e-CRM에서 개인화 향상을 위한 의사결정나무 사용에 관한 연구

- Study on the Application of Decision Trees for Personalization based on e-CRM -

양 정 회 *
Yang Jeong Hoe
한 서 정 **
Han Seo Jeong

Abstract

Expectation and interest about e-CRM are rising for more efficient customer management in on-line including electronic commerce. The decision-making tree can be used usefully as the data mining technology for e-CRM. In this paper, the representative decision making techniques, CART, C4.5, CHAID analyzed the differences in personalization point of view with actuality customer data through an experiment. With these analysis data, it is proposed a new decision-making tree system that has big advantage in personalization techniques. Through new system, it can get following advantage. First, it can form superior model more qualitatively in personalization by adding individual's weight value. Second it can supply information personalized more to customer. Third, it can have high position about customer's loyalty than other site of similar types of business. Fourth, it can reduce expense that cost marketing and decision-making. Fifth, it becomes possible that know that customer through smooth communication with customer who use personalized service wants and make from goods or service's quality to more worth thing.

Key-word : e-CRM, Decision-making Tree, personalization

* 인덕대학 산업시스템경영과
** 호서대학교 디지털비즈니스학부

1. 서 론

오늘날 기업들은 여러 해 동안 기간업무시스템(Line of Business), ERP(Enterprise Resource Planning) 및 다양한 운영시스템을 통해서 대량의 데이터베이스를 축적해 오고 있다. 인터넷의 확산과 함께 급속도로 확장되고 있는 전자상거래의 물결은 기업에게 시장의 변화에 빠르게 대처하는 능력을 요구하고 있으며, 이러한 변화는 OLAP(Online Analytical Processing)을 기반으로 한 분석시스템 및 데이터 마이닝에 대한 기업의 관심을 고조시키고 있다. 기업에 있어 다년간의 기업활동의 결과로 축적된 데이터베이스에서 일정한 패턴과 규칙을 찾아낼 수만 있다면 고객과 시장을 이해하는 데에 큰 도움이 될 뿐 아니라 그들의 마케팅 및 판매, 고객지원을 포함한 기업 전반의 활동을 개선하고 그 결과를 예측하여 효과를 극대화시킬 수 있는 기회를 얻을 수 있기 때문이다. 이를 위해서는 정확한 분류와 예측을 위한 데이터 마이닝 기술이 주요한 관건이지만 그 기대와 관심의 고조에 비해 적재적소에 바람직하게 쓰이고 있는가에 대한 평가와 근거 제시는 미흡한 실정이다.

본 논문에서는 예측과 분류를 위한 강력한 데이터 마이닝 도구인 의사결정나무의 주요 알고리즘들인 CHAID, CART, C4.5에 대해 전자상거래와 개인화 관점에서 유효성을 분석하고 실제 데이터 셋을 이용하여 보다 효율적인 개인화 의사결정나무를 제안하고자 한다.

2. 의사결정나무 알고리즘

세 가지 알고리즘을 비교하기 위해 사용된 데이터는 골프를 칠 수 있는가에 대한 날씨 데이터로 목표변수는 Play에 대한 Yes와 No이며, 설명변수는 <표 1>과, 실제 training data는 <표 2>와 같다.

<표 1> 실험 데이터의 설명변수

ATTRIBUTE	POSSIBLE VALUES
outlook	sunny, overcast, rainy
temperature	continuous
humidity	continuous
windy	true, false

<표 2> 실험 데이터

OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
sunny	85	85	false	No
sunny	80	90	true	No
overcast	83	78	false	Yes
rainy	70	96	false	Yes
rainy	68	80	false	Yes
rainy	65	70	true	No
overcast	64	65	true	Yes
sunny	72	95	false	No
sunny	69	70	false	Yes
rainy	75	80	false	Yes
sunny	75	70	true	Yes
overcast	72	90	true	Yes
overcast	81	75	false	Yes
rainy	71	80	true	No

2.1 CHAID (Chi-squard Automatic Interaction Detection)

CHAID는 1975년 J. A. Hartigan에 의해 소개된 알고리즘으로 통계 package에서는 가장 보편적인 프로그램이다. 이 알고리즘은 AID(Automatic Interaction Detection system)에 기원을 두고 있다. AID란 두 변수간의 통계적 관계를 찾는 것인데 의사결정나무 형성을 위해 이 알고리즘을 사용한다. CHIAD는 카이제곱-검정(이산형 목표변수) 또는 F-검정(연속형 목표변수)을 이용하여 다지 분리(multiway split)를 수행하는 알고리즘이다. 카이제곱 분포표는 r×c 분할표라고도 하며 다음과 같다.

<표 3> r×c 분할표

목표 / 설명	목표변수1	목표변수2	...	목표변수c	합계
설명변수1	f_{11}	f_{12}	...	f_{1c}	$f_{1.}$
설명변수2	f_{21}	f_{22}	...	f_{2c}	$f_{2.}$
...	...	...	...	...	...
설명변수r	f_{r1}	f_{r2}	...	f_{rc}	f_{11}
합계	$f_{.1}$	$f_{.2}$	...	$f_{.c}$	$f_{..}$

자유도(degree of freedom) df = (r-1)(c-1)

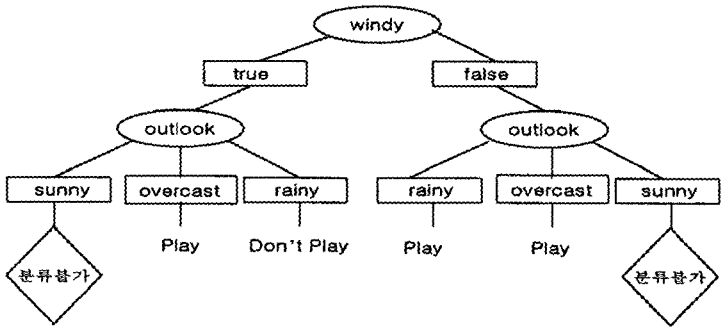
설명변수 i 가 목표변수 j 범주에 속하는 빈도수 f_{ij}

e_{ij} : 분포의 동일성 또는 독립성의 가설하의 기대도수 $e_{ij}=\frac{f_{i.}\times f_{.j}}{f_{..}}$

CHAID는 $P=\frac{df}{x^2}$ 값이 작은 변수부터 선택하여 자식마디를 생성하게 된다. 목표변수가 이산형일 때에는 Pearson의 χ^2 통계량을 $x^2=\sum_{i,j}\frac{(f_{ij}-e_{ij})^2}{e_{ij}}$, 목표변수가 순서형이나 그룹화 된 연속형일 때에는 우도비 χ^2 통계량 $x^2=2\sum_{i,j}f_{i,j}\times\log_e\left(\frac{f_{ij}}{e_{ij}}\right)$을 쓰게 된다. CHIAD는 각 설명변수의 범주들이 자료를 반응변수의 각 범주들로 구분하는 판별력의 크기에 따라 설명변수의 범주들을 이용하여 나무구조를 만드는 분석방법으로 전체 자료를 둘 이상의 하위노드(child node)로 반복적으로 분할한다. 이 과정에서 설명변수의 범주의 쌍에 대한 반응변수의 유의한 차이가 없으면 설명변수의 범주들을 병합하며, 유의적이지 않은 쌍들이 없을 때까지 과정을 계속한다. 각 설명변수에 대한 최고의 분할을 찾고, 모든 설명변수에 대한 유의성을 조사하여 가장 유의적인 설명변수를 선택한다. 선택된 설명변수의 범주들의 그룹을 사용해 자료를 상호 배반인 부분집합으로 분할하며 각 부분집합에서 정지규칙중의 하나가 만족될 때까지 이 과정을 독립적으로 순환, 반복한다.

2.2 CART

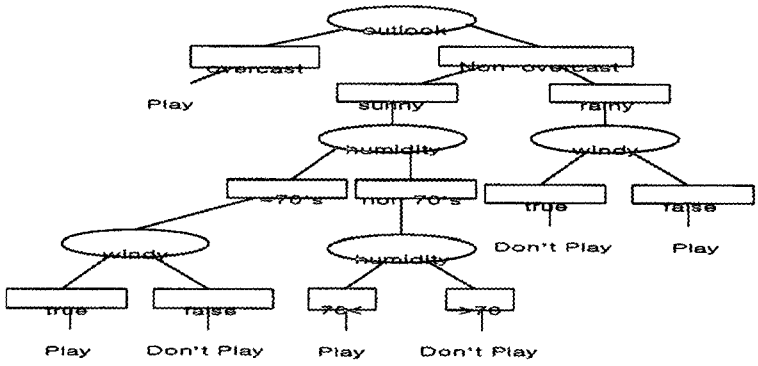
CART알고리즘은 의사결정나무분석을 형성하는데 있어서 가장 보편적인 알고리즘이라고 할 수 있다.[4] 1984년 L. Brieman에 의해 발표되어 machine-learning 실험의 시초가 되고 있다. 모형의 형성은 training data set을 가지고 한다. 목표변수는 이미 그 분류가 알려져 있으며, 우리는 나머지 설명변수를 가지고 이 목표변수를 잘 분류할 수 있는 모형을 만들어 새로운 데이터 세트에 적용시킨다.



<그림 1> CHAID에 의해 생성된 의사결정나무

CART알고리즘은 이진트리구조로 모형을 형성하는데 첫 번째 과제는 목표변수를 가장 잘 분리하는 설명변수와 그 분리시점을 찾는 것이다. 이 측도의 하나를 다양성(diversity)라고 하는데, 노드의 다양성을 가장 많이 줄이는 설명변수를 선택한다. 분리기준은 diversity (before split) - (diversity (left child) + diversity (right child))를 크게 하는 곳을 분리 기준을 정한다.

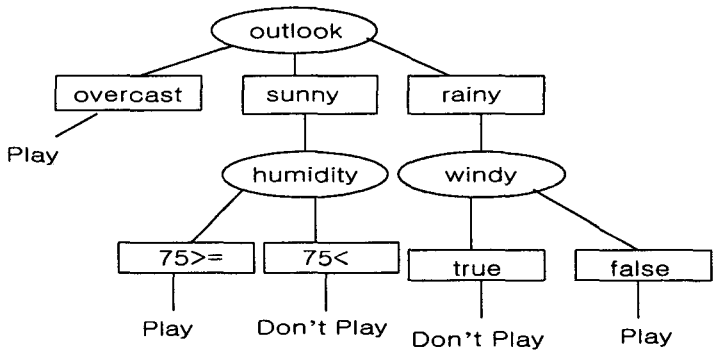
Gini Index(G)는 n개의 원소 중 임의로 두 개를 추출하였을 때, 추출된 두 개가 서로 다른 그룹에 속해 있을 확률로 각 마디에서의 impurity나 diversity를 재는 측도로 사용한다. P(i)는 각 마디에서 한 개체가 목표변수의 I번째 범주에 속할 확률이고, P(i)P(j) : i번째 변수에서 추출된 임의의 한 개체를 j번째 범주에 속한다고 잘 못 분류할 확률이라 할 때, $G=\sum_{j=1}^{c} P(j)(1-P(j))$이다. CART의 기본 아이디어는 Gini Index를 가장 감소시켜주는 예측변수와 그 변수의 최적분리를 Child Node로 선택한다. Gini Index의 감소량은 n을 부모노드의 관측치 수라하고, n_R , n_L 을 자식 노드의 관측치 수라 할 때 $\Delta G = G-\left(\frac{n_L}{n} G_L + \frac{n_R}{n} G_R\right)$ 이다. CART에 의해 생성된 의사결정나무는 <그림 2>와 같다.



<그림 2> CART에 의해 생성된 의사결정나무

2.3 C4.5

C4.5는 J. Ross Quinlan에 의해 정립된 의사결정나무 알고리즘으로 machine learning 분야의 ID3 알고리즘과 유사하다. C4.5가 CART와 다른 점은 CART는 이진분리를 하지만 C4.5는 가지의 수를 다양화 할 수 있다는 것이다. 이 알고리즘은 연속변수에 대해서는 CART와 비슷한 방법을 사용하지만 범주형에서는 다른 방법을 사용하게 된다. 가지치기 방법도 CART와는 다르게 training dataset과 멀리 떨어져있는 데이터에 대해서는 언급하지 않고 가지치기를 할 때에 training dataset과 같은 데이터를 적용한다.
Gain이 가장 큰 속성이 첫 번째 노드의 query가 된다. Gain은 Gain(X,T) = Info(T) - Info(X,T)과 같이 정의할 수 있다. C4.5에 의해 생성된 의사결정나무는 <그림 3>과 같다.



3. 알고리즘의 유효성 분석

3.1 전자상거래 관점에서의 비교

3.1.1 모형의 정확성

단일변수의 가지 분류에서 분류 규칙은 선택된 변수와 그 시점에 의해 특징지어진다. 그러나 CART의 탐색 기법은 많은 값을 지닌 목표변수 쪽으로 심하게 기울어지는 경향을 나타내고 있다. 이에 반해 C4.5는 그러한 치우침이 거의 없다. 또한 의사결정나무가 연속형 변수를 비연속형 변수로 취급하여 발생할 수 있는 예측오류가 상대적으로 낮다. 반면 나무를 통해 생성된 규칙은 CART와 C4.5가 거의 비슷한 결론을 보여주고

있는데 CHAID의 경우에는 전혀 다른 규칙이 생성되고 있다. 이것은 CART와 C4.5는 모두 데이터의 불순도를 측정하여 나무의 자식 노드를 결정하고 CHAID는 두 변수 간의 연관도를 측정하여 자식 노드를 결정하기 때문이라고 볼 수 있다. 결국 규칙의 생성에 가장 많은 영향을 준 변수가 무엇인가를 찾는 문제에 있어서 CHAID와 CART의 방법은 비교적 쉽게 답을 얻을 수 있지만 C4.5에서는 비트 개념을 이용하여 엔트로피 지수를 수리적으로 보다 더 정교하게 나타낸다.

3.1.2 나무의 간결성

CHAID와 C4.5에 의해 생성된 의사결정나무의 깊이는 각각 3단계이 CART에 의한 나무의 깊이는 4이다. 또한 말단 노드의 수는 C4.5가 5개, CART가 7개, CHAID의 경우 8개 노드 이상의 가능성을 가지고 있다. 더욱이 CART와 같은 이진 분리 알고리즘은 너무 많은 가지가 발생하게 되고 error-prone의 우려가 있다. 이러한 현상은 데이터 레코드의 수가 많아질수록 더욱 심화되게 되는데 가지치기를 통해 유의한 수준으로 모형을 형성하게 된다. 그러나 가지치기 시, 분리기준 설정 시 각각의 경우를 다 고려해야 되므로 각 조합의 경우를 모두 고려할 경우 그리고 가지치기를 할 경우 상당한 컴퓨팅이 제공되어야 한다.

3.1.3 다른 분야에의 확장성

의사결정나무는 명목형 변수를 위주로 분류하게 되어 있다. 연속형 변수의 경우 명목형 변수와 같이 변형하여 명목형 변수에서 사용하는 분류규칙을 적용하도록 하고 있다. 그러나 전자상거래 분야를 비롯한 실세계의 데이터는 수많은 연속형 변수에 의해 이루어져 있다. 따라서 이를 명목형으로 변환하기 위한 컴퓨팅 비용이 소모되는데 그런 점에서는 C4.5보다 계산이 용이한 CART가 더 효율적이다. 그러나 C4.5를 언급할 때 따라다니는 지수는 엔트로피(Entropy)라는 개념을 고려해야 한다. CART와 마찬가지로 C4.5도 마디의 순수함을 재는데 비트(bit)개념을 이용한다. 비트 개념을 다음의 예로 이해할 수 있다. 만약 8개의 카테고리로 이루어진 마디가 있다면 그 각각의 카테고리를 표시할 비트는 총 log28 = 3 개가 필요하게 된다. 그 마디에 가지가 쳐져서 4개의 카테고리가 있는 자식마디가 생기면 그 자식마디는 log24 = 2 개의 비트만 있으면 표현이 가능하다. 결국 가지가 쳐지면서 3개가 필요했던 비트가 2개로 줄어들어 엔트로피 1의 이득을 보게 된다. 이러한 수리적인 계산에 의해 C4.5의 계산 비용을 줄일 수 있다.

3.2 개인화 관점에서의 비교

3.2.1 개인화 용도로의 사용적절성

사용자들의 측면에서 개인화를 원하는 첫 번째 이유는 시간의 절약이다. 의사결정나무는 누구나 쉽게 이해할 수 있는 모형으로써 전자상거래 사이트의 사용자들로 하여금 자신이 원하는 내용을 복잡한 과정을 거치지 않고 얻을 수 있도록 해준다. 두 번째 이유는 개인화가 사용자들에게 자신만의 선택을 가능하게 해 주는 데 있다. 사용자들은 그들의 관심에 맞는 내용을 얻기를 원할 뿐 아니라 자신이 결정할 수 있는 부분을 갖게 되기를 원한다. 의사결정나무는 웹서비스 제공자들뿐만 개인 사용자들에게도 쉽게 자신의 관심과 선호를 밝힐 수 있도록 돕는다. 개인화의 세 번째 장점은 개인화 된 서비스를 받을 수 있다는 점이다. 개인별로 생성된 의사결정나무는 고객 개인에게 알맞은 서비스를 전달할 수 있다.
그러나 의사결정나무는 누적된 데이터를 통해 모형을 형성하여 규칙을 결정하는 것이므로 한 번 결정된 규칙은 변경하기 어렵다는 특징을 가지고 있다. 개인화를 위해서는 이러한 나무모형의 유연성 부족이 문제가 된다. 의사결정나무는 기본적으로 DW의 데이터를 이용해 구성되지만 수시로 들어오는 개개인의 입력 값에 의해 달라질 수 있는 모형이 필요하다. 이렇게 반복적으로 나무를 형성하기 위해서는 비용이 적게 들고 보다 타당성이 규칙을 생성하는 방법을 이용해야 한다. CART의 경우 계산 비용은 적게 들고 C4.5의 경우 가장 간결한 모형을 제시해주고 있다. 이러한 각각의 장점을 병합하면 보다 효율적인 개인화 의사결정나무의 생성이 가능하다.

3.2.2 e-CRM 분야로의 적용용이성

의사결정나무 분석은 예측과 분류를 위해 보편적이고 강력한 도구로써 신경망 구조 분석과는 달리 나무구조로 규칙을 표현하기 때문에 이해하기가 쉽다. 어떤 적용에서는 얼마나 잘 분류하거나 예측하느냐 만이 문제화되기도 한다. 즉, DM 발송회사는 모델이 어떻게 구성되었는지 보다는 얼마나 자신의 메일에 잘 대답을 해줄 수 있는 집단을 분류해줄 수 있는지에 관심을 가지고 있다. 하지만, 어떤 경우에는 왜 이런 결정을 하게 되었는지 설명하는 것도 중요하며 의사결정나무 분석은 이러한 경우에 유용하다. 예를 들면, 카드신청자의 카드 발급을 거절해야 하는 경우 그것의 결과를 설명할 수 없는 신경망 구조 분석보다 이유를 설명해 줄 수 있는 의사결정나무 분석이 더 유용하다. 따라서 고객세분화를 통한 개인화 서비스 분야와 같은 곳에 쓰기에 적절하다.
또한 의사결정나무에 의해 내려지는 결과는 이산적인 모습으로 나타나는 극단적인 형태를 보여주므로 모형의 안정성에 있어서도 기존의 통계적 모형보다 우수하다고 할 수 있다.

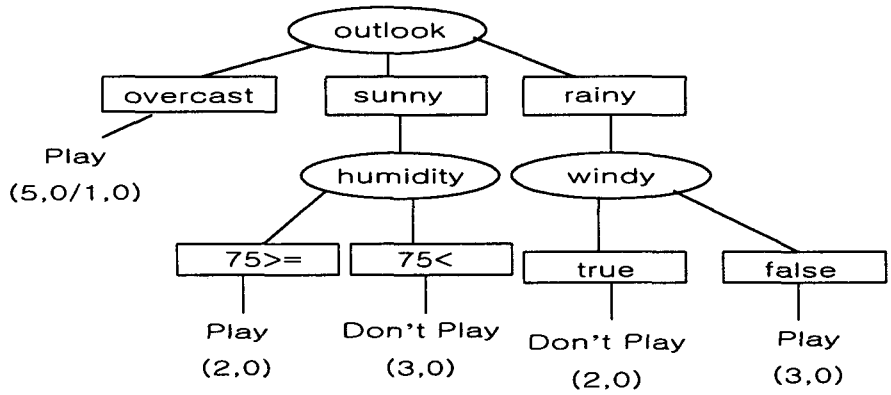
4. 제안 시스템

4.1 기본 개념

본 논문에서는 정확성, 간결성, 적절성 등에서 효과가 있는 C4.5 알고리즘을 이용한다. 기본 아이디어는 우선 DW의 historical data를 이용하여 1차 의사결정나무를 생성한다. 나무 모형에서 stop splitting rule을 가지고 정확한 구조의 나무 모형을 만드는 것은 곤란하다. 그래서 제안된 것이 충분한 크기의 나무모형을 생성한 후 가지치기를 하는 것이다. 다른 하나는 cross validation 혹은 test sample을 이용하여 가지의 변화에 따라 error rate의 변화를 살피고 이를 최소화하는 가지를 선택하거나 혹은 약간의 오차를 허용하여 보다 간단한 모형을 선택하는 방법이다. 실제로 CART는 두 가지를 모두 옵션으로 사용할 수 있도록 하였다.
그러나 C4.5에서는 후자의 경우만 가능하다. 또한 C4.5에서 사용하는 test sample은 training set과 같은 것이다. 따라서 C4.5의 divide and conquer 방법을 이용하여 나무를 형성하되 test sample을 이용해 가지치기를 할 때 서비스하고자 하는 개인의 정보를 이용하여 error rate에 의한 validation을 시행한다. 그러나 개인의 정보는 test set으로 너무나 적은 데이터이기 때문에 개인 데이터는 전체 test set의 일부로 쓰되 패널티의 반대 개념인 가중치를 두도록 한다. C4.5 알고리즘은 너무 동떨어진 데이터에 대해서는 분류 규칙을 적용하지 않으므로 크게 치우칠 우려가 적다. 이렇게 해서 얻어진 의사결정나무는 historical data를 반영하면서도 개인의 취향을 더 고려한 결정을 보여줄 수 있다.

4.2 제안 시스템의 효과

위에 제안된 시스템을 구현함으로써 얻을 수 있는 효과로는 다음과 같은 것들이 있다. 우선은 전자상거래를 위한 우수한 모델을 제시해 준다는 것이다. 데이터 마이닝과 e-CRM을 위한 상업용 소프트웨어들이 많이 나오고 있지만 그 효과가 입증되지 못하고 있다. 그것은 기존의 데이터 마이닝 알고리즘들이 주로 통계적 분석을 위한 모형들을 위한 것이기 때문에 유동적인 e-CRM의 상황에 적용하기 어려운 까닭이다.
둘째로 고객에게 보다 개인화 된 정보를 제공할 수 있다. 기존의 경직된 의사결정나무 모형으로 분류에 대한 설명만 해주는 것이 아니라 개인의 데이터를 적용하여 새로운 의사결정나무를 제시하여 준다.



<그림 4> 제안 시스템에 의해 생성된 의사결정나무

4.3 e-CRM의 효과

이 시스템을 통해 e-CRM의 효과 중 특히 다음과 같은 특징들을 기대할 수 있다. 고객에 대한 맞춤 서비스를 제공함으로써 고객의 로열티(loyalty : 충성도)에 대한 우위를 점유할 수 있다. 고객의 로열티란 기존 고객의 재구매율을 높임으로써 경영성과에 직접적으로 영향을 미치는 요소를 말한다. 인터넷 비즈니스에서는 신규고객 위주로 기존 고객을 무시하는 경향이 있었다. 기업이 신규고객 위주로만 Promotion을 하면 기존 고객의 Customer Loyalty는 흔들리게 된다. 이러한 로열티는 CRM, 즉 고객관계관리를 통해서 얻을 수 있다.또한 통계 나무모형의 특징으로 인해 마케팅과 의사결정에 드는 비용을 절감할 수 있다. 의사결정나무는 중요한 입력변수를 찾기가 쉽고 모형이 간단하다. 또한 통계학적인 용어를 쓰지 않고도 쉽게 변수들 간의 관계를 설명해주어 이해하기 쉽다. 따라서 전문가가 아니더라도 결과와 원인을 쉽게 분석할 수 있다. 마지막으로 고객의 선택이 실시간으로 반영되므로 고객과의 원활한 의사소통을 기대할 수 있다. 이를 통해 상품이나 서비스의 질을 개선하는 것 뿐 아니라, 사용자 및 고객 개인에게 있어서 더 가치 있는 것을 제공할 수 있다.

4.4 제안 시스템의 실험평가

원래의 데이터에 새로운 레코드 (overcast, 83, 86, false, no)를 더하여 2차 실험용 데이터 셋을 마련하였다. 분류의 적절성을 어떻게 평가할 것인가 하는 문제에 있어서 TP Rate, FP Rate, TN Rate, FN Rate를 따져 볼 수 있는데, 이번 실험에서는 TP

Rate와 FP Rate를 이용하여 Precision, Recall, F Measure값을 계산했다. TP Rate는 예측값과 실제값이 동일한 경우의 확률이고 FP Rate는 예측값이 실제와 다른 경우의 확률이다. FN Rate는 예측하지 못한 Yes의 확률이다.

$$\text{Precision} = TP / (TP+FP)$$

$$\text{Recall} = TP / (TP+FN)$$

α = P와 R의 가중치를 결정하는 요소(보통 0.5)

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}}$$

제안된 시스템과 기존의 시스템과 성능평가를 위해서, 실험은 C4.5 알고리즘에 의해 시행되고 10 cross validation 방법으로 테스트를 거쳤다. 그 결과는 다음과 같다.
<표 4>와 <표 5>는 기존의 의사결정나무와 제안된 시스템의 에러율의 차이를 보여준다.

<표 4> C4.5에 의한 의사결정나무의 에러율

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.778	0.8	0.455	0.556	0.5	yes
0.2	0.222	0	0	0	no

<표 5> 제안 시스템에 의한 의사결정나무의 에러율

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.556	1	0.455	0.556	0.5	yes
1	0.444	0	0	0	no

본 연구에서 제안한 시스템과 기존의 C4.5와 비교했을 때 동일한 예측율과 취소율을 보였으며, 제안 시스템이 에러율에서 정분류의 확률을 높이고 오분류의 확률을 낮추어, 정확도 높은 예측결과를 보였다. 제안 시스템은 복잡한 계산을 지양하고 해석이 용이하면서도 개인의 데이터에 대해 가중치를 두어 고객 맞춤을 실현시킬 수 있도록 하였다.

5. 결 론

인터넷의 확산과 데이터의 전산화로 인한 DW, 데이터 마이닝의 보급과 그 활용은 전자상거래 시장에서 중요한 문제임을 인식할 수 있다. 이론적으로만 정립되어 있는 각종 분석 도구들을 실세계에서 활용한다면 더욱 그 가치는 높아질 것이다. 지금까지 실제 데이터를 적용하여 CHAID, CART, C4.5의 세가지 알고리즘을 적용시켜 의사결정나무를 생성하고, 그 차이점과 모델의 정확성, 간결성, 확장성 등에 대하여 살펴 보았다. 의사결정나무는 e-CRM을 위해 효과적으로 쓰일 수 있는 중요한 데이터 마이닝의 도구임을 알 수 있다.

본 논문에서는 개인화라는 목적에 맞추어 의사결정나무의 경직성을 보완하여 주면서도 예측의 정확도를 높일 수 있는 시스템을 제안하였다. 제안 시스템은 C4.5와 비교했을 때 동일한 예측율과 취소율을 보였다. 또한 FP Rate가 상승하였으며, 반면에 TP Rate는 내려갔다. 이 시스템은 복잡한 계산을 지양하고 해석에 용이하면서도 개인의 데이터에 대해 가중치를 두어 고객 맞춤을 실현시킬 수 있도록 하였다. 실제로 전자상거래 사이트에서 적용할 때에는 고객 분류, 상품 추천 등에 쓰일 수 있을 것이다.

6. 참 고 문 헌

[1] A. Berson, S. Smith and K. Thearling, Building data mining applications for CRM, McGraw-Hill, 1999.

[2] J. Ross Quinlan, "Simplifying Decision Trees", Int. J. Man-Machine Studies, 27, pages 221-234, 1987.

[3] Kim, H. and Loh, W.-Y. Classification trees with unbiased multiway splits, Technical Report 1012, Department of Statistics, University of Wisconsin-Madison, 1999.

[4] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Classification and Regression Trees, Wadsworth, Belmont, 1984.

[5] Loh, W.-Y. and Shih, Y.-S. Split selection methods for classification trees, Statistica Sinica, 7, 815-840, 1997

[6] Salford Systems, CART, Salford Systems, San Diago, CA, 1997.

[7] Song, M., Yoon, Y., "A comparative study on variable selection methods in data mining software packages", 한일통계학회, 2000.

[8] Quinlan, J.R., C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA, 1993.

저 자 소 개

양정회 : 성균관대학교를 졸업하고, 동 대학원에서 석사 및 박사학위를 취득.
현재, 인덕대학 산업시스템경영과에서 교수로 재직중이며,
주요관심분야는 QM, SCM, RAM, CRM 등이다.

한서정 : 성균관대학교 대학원 컴퓨터공학부에서 석사학위를 취득,
현재, 서울대학교 컴퓨터공학부에 박사과정 중이며,
주요관심분야는 CRM, e-business 등이다.