다.

(단위 %)	Train		Evaluation	
	SPA	SUA	SPA	SUA
규칙 #1	94.2	94.5	83.0	93.0
규칙 #2	85.7	90.7	84.0	90.0
규칙 #3	87.0	91.7	78.0	83.0
규칙 #4	88.0	92.2	79.0	85.0
규칙 #5	86.5	93.2	72.0	92.0
규칙 #6	88.0	90.2	71.0	89.0
규칙 #7	87.7	90.5	84.0	74.0
규칙 #8	83.5	91.7	73.0	91.0
규칙 #9	87.0	89.2	72.0	86.0
규칙 #10	88.5	92.2	72.0	76.0

[표 5] C5.0의 예상 정확도

4.3 모형별 성과 비교

모형별 성과 실험은 IRANN모형의 결과와 ANN의 비교하기 위한 것이고, 그 결과 IRANN은 C5.0의 해석 능력뿐만 아니라 예측률에 있어서도 우수한 결과를 보이고 있다.

모형 / 출력변수	ANN	C5.0	IRANN
X_{spa}	71.74	84.00	94.00
X_{sua}	85.25	93.00	98.00

[표 6] 모형의 예측 정확도

5. 결론

인공신경망의 연구는 연속형 변수를 사용할 수 있는 장점을 부각시킨 신용도 문제나 재무관련 연구가 활발하였다. 본 연구는 마케팅 관련 연구에서 연속형 변수뿐만 아니라 범주형 변수 사용과 복잡한 영역에서도 훌륭한 결과를 도출할 수 있는 점을 이용해 인공신경망을 적용시켰고, 마케팅의 널리 사용되어지고 있지만 출력변수가 문자형일 때만 적용되는 단점과 해석이 용이한 장점을 가진 의사결정트리의 C5.0을 적절히 조합한 IRANN 모형을 통해 기업의 제품에 대한 고객특성을 예측해보았다. 그러나 타겟으로 하는 고객의 리스트에 접근이 어렵기 때문에 예측률에 대한 의문이 제기될 수 있다. 제품을 필요로 하는 잠재고객의 리스트를 사전에 정확히 선별해 낼 수만 있다면 그만큼의 개봉률을 극대화 시키는 효과적인 매체로 자리 잡을 수 있을 것이다.

참고 문헌

[1] 이건창,정남호,신경식, "신용카드 시장에서 데이터마이닝을 이용한 이탈고객 분석", 한국지능정보시스템학회논문지 제8권 2호 pp.15-35, 2002.
[2] 허준, "데이터마이닝에서 신경망 분석과 의사결정나무 분석의 비교", 수학.통계논문집 제6권, pp. 47-71, 1999.
[3] 백길호, "매스마케팅 한계 극복할 DB 마케팅", 제일기획마케팅연구소,1995.
[4] 정환묵, "지능정보시스템원론", 21세기사, 1999.
[6] 최인규, "나무구조 分析을 이용한 신경망 分析의 보완", 중앙대학교 석사논문, 1999.
[7] 최용석, http://home.pusan.ac.kr/~yschoi/Datamining
[8] 전응섭, "인공신경망을 이용한 소프트웨어 개발공수 예측모델에 관한 연구", 한국정보처리학회 논문지 제3권 제1호, 1996.
[9] 정영찬외, "SPSS 프로그램을 활용한 따라하는 통계분석", 크라운출판사, 2002.
[10] Quinlan, R., "Induction of Decision Tree", Machine Learning, Vol.1, pp81-98, 1996.
[11] R.P.Lipmann, "An Introduction to Computing With Neural Nets", IEEE ASSP Magazine, Vol.3, No.4, pp.4-22, 1988.
[12] Quinlan, J.R., and Quinlan, J., C4.5:Programs for Machine Learning, Morgan Kaufman Publishers, 1997.
[13] McKinsey & Media Metrix, Bain & Company, LG주간경제, 2000.
[14] Neuralware社의 NeuralWorks S/W.
[15] Jain, B.A. & N.B. Nag, "Performance Evaluation of Neural Network Decision Models", Journal of Consumer Research, Vol.9, pp179-180, 1982
[16] Agrawal, D. & C. Schorling, "Market share forecasting: An Empirical Comparison of Artificial Neural Networks and Multinomial Logit model", Journal of retailing, Vol.72, No.4,pp383-407, 1996