한국산학기술학회논문지
Vol. 7, No. 3, pp. 518-523, 2006

하이브리드 의사결정나무와 인공신경망 모델을 이용한 방문학습지사의 고객세분화

서광규[1], 안범준[1*]

Customer Segmentation of a Home Study Company using a Hybrid Decision Tree and Artificial Neural Network Model

Kwang-Kyu Seo[1] and Beum-Jun Ahn[1*]

요 약 본 논문은 하이브리드 의사결정 나무(CART)와 인공신경망 모델을 개발하여 고객의 이탈에 대한 예측을 높이기 위하여 가정방문 학습지 고객의 패턴을 분류하고, 분석하는 새로운 방법에 대하여 연구하였다. 의사 결정나무(CART5)를 형성하여 선택된 결정변수들은 인공신경망의 입력벡터 값으로 선택되는 새로운 방법을 제시하였다. 고객관리측면에서 본 논문은 가정방문 학습지 회사의 기존고객을 분류하여 패턴을 분석함으로써 우수한 고객의 지속적인 관리와 이탈 가능성이 많은 고객을 차별 관리하여 기업이익을 증대시킬 수 있을 것이다. 새롭게 제안한 하이브리드 모델은 기존의 의사결정트리모델(CART), 회귀모형, 인공신경망 모델과 비교한 결과 그 예측 정확성이 높음을 확인할 수 있었다.

Abstract Due to keen competition among companies, they have segmented customers and they are trying to offer specially targeted customer by means of the distinguished method. In accordance, data mining techniques are noted as the effective method that extracts useful information. This paper explores customer segmentation of the home study company using a hybrid decision tree and artificial neural network model. With the application of variance selection process from decision tree, the systemic process of defining input vector's value and the rule generation were developed. In point of customer management, this research analyzes current customers and produces the patterns of them so that the company can maintain good customer relationship. The case study shows that the predicted accuracy of the proposed model is higher than those of regression, decision tree (CART), artificial neural networks.

Key Words : Customer Segmentation, Decision Tree, Artificial Neural Network, Hybrid Model, Customer Relationship Management.

1. 서론

현대의 많은 기업들은 이미 시장의 포화상태에서 경쟁을 하고 있고, 이러한 경쟁사회에서는 작은 정보하나라도 기업에게 큰 영향을 끼치게 된다. 대부분의 기업들은 무형적으로든 유형적으로든 방대한 양의 정보를 축적하고 있는데, 경쟁력을 확보하기 위해서는 축적된 데이터들 중에서 유용한 것들을 골라내고 이것을 분석하는 것이 중요하다. 데이터마이닝은 대용량의 데이터로부터 유용하게 활용될 수 있는 지식을 효과적으로 찾아내는 지식 탐사의 한 연구 분야로써, 수많은 형태의 방대한 데이터에서 그 각각의 목적에 도움이 될 수 있는 유용한 지식을 추출하는 것이다. 최근에는 기업 업무의 효율적 수행을 위해 데이터베이스를 이용하고, 그 결과를 활용하는 단계로부터 데이터 자체의 분석을 통해 행동 패턴을 추출해내고 이 결과를 업무와 생산의 효율성 증대를 위해 이용하는 단계로 넘어서는 추세이다. 이러한 추세에 맞추어 데이터마이닝 기술이 소개되었고, 이는 기업의 경쟁력 확보와 문제점 파악을 위한 기반 기술로 많은 발전이 이루어지고 있다 [1, 2, 3, 6].

사교육적 차원에서 학원이나 과외 말고는 마땅히 교육

[1]상명대학교 산업정보시스템공학과
*교신저자: 안범준(bjahn@smu.ac.kr)

시스템이 없었던 초기의 방문 학습지 시장은 비교적 경쟁이 약한 시장이었다. 굵직한 몇 개의 기업만이 업계의 전부였고, 다양하지 못한 상품 등으로 고객들은 선택의 폭이 좁았다. 학부모들은 다른 아이들이 하는 것을 내 자식이 하지 않으면 혹시나 뒤쳐질까 하는 불안 심리를 가지고 있으므로, 좋든 싫든 방문 학습지를 시키는 경우가 많았다.

그러나 근래에 들어서는 여러 기업이 업계에 뛰어들고 있고, 그만큼 다양한 상품과 구성으로 소비자들의 입맛을 사로잡고 있는 상황이 되면서 이들 기업간의 경쟁은 점점 심화되었다. 그리고 방문 학습지 시장의 주된 대상인 초등학생층은 양적 팽창의 한계가 있기 때문에 고객 유치를 위한 쟁탈이 심화되고 있다. 이러한 시점에서 다른 여러 업계와 마찬가지로 방문 학습지 업계 또한 신규고객의 유치보다는 비용 효율이 높은 기존고객의 이탈을 방지하는 측면에 한층 무게중심을 두고 있다.

본 연구는 데이터마이닝 기법 중에 의사결정 나무를 이용하여 데이터를 분류하고, 의사결정나무로부터 생성되는 결정변수들을 선택하여 인공신경망의 새로운 입력변수로 선정함으로써 분류 모델의 예측 정확도를 향상시키기 위한 하이브리드 모델을 제안한다. 제안하는 모델을 검증하기 위하여 G학습지의 기존고객과 이탈고객의 성향을 분석하고, 이탈고객의 패턴을 추출하여 향후 마케팅 전략 수립에 도움을 주고자 한다.

2. 이론적 고찰

2.1 CART 의사결정나무

CART(Classification and Regression Tree) 알고리즘은 의사결정나무를 형성하는데 있어서 가장 보편적인 알고리즘이다 [4]. CART는 지니(범주형 목표변수인 경우 적용) 또는 분산의 감소량(연속형 목표변수인 경우 적용)을 이용하여 이진 분리(binary split)를 수행하는 알고리즘이다.

지니지수(Gini Index)는 불순도(impurity)를 측정하는 하나의 지수이다. 임의의 한 개체가 목표변수의 i 번째 범주로부터 추출되었고, 그 개체를 목표변수의 j 번째 범주에 속한다고 오분류(misclassification)할 확률은 $P(i)P(j)$가 된다. 여기에서 $P(i)$는 각 마디에서 한 개체가 목표변수의 i번째 범주에 속할 확률이다. 이러한 오분류 확률을 모두 더하여 식 (1)을 얻을 수 있다.

$$G=\sum_{j=1}^{c}\sum_{i\neq j}P(i)P(j) \tag{1}$$

이는 위와 같은 분류 규칙하에서 오분류 확률의 추정치가 된다. 여기서 c 는 목표변수의 범주 수를 말한다.

일반적으로 CART는 범주형 목표변수에 대해서는 지니지수를 분리기준으로 사용한다. 지니지수는 다음의 식 (2)와 같이 표현된다.

$$G=\sum_{j=1}^{c}P(j)(1-P(j))=1-\sum_{j=1}^{c}P(j)^2=1-\sum_{j=1}^{c}(\frac{n_j}{n})^2 \tag{2}$$

여기에서 n은 그 마디에 포함되어 있는 관찰치수를 말하고, n_j는 목표변수의 j 번째 범주에 속하는 관찰치 수를 말한다. 지니 지수는 n개의 원소 중에서 임의로 2개를 추출하였을 때 추출된 2개가 서로 다른 그룹에 속해 있을 확률을 의미하며 Simpson의 다양도 지수(diversity index)로도 알려져 있다. 목표변수의 범주가 2개인 경우에는 지니 지수는 다음 식 (3)과 같이 표현될 수 있다.

$$G=2P(1)P(2)=2(\frac{n_1}{n})(\frac{n_2}{n}) \tag{3}$$

이는 카이제곱 통계량을 사용하는 것과 같은 결과를 갖는다. CART 알고리즘은 지니 지수를 가장 감소시켜 주는 예측변수와 그 변수의 최적분리를 자식마디로 선택하는데, 지니 계수의 감소량은 다음 식 (4)와 같이 계산된다.

$$\Delta G=G-\frac{n_L}{n}G_L-\frac{n_R}{n}G_R \tag{4}$$

여기서 n은 부모마디의 관측치 수를 말하고 n_R 과 n_L는 각각 자식마디의 관측치 수를 의미한다. 즉, 자식마디로 분리되었을 때의 불순도가 가장 작도록 자식마디를 형성하는 것이다. 이는 다음 식 (5)와 같은 자식마디에서 불순도의 가중합을 최소화하는 것과 동일하다.

$$P(L)G_L+P(R)G_R=\frac{n_L}{n}G_L+\frac{n_R}{n}G_R \tag{5}$$

2.2 인공신경망

신경망 분석은 인간의 신경망을 모방하여 실제 자신이 가진 데이터로부터 반복적인 학습과정을 거쳐 데이터에 숨어 있는 패턴을 찾아내는 모델링 기법이다. 자료 분석 분야에서 신경망은 복잡한 구조를 가진 자료에서의 예측

문제를 해결하기 위해서 사용되는 유연한 비선형모형의 하나로 분류될 수 있다. 신경망은 은닉마디라고 불리는 독특한 구성요소에 의해서 일반적인 통계모형과 구별되어진다. 은닉마디는 인간의 신경세포를 모형화한 것으로써, 각 은닉마디는 입력변수들의 결합을 수신하여 목표변수에 전달한다. 이때 결합에 사용되는 계수들의 연결강도라고 부르며, 활성함수는 입력값을 변환하고 이를 입력으로 사용하는 다른 마디로 출력하게 된다. 신경망에서 학습 알고리즘의 기본원리는 입력층의 각 유니트에 입력자극을 주면, 이 신호는 각 유니트에서 변환되어 은닉층에 전달되고 최후에 출력층에서 결과를 출력하게 된다. 또한 관리학습에서는 입력 및 원하는 출력패턴이 네트워크에 제시된다. 네트워크 입력층에 주어진 입력자극이 출력층에 전파되면서 변한 출력패턴을 목표패턴과 비교한다. 네트워크에서 출력된 패턴이 목표패턴과 일치하는 경우에는 학습이 일어나지 않으며 그렇지 않은 경우는 얻어진 출력패턴과 목표패턴의 차이를 감소시키는 방향으로 네트워크의 연결 강도를 조절하여 학습을 한다.

3. 하이브리드 의사결정나무와 인공신경망 모델

본 연구에서는 의사결정나무인 CART와 신경망 알고리듬을 결합한 하이브리드 모델을 제안하였는데, 의사결정나무의 장점인 규칙생성과 이를 이용 한 변수선정으로 신경망에 결합시킴으로써 보다 좋은 예측의 결과를 얻고자 한다.

3.1 CART 의사결정나무의 형성

먼저, CART를 이용하여 고객들을 분류하게 되는데, CART는 각 마디 데이터 분할을 형성하며 이지분리 나무구조를 만들게 된다. CART는 순환적 분리를 수행하므로 특정 입력변수에 의해 나누어진 뒤에도 재차 동일한 변수에 의해 나누어진다.

의사결정나무 형성 결과를 이용하여 생성된 결정변수들은 인공신경망의 입력값으로 선택된다. 결정변수는 각 마디를 형성하는 변수들로 구성된다. 이는 CART에 의해 부모와 자식마디까지 형성되었던 최고의 이익비율을 가지는 각 마디의 변수들을 이용하여 선택하게 된다. 이 결과를 이용하면 의미 없는 변수들을 제거되므로, 기존의 고려되었던 변수보다 적은 변수가 새로운 모델의 형성에 필요한 결정변수가 되어 학습 효율이 증대되며 더 정확한 예측 결과를 얻을 수 있게 된다.

3.2 하이브리드 모델의 구축

하이브리드 모델을 개발하기 위하여 학습데이터가 입력된다. 입력된 학습데이터는 CART알고리즘에 의해 필요한 결정변수와 규칙 생성을 위하여 의사결정나무를 형성한다. 나무 전체의 예측율을 높게 형성해주는 정지 기준과 가지치기를 실행하여 의사결정나무를 형성하고, 규칙을 산출하게 된다. 이렇게 형성된 의사결정나무에 제안된 방법으로 결정 변수를 구하고, 이 결과를 인공신경망 구조의 새로운 입력 변수로 선정한다.

결합모델을 형성하는데 사용되는 신경망 알고리즘은 현재까지 유용하게 활용되는 오류 역전파 알고리즘을 적용하였다. 오류 알고리즘은 학습용 자료가 주어지면 임의

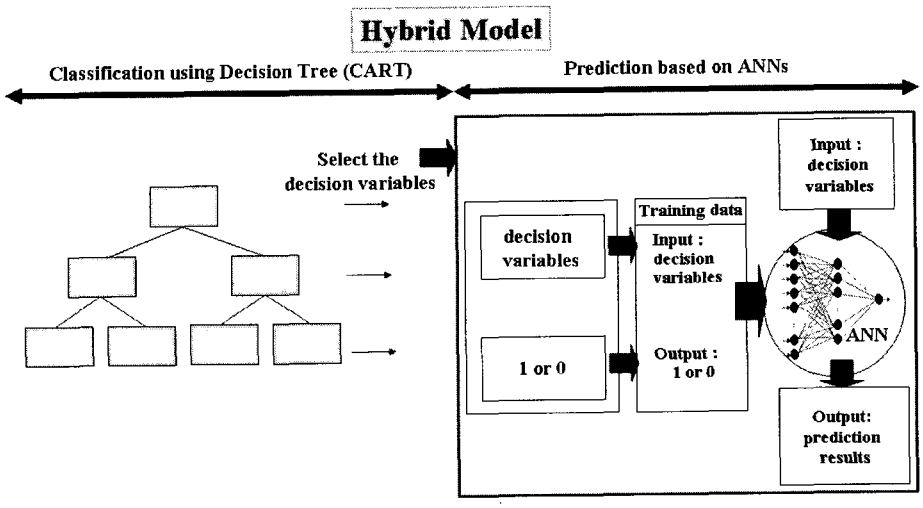


그림 1. 제안된 하이브리드 알고리듬

로 주어진 연결강도를 이용하여 결과값을 계산한다. 그리고 계산된 결과와 실제 값의 차이인 오차를 계산하여 오차신호를 산출하고 이를 통하여 은닉층과 출력층으로 역전파시켜 연결 강도를 조정한다. 즉, 출력층 오차신호를 은닉층에 역전파하여 입력층과 은닉층사이의 연결강도를 변경하는 학습방법이다. 출력노드는 하나로 이루어져 있고 목적함수를 최적화하기위하여 시행착 오 방법을 사용하여 은닉층의 수와 노드를 결정하고, 목적함수가 최소일 때, 각 파라미터의 값을 선정하도록 하였다. 제안된 하이브리드 알고리즘의 구조는 그림 1과 같다.

4. 사례 연구

본 논문에서는 G 방문학습지의 학습지를 현재 구독하고 있는 고객들을 대상으로 실시한 설문조사 자료를 이용하여 이탈고객에 대한 분석을 수행하였다. 유지고객의 만족도와 불만도 등 여러 항목들을 통해 패턴을 읽어내어 G 방문학습지사에서 어떠한 요인들이 고객유지나 이탈에 영향을 미치고 있는지 알아보는 것이 주요한 관점이다. 본 연구에서는 제안한 연구결과의 정확도를 평가하기 위해 데이터 마이닝 도구인 Clementine 7.0 [7]의 여러 모델들과 비교 분석하였다.

표 1. 설명 변수 구분

변 수 명	구 분
1. 살고 있는 곳	연속형 변수
2. 이탈여부	유/무 (목표변수)
3. 자녀의 나이	연속형 변수
4. 이용 기간	연속형 변수
5. 자녀의 성격	내향적/외향적/보통
6. 자녀의 집중도	연속형 변수
7. 자녀의 집중도	연속형 변수
8. 소득 수준	50~100/100~150/150~200/250~300/기타
9. 부모의 학력	중졸/고졸/대졸/대졸이상
10. 선택 경로	전화/주변사람/인터넷/광고매체/잡지/기타
11. 가격 만족도	연속형 변수
12. 희망 가격	연속형 변수
13. 사은품 만족도	연속형 변수
14. 방문횟수 만족도	연속형 변수
15. 희망 방문횟수	1주2번/1주3번/1주4번/1주4번 이상/기타
16. 학습시간 만족도	연속형 변수
17. 희망 학습시간	10분이내/20분이내/30분이내/30분이상/기타
18. 선생님 만족도	연속형 변수
19. 선생님 조건	선생님의 자질/학력/성격/생김새/옷차림/기타
20. 학습지 내용 만족도	연속형 변수
21. 내용 면에서 우선시 하는것	문제풀이/그림위주/시험대비/원리이해/기타
22. 디자인 만족도	연속형 변수
23. 만족스럽지 못한 점	연속형 변수
24. 디자인에서 우선시 하는 것	책의 크기/색깔/글씨크기/그림/기타
25. 광고 접촉 유무	유/무 (목표 변수)
26. 광고 만족도	연속형 변수
27. 광고 영향 정도	연속형 변수
28. 인지도의 중요성	연속형 변수
29. 기업 이미지의 중요성	연속형 변수
30. 홈페이지 이용할 의향	연속형 변수
31. 학습지를 바꾸는 이유	성적부진/선생님능력부족/내용부실/많은회비/기타
32. 변경 여부	유/무 (목표 변수)
33. 변경 시기	연속형 변수
34.선택 시 가장중요한 사항	자녀의선호도/선생님능력/학습지내용/가격/성적향상/방문횟수/사은품/학습시간/책디자인/광고/인지도/기업인지도/온라인서비스/기타

4.1 분석자료 및 방법

분석에 사용된 자료는 고객들을 대상으로 설문 조사한 400명의 자료를 이용하였다. 이중 유지고객 250명(62.5%), 이탈고객 150명(37.5%)이다. 목표변수는 이탈고객은 1(무)의 값을 주고 유지고객은 0(유)의 값을 갖는 변수로 설정하였다. 설명변수로는 설문지에 작성된 항목들로 구성되었다. 설명변수는 총 34개로 연속형 변수와 명목형 변수가 함께 사용되었는데(표 1 참조), 초기분석을 위해 SAS를 활용하였다[3].

실험을 위하여 학습데이터는 전체데이터의 80%를, 검증데이터는 전체데이터의 20%를 임의로 선택하였고, 각 데이터의 유지고객과 이탈고객으로 분할된 데이터들의 분포는 전체데이터와 비슷하게 구성되도록 하였다.

4.2 실험결과

CART 의사결정나무를 이용하여 분석한 결과, 고객이탈여부에 영향을 미치는 중요한 변수로는 학습시간, 선생님의 조건, 가격, 학습지 내용, 방문횟수, 선택경로임이 확인되었고, 이렇게 확인한 주요 변수를 하이브리드 모델의 입력변수로 선택하였다.

제안한 하이브리드 모델을 정확도를 평가하기 위하여 다양한 예측 기법인 회귀모형, CART, 인공신경망과 비교하였는데, 그 예측결과는 표 2와 같다.

표 2. 다양한 모델의 예측 결과 비교

모델 / 데이터	회귀분석	CART	인공신경망	하이브리드 모델
Training Data	85.31%	86.88%	91.88%	93.12%
Test Data	82.50%	83.75%	88.75%	91.25%

표 1에서 확인할 수 있듯이 본 연구에서 제한한 하이브리드 모델의 예측결과가 가장 높음을 확인할 수 있었고, 인공신경망도 상대적으로 우수한 예측결과를 보여 주었다.

5. 결론

본 연구에서는 현존하고 있는 G학습지의 유지고객과 이탈고객을 대상으로 의사결정나무와 인공신경망을 결합한 새로운 하이브리드 모델을 적용해 고객이탈에 영향을 주는 요인과 이탈고객집단을 분류하는 방법을 살펴보았다. 의사결정나무를 이용하여 고객이탈과 관련이 깊은 변수로는 학습시간, 선생님의 조건, 가격, 학습지 내용, 방문횟수, 선택경로 등으로 확인할 수 있었으며, 이를 결정변수로 선택하여 인공신경망의 입력값으로 사용하여 학습을 수행하는데, 그 예측 결과는 기존의 기법인 회귀모형, CART, 인공신경망보다 우수함을 확인할 수 있었다.

제안한 모델을 이용하면, 고객데이터의 패턴을 분석함으로써 우수고객의 지속적인 관리와 이 탈가능고객의 분류를 통하여 회사의 손실을 줄일 수 있을 것이다. 회사는 이러한 분석을 통하여 얻은 유익한 정보를 활용하면 철저한 고객관계관리가 가능하리라 생각되고, 이뿐만 아니라 신규고객을 유치한 후부터 친절한 서비스와 고객 개개인의 정보를 신속히 하고 지속적으로 파악하여 고객의 니즈를 만족시켜야 하고, 불만의 처리도 신속하고 정확하게 받아들여야만 할 것이다.

참고문헌

[1] 강현철 외, 데이터마이닝 - 방법론 및 활용, 자유아카데미 2001.

[2] 김영아, 시장세분화를 위한 데이터마이닝 응용에 관한 연구, 서강대학교 석사학위 논문, 2001.

[3] 노형진, 한글 SPSS 10.0에 의한 조사방법 및 통계분석, 형설출판사, 2004.

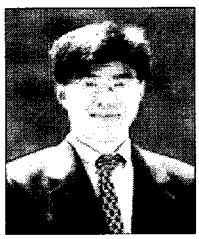
[4] Quinlan, J.R, "Decision Trees and multi-valued attributes", Machine Intelligence 11 pp. 305-318, 1988.

[5] 이현정, 데이터마이닝을 이용한 보험회사 고객이탈분석에 관한 연구, 중앙대학교 석사학위 논문, 2001

[6] 조혜정, 고객세분화를 위한 데이터마이닝 기법 비교, 동아대학교 석사학위 논문, 2001.

[7] 허준 외, Clementine 7 매뉴얼, SPSS 아카데미, 2003.

안 범 준(Beum-Jun Ahn) **[종신회원]**



- 1989년 8월 : 고려대학교 산업공학과 (공학사)
- 2002년 2월 : 일본히로시마대학 경영정보전공(경제학석사)
- 1998년 2월 : 일본히로시마대학 경영정보전공(경제학박사)
- 1999년 3월 ~ 현재 : 상명대학교 산업정보시스템공학과 부교수

<관심분야>
생산관리, 공급사슬망관리, 품질관리

서 광 규(Kwnag-Kyu Seo) **[정회원]**

- 2002년 8월 : 고려대학교 산업공학과 (공학박사)
- 1997년 9월 ~ 2003년 2월 : 한국과학기술연구원(KIST) 시스템연구부 선임연구원
- 2003년 3월 ~ 현재 : 상명대학교 산업정보시스템공학과 조교수

<관심분야>
정보시스템, 데이터마이닝과 CRM, 생산공학, 인공지능