의사결정나무 기법을 활용한 백화점의 고객세분화 사례연구

채 경 희*, 김 상 철**

국문초록

기업에서는 마케팅 비용대비 효과를 극대화하기 위하여, 고객을 세분한 후, 목표고객을 선별하여 해당 고객에 적절한 캠페인을 실시하고 있다. 특히 고객세분화 방법으로 통계 모형을 비롯하여 데이터마이닝 방법 등 다양한 방법들이 활용되고 있다. 그 중에서도 데이터마이닝은 1990년대 초에 도입되어 다양한 경영 문제를 해결하고 있다.

본 논문에서는 이와 같은 고객세분화에 활용되고 있는 데이터마이닝 방법에 대해 살펴본 후, 실제 백화점 사례를 기반으로 고객세분화에 주로 활용되고 있는 의사결정나무 분석 방법의 효과 및 장단점에 대해 논의해보고자 한다.

주제어 : 고객세분화, 데이터마이닝, 의사결정나무, STP전략

Ⅰ. 서 론

세분화 과정은 전체 고객을 확인 가능한 일정 규모의 집단으로 분할하는 것이다. 이와 같이 고객을 세분함으로써 기업에서는 다음과 같은 이점을 확보할 수 있다. 첫 번째, 마케팅 비용을 감소시킬 수 있다. 캠페인이 필요하거나, 캠페인에 반응할 가능성이 높은 고객을 선별하여 캠페인을 실시함으로써 비용 지출을 줄이고 효과는 극대화할 수 있다. 두 번째, 캠페인에 대한 부정적인 반응을 줄일 수 있다. 홍보 활동 및 정보제공 활동이 필요한 고객에게는 유용한 정보가 될 수 있지만, 필요로 하지 않는 고객에게는 번거로운 스팸(Spam)에 지나지 않을 수 있다. 고객세분화를 통해 목표고객을 설정하게 되면 이와 같은 부작용을 줄일 수 있다. 세 번째, 고객 가치를 향상시킬 수 있다. 소수의 고객을 대상으로 차별화된 관심과 서비스를 제공함으로써 고객 만족을 증대시키고, 충성도를 향상시킬 수 있다.

대부분의 고객은 일반적으로 제품 및 서비스 범주 내에서 가치를 부여하는 것 또는 그들이 지불하는 비용, 혹은 시장 접근 특성이 유사한 소비자들로 이루어진 하위집단으로 나누어질 수 있다. 이러한 하위집단을 세분고객이라고 한다. 고객세분화의 목표는 실행 가능한 세분화와 의미 있는 세분화를 하는 것이다(박명호 외 3인, 2005). 이는 모든 고객세분화가 의미 있는 것은 아님을 의미한다. 세분화가 효과적이기 위해서는 다음과 같은 세 가지 전제조건을 따라야 한다. 첫 번째, 의미가 있어야 한다. 고객들의 행동 특성을 기술하고 설명하는 데 도움이 되어야 한다. 두 번째, 실행가능성이 있어야 한다. 세분 고객에 대한 타깃팅과 포지셔닝 의미에서 현실적으로 실행 가능해야 한다. 세 번째, 금전적으로 매력적이어야 한다. 목표 세분고객은 경제적으로 가치가 있어야 한다. 따라서 본 논문에서는 실제 A백화점에서 실시하고 있는 세분화 방법이 위와 같은 전제조건 기준, 즉 의미, 실행가능성, 금전적 측면에서 유용한지 평가해 보고자 한다.

A백화점을 포함하여 많은 분야의 기업들이 데이터마이닝 기법을 활용하여 의사결정을 하고 있다. 그러나 대부분 전문가의 지식에 의해 일련의 패턴이 구성되는 규칙기반 시스템에 대한 의존도가 높고, 그 외의 데이터마이닝 기법은 참고 자료로 활용되고 있다. 따라서 본 논문에서는 규칙기반 시스템의 적용 결과와 다른 데이터마이닝 기법의 적용 결과를 비교해 봄으로써, 데이터마이닝 기법 활용도에 대한 효과와 시사점에 대해서도 살펴보고자 한다.

데이터마이닝은 다양한 분야에 도입되어 그 효과성을 입증 받고 있다(김근형, 오성렬, 2009; 방정혜 외 2인, 2007; 이건창 외 2인, 2001). 이는 현대 사회의 복잡성으로 우리가 엄청난 양의 정보를 접하며 살고 있으며, 컴퓨터의 발전과 더불어 이러한 정보는 데이터베이스로 구축되어 주어지게 되었기 때문이다. 방대하고 다양한 형태의 데이터로부터 의사결정에 유용한 정보 및 지식을 발견하는 일련의 데이터 분석 및 모형 선정과정이 데이터마이닝(Data Mining)이다. 데이터마이닝을 사용한 사례로는 보험요율산정, 개인신용평가, 신용카드 부정거래자 색출, 데이터베이스마케팅, 텔레커뮤니케이션 서비스 등을 들 수 있다(구자용 외 2인, 2000). 이러한 사례를 통하여 볼 때 데이터마이닝 과정은 목표설정, 데이터 탐색 및 전처리, 분석, 결과해석 등의 일련의 과정을 통해 이루어지게 되며, A백화점의 경우에도 이와 같은 일련의 과정을 통해 살펴보도록 한다.

본 논문의 구성은 다음과 같다. 먼저 2장 문헌연구에서 일반적인 세분화의 기준에 대해 살펴봄으로써, 현재 A백화점에서 사용하고 있는 세분화 기준과 비교하도록 한다. 또한 세분화 기준으로써 활용될 수 있는 데이터마이닝 기법들을 소개한 후, A백화점에 적절하다고 판단되는 의사결정나무 분석의 선택 이유 및 장단점에 대해 대해서 논의할 것이다. 다음으로 3장에서는 의사결정나무를 활용한 세분화 방법을 A백화점의 사례를 기반으로 살펴보고, 분석결과 및 일련의 데이터마이닝 과정을 통해 획득 가능한 A백화점 고객의 특성에 대해 살펴본다. 그리고 기존의 분석 방법과 비교해 봄으로써 제안하는 기법의 유용성 및 가치에 대해 논의해 보고자 한다. 그리고, 마지막 4장 결론을 통해 본 논문의 의의 및 한계점에 대해 논의할 것이다.

* 이리스컨설팅(주) 컨설턴트, 주저자, E-mail: khchae@lirsconsulting.com
** 유한대학 유통물류학과 부교수, 교신저자, E-mail: kimsc@yuhan.ac.kr

Ⅱ. 문헌연구

1. 고객 세분화 기준

일반적인 마케팅 문헌에서 제시되고 있는 포괄적인 세분화 기준은 인구통계적 변수, 지리적 변수, 심리적 변수, 인지 및 행동변수 등이 있다(박명호 외 3인, 2005; 이두희, 2006).

인구통계적 변수는 나이, 성, 직업, 인종, 소득, 가족상황, 생애단계 등을 의미하며, 지리적 변수는 국가, 지역, 온라인 또는 오프라인 등을 의미한다. 인구통계적 변수는 측정하기가 비교적 용이하며, 각종 통계자료를 이용하여 파악할 수 있기 때문에 일반적으로 많이 사용되는 세분화 변수이다. 그러나 A백화점의 경우, 고객의 인구통계적 정보는 활용하기 어렵다. 이는 현금 사용 고객의 경우, 정보 획득 및 구매 추적이 불가능하며, 카드 사용 고객의 경우에는 본인의 카드가 아닌 가족의 카드를 사용하는 경우가 많기 때문에, 인구통계적 변수를 기반으로 세분화 하는 것은 바람직하지 않다.

심리적 변수는 라이프 스타일, 사회계층, 성격유형 등을 의미하며, 객관적으로 측정이 용이한 인구통계적 변수에 비해 추상적이어서 고객 규모의 측정이 어렵고, 세분 고객에 대한 접근 가능성을 찾기가 어렵다. 반면, 인구통계적 변수보다는 소비자의 행동을 보다 근원적으로 설명해 줄 수 있는 변수이다.

인지 및 행동 변수는 제품 및 서비스 효익에 대한 태도, 이용률, 충성도, 제품 및 서비스 구매패턴 등을 의미한다. 이러한 인지 및 행동 변수는 구매행동을 직접적으로 나타내거나 구매 행동과 밀접하게 관련이 있는 변수들이므로 고객을 세분하는데 효과적이다. 실제 A백화점의 경우에도 행동 변수의 하나로써, 내점일수와 구매금액을 기준으로 고객을 세분해오고 있었다. 그러나 최근 이와 같은 세분화 기준에 대한 유용성 및 정확도에 대한 문제가 제기됨으로써, 보다 정확도가 높은 새로운 세분화 방법에 대한 필요성이 대두되었고, 그 대안으로써 데이터마이닝 방법의 하나인 의사결정나무 분석 방법에 대한 도입이 고려되었다.

2. 세분화에 활용 가능한 데이터마이닝 기법

데이터마이닝은 컴퓨터 과학의 인공지능(artificial intelligence), 패턴인식 등에 활용되는 기계학습(machine learning), 문헌연구에서 활용되는 정보 필터링(information filtering) 이론에서부터 시작되었다. 예를 들어 도서관에 방문한 고객이 입력한 단어를 기반으로 고객이 찾고자하는 키워드와 가장 적합한 책의 목록을 나열한다거나, 신용도에 따른 고객의 과거 거래 패턴을 인식하여 모형을 수립한 뒤, 개인의 신용도 등을 미리 예측하는 작업을 데이터마이닝이라 명명하기 시작한 것이다. 기계학습 이론에서는 신경망 이외에도 의사결정나무 모형 등 다양한 알고리즘들이 있다(구자용 외 2인, 2000).

고객세분화에 활용될 수 있는 데이터마이닝 기법들로는 규칙기반 시스템, 의사결정나무, 인공신경망 등이 주로 사용되고 있으며, 각 각에 대해 살펴보면 다음과 같다. 그 외에 간혹, 사례기반추론이나 유전자알고리즘 등이 활용되는 경우도 있으나, 본 논문에서는 주로 활용되는 기법을 중심으로 설명하고자 한다.

1) 규칙기반 시스템

규칙기반 시스템은 전문지식을 끌어내어 일련의 규칙 형태로 표현하는 것이다. 이러한 규칙들을 끌어내어 구축한 시스템이 특정 분야에서 전문가 역할을 대신하도록 하는 것이다. 이와 같은 시스템에서 대부분의 지식은 가정-결과(IF-THEN) 형식을 따르는 규칙의 형태로 저장되어 있다. A백화점 또한 규칙기반 시스템의 하나로 볼 수 있으며, 규칙기반 시스템을 적용한 사례로는 Elmer & Borowski(1988)의 방법이 있다. 이들은 예금대부 기관의 재무적 건전성을 평가하고, 도산 가능성을 예측하는 시스템을 구축하였다. 구축된 시스템은 공개된 정보만을 이용하여 기관들의 건전성을 측정하는 단일 점수를 산출하는 방식이었다.

모형의 유용성을 평가하기 위해 기존의 로짓모형과 성과를 비교하였을 때, 로짓의 결과가 미세하게 좋은 성과를 나타내기도 하였으나, 도산예측 기간을 늘림에 따라 Elmer & Borowski(1988)의 성과가 더 좋게 나타남으로써, 기업의 도산예측 가능성의 대한 장기 예측에는 규칙기반 시스템이 로짓에 비해 적절하다고 주장하였다(Elmer & Borowski, 1988; 조홍규, 2003). 현재까지도 규칙기반 시스템 및 로짓 기법은 도산예측 이외에도 구매예측, 신용도평가 등에서 유용한 분석 방법으로 활용되고 있다.

2) 의사결정나무

의사결정 나무는 누구나 이해할 수 있고 쉽게 설명되어질 수 있는 결과의 간결함으로 많은 분야에서 선호되고 있다. 고객의 의사결정 패턴을 분석해야하는 상품개발, 마케팅 부서, 그리고 문자, 지문인식 등을 연구하는 기계학습 이론분야에서 사용되고 연구되고 있다. 그밖에 누락된 관측값에 대한 처리가 다른 모델보다 우수하고 변수들 간의 교호작용(interaction)의 설명과 처리가 용이하다는 장점이 있다.

의사결정 나무의 시초는 통계학에서 시작되었다고 할 수 있다. 1980년에 Kass는 CHAID(CHi-squared Automatic Interaction Detector)라는 알고리즘을 소개하였는데, 카이제곱 적합성 검정에 근거한 의사결정 나무로써 현재까지 널리 사용되고 있다. 그 외에도 Quinlan(1986; 1993)의 ID3와 C4.5가 있으며, 이 또한 대표적인 알고리즘들이다. 의사결정 나무의 역사상 가장 중요한 학문적 결과는 Breiman 외 3인(1984)의 CART(Classification And Regression Tree)라고 할 수 있다. Breiman등은 CART를 통하여 의사결정 나무의 분할규칙에 의한 나무모형의 성장, 최종 모형의 선정을 위한 의사결정 나무의 가지치기 등을 이론적으로 정립하였다. CART 이후 의사결정 나무에 대한 연구가 더욱 활발히 진행되었으며 현재 CART와 함께 의사결정 나무의 양대 산맥을 이루는 Quinlan의 C4.5도 CART의 기본 하에 이루어진 것이다.

의사결정 나무는 각 노드에서 분할이 일어나면서 자라나게 된다. 자료의 분할과정은 분할 후 각 노드에 속하는 자료의 순수도(purity)가 가장 크게 증가하도록 진행되는데, 이를 위해 분할 기준이 되는 변수와 분할의 위치를 결정하여야 한다. 순수도의 증가란, 분할 후 각 노드에 속하는 자료의 구성이 어느 한 부류 만에 속하는 자료의 비율이 높다는 의미이다.

의사결정 나무 T 의 전체 불순도 $D(T)$는 다음과 같이 구한다.

$$D(T) = \sum_{g \in G} \Phi(g) p(g)$$

여기서 G 는 의사결정 나무 T 의 종료 노드의 집합이고 $p(g)$

는 종료 노드 g 에 속할 확률이다. 그리고 $\Phi(g) = \Phi(p_1(g), ..., p_J(g))$ 는 불순도 함수이며, 함수 Φ로 표현될 수 있는 불순도 함수로는 Gini 지수, Entropy 지수, Deviance 등이 있다. Gini 지수는 CART에서, ent ropy 지수는 C4.5에서 사용되는 분할기준이며, Deviance는 AT&T의 Chipman등에 의하여 제안된 함수이다(Han & Kamber, 2001).

3) 인공신경망

인간두뇌에 대한 연구는 상당히 오래 전부터 계속되어 왔으며 정보처리기술의 발달과 더불어 인간과 같이 사고하고 판단하며, 사물을 인식할 수 있는 인공지능을 실현시키기 위해 많은 노력이 있어 왔고, 덕택에 우리는 특정 분야의 전문가가 가진 전문지식을 컴퓨터에 옮기고 이를 자유로이 이용할 수 있는 단계에까지 이르렀다. 현재 제한된 범위 내에서라는 전제만 한다면 인간보다 더 빠르고 정확하게 사고하고 판단하는 인공지능 컴퓨터의 구현은 어려운 일이 아니다. 하지만 사물의 인식, 특히 시각적 패턴인식의 측면에서는 아직도 큰 진전이 없었다. 다시 말해, 한두 살 짜리 어린아이의 얼굴인식 능력이나 연상 능력은 현재 가장 우수하다는 수퍼컴퓨터를 여러 대 동원하여도 따라가기 힘든 실정이다(조홍규, 2003).

인공신경망은 인간의 신경세포와 유사한 PE(Processing Element)로 이루어져 있다. PE는 입력, 출력, 가중치, 뉴런함수의 네 부분으로 구성되어 있다. 입력과 출력은 0/1, 연속치, 집단 등의 다양한 형태를 가질 수 있다. 각 각의 입력은 효과에 대한 상대적인 가중치를 가지고, 가중치는 입력신호의 정도를 나타내기 위해 모형내에서 결정된다. 뉴런함수는 합산(summation), 활성화(activation), 전이(transfer), 학습(learning)의 네 가지가 있다. 합산함수는 가중치에 따라 입력을 더하고, 그 결과는 활성화함수의 입력이 된다. 전이함수는 활성화함수의 결과를 받아 이미 정해진 크기와 비교한 후 다음 PE에게 적절한 가중치를 가진 출력을 보내준다. 학습함수는 현재의 출력과 원하는 출력을 비교하여 오차를 감소시킨다. 병렬적으로 동시에 실행되는 PE의 집합을 층(layer)이라 부른다. 입력층과 출력층이 문제에서 정의된 변수에 따라 확정적인데 반해 은닉층 PE의 수는 유동적이다.

인공신경망은 두개의 유용한 특징을 가지고 있는데, 결점포용성(fault tolerance)과 적응성(adaptability)이다, 인공신경망은 하나의 처리 장치가 아닌 다수의 PE들의 상호작용을 통해 처리되므로, 불완전한 자료로부터 학습과 의사결정을 할 수 있는데, 이러한 특성을 결점포용성이라 한다. 적응성은 인공신경망의 자기변화를 의미하는 것으로, 지속적인 학습을 통해 가중치를 변화시키면서 적합한 모델을 만드는 과정을 말한다(조홍규, 2003; Han & Kamber, 2001).

인공신경망은 카드사의 사기적발(Fraud detection), 신용평가 등에 주로 활용되며, 다른 데이터마이닝 기법들에 비해 정확도가 높은 것으로 알려져 있다(Han & Kamber, 2001). 그러나 결과가 이해하기 쉽고 간결한 의사결정나무에 비해, 결과 도출 과정이 불투명하고 분석 과정이 복잡하여 이해하기 어려운 단점으로 인해 활용도는 비교적 낮은 편이다. 따라서 A백화점의 사례에는 비전문가인 마케팅 담당자 등의 의사결정자가 보고 쉽게 이해하고 판단할 수 있도록 하기 위하여 의사결정나무 기법을 활용하여 분석하였다.

Ⅲ. 의사결정나무를 활용한 세분화 사례

1 분석자료 및 변수선택

본 논문에서는 SAS E-miner를 이용하여 자료를 분석하였으며, 해당 자료는 A백화점의 2007년 9월부터 2008년 8월까지 고객의 카드사용 내역을 기반으로 한다. 고객의 카드는 백화점에서 발급되는 인하우스 카드로써, 고객의 거래 데이터 확보가 가능한 카드를 대상으로 한다.

A백화점에서 실시하는 고객 세분화의 목적은 첫 거래 후, 4개월간의 거래패턴을 분석하여, 향후 우수고객이 될 가능성이 있는 고객을 대상으로 구매금액 및 내점 횟수를 높이기 위한 프로모션을 제공하는 것에 있다. 기존의 방법은 구매금액과 내점일수를 기준으로 특정 금액 및 일수를 만족하면 모든 고객에게 일괄적으로 프로모션을 제공하고 있었으나, 정확도가 낮음으로 인한 비용의 낭비에 대한 문제가 제기되었다. 즉, 우수고객의 가능성이 낮은 고객에게도 프로모션이 제공되고 있는 것이 문제점으로 지적되었다. 따라서 본 논문에서는 구매금액 및 내점일수 이외에 다양한 변수를 생성하고 파생하여 의사결정나무 분석에 활용함으로써, 정확한 예측이 가능하도록 하였다. <표 1>은 본 논문을 통해 추가로 생성된 변수의 일부로써 변수에 대한 설명 및 단위를 나타내고 있으며, 각 변수는 고객이 첫 구매 후, 4개월간의 요약 자료이다.

<표 1>에 나타난 변수를 포함하여, 총 40개의 독립변수가 파생되었으며, 각 변수에 대한 선별과정은 다음과 같이 진행하였다.

1단계: 데이터의 분포를 파악하여, 결측치 및 이상치의 처리방법 결정
2단계: 단변량 회귀분석을 실시하여, 각 변수별 설명력 확인
3단계: 인공신경망분석을 실시하여, 각 변수별 중요도 확인
4단계: 변수별 설명력 및 중요도를 고려하여 변수 선별

1) 결측치 및 이상치 처리

결측치 및 이상치가 모두 제거하거나 처리해야 할 대상은 아니며, 그 자체로도 의미가 있는 경우가 있다. 예를 들어, 구매금액이 매우 큰 고객과 같은 이상치의 경우에는 따로 선별하여 관리할 필요가 있기도 하다. 따라서 각 변수별로 변수의 특성을 고려하여 결측치 및 이상치의 존재 여부를 파악하고, 해당 데이터 또는 변수 자체를 제거할 것인지, 평균이나 중앙값 또는 특정 값(예를 들어 0)으로 대체할 것인지, 아니면 그대로 보존하고 분석을 실시할 것인지 결정하고 처리하는 과정을 거쳐야 한다.

쇼핑몰매출액 분포를 살펴보면 모든 값이 0 또는 결측치임을 알 수 있다. 이는 쇼핑몰 판매실적이 없는 것을 의미하며, 쇼핑몰 매출액을 기반으로 파생되는 쇼핑몰매출비율 또한 동일한 결과로 나타난다. 따라서 쇼핑몰매출액 및 매출비율은 분석에서 제외하였다.

<표 1> 분석에 사용된 변수목록의 예

No	한글변수명	변수설명	단위	변수 구분
0	1년후우수고객여부	해당 고객의 1년 후, 등급이 우수고객 이상인지의 여부	-	타겟
1	일반매출	식품을 제외한 구매금액의 합	원	매출
2	일반매출건수	식품을 제외한 총 구매횟수(영수증 건수 기준)	회	
3	식품매출액	식품구매금액의 합	원	
4	식품매출건수	식품을 구매한 총 횟수	회	
5	쇼핑몰매출액	온라인 쇼핑몰에서 이루어진 구매금액의 합	원	
6	쇼핑몰매출비율	(쇼핑몰매출액/전체매출금액)X100	%	
7	일시불구매금액	일시불 결제 금액의 합	원	
8	일시불구매건수	일시불 결제횟수	회	
9	할부구매금액	할부 결제금액의 합	원	
10	할부구매건수	할부 결제횟수	회	
11	할인구매건수	할인 구매한 횟수(전체구매건수 중 정상구매건수 제외)	회	
12	정상구매건수	정상 구매한 횟수(전체구매건수 중 할인구매건수 제외)	회	
13	할인건수비율	(할인구매건수/전체구매건수)X100	%	
14	할인구매금액	전체매출금액 중 정상구매금액을 제외한 금액의 합	원	
15	정상구매금액	전체매출금액 중 할인구매금액을 제외한 금액의 합	원	
16	할인금액비율	전체매출금액 대비 할인구매금액의 비율	%	
17	전체매출금액	전체 구매금액의 합	원	
18	주중구매금액	주중에 구매한 금액의 합(주중 : 월~목)	원	
19	주말구매금액	주말에 구매한 금액의 합(주말 : 금~일)	원	
20	일반구매일수	식품을 제외한 실 구매일수	일	구매 기간
21	식품매출일수	식품구매가 이루어진 총 일수	일	
22	구매일수	일반매출과 식품매출을 포함하여 구매가 이루어진 총 일수	일	
23	구매개월수	일반매출과 식품매출을 포함하고 구매가 이루어진 개월 수	월	
24	구매중MD수	구매한 상품의 중복되지 않은 중MD 가지 수	개	제품 다양성
25	구매소MD수	구매한 상품의 중복되지 않은 소MD 가지 수	개	
26	구매브랜드건수	구매한 상품의 중복되지 않은 브랜드 가지 수	회	
27	캠페인대상선정수	캠페인의 대상으로 선정된 횟수의 합	회	캠페인
28	캠페인성공수	캠페인에 반응한 횟수의 합(쿠폰을 사용 횟수)	회	

한편, 일시불로 구매하지 않은 고객의 경우에는 결측치가 발생하므로, 이와 같은 경우에는 결측치를 0으로 대체할 수 있다. 또한 매우 큰 금액이 소수 존재하므로 연속변수를 그대로 사용하지 않고, 0, 0 이상~100 만원 미만, 100 만원이상~200 만원 미만, 200 만원이상~600 만원미만, 600 만원이상과 같은 구간변수로 변환하여 사용하였다. 이와 같은 전처리 과정을 통해 7개 변수가 제거되고 최종 분석에 활용되는 변수는 33개이다.

2) 변수의 설명력 확인

일반적으로 회귀분석과 같은 선형모형을 활용할 경우에는 변수간의 다중 공선성, 변수의 정규분포 등을 파악하기 위하여 다양한 사전 분석들을 실시하나, 본 논문에서 제시하고자 하는 의사결정나무 분석의 경우 비모수를 기반으로 함으로써, 변수간의 다중공선성 및 정규분포 가정을 필요로 하지 않는다. 그러나 변수가 지나치게 많은 경우에는 분석 속도 및 예측력 향상을 위해 일정부분의 변수제거 과정이 필요하므로 각 독립변수의 종속변수에 대한 설명력을 확인함으로써 R^2 값이 0.4 이하로 낮은 변수는 제거하였다. 40개 변수 중 A백화점의 각 지점별 잔존 변수를 살펴보면 전체 33개 변수 중에, 제1지점이 28개, 제2지점이 26개, 제3지점이 26, 제4지점이 19개인 것으로 나타났다.

3) 변수의 중요도 확인

변수의 중요도는 예측력이 뛰어나다고 평가받고 있는 인공신경망 분석을 통해 평가하였다. 인공신경망 분석기법은 변수간의 교호작용을 고려하지 않아도 되며, 의사결정나무 분석 기법과 마찬가지로 비모수를 기반으로 하기 때문에 인공신경망 기법을 통해 변수들의 중요도를 평가해 봄으로써, 의사결정나무 분석 결과와 인공신경망 기법의 결과를 비교할 수 있도록 하였다.

전처리 과정을 통해 선별된 33개 변수 중에서, 인공신경망 기법에 의해 선택된 상위 20%, 즉 중요도가 높은 상위 7개 변수들을 살펴보면 다음과 같다. 제1지점은 할인금액비율, 할인건수비율, 전체식품구매금액, 식품건수비율, 전체주중구매금액, 전체주말구매금액, 점포구매주기 등이며, 제2지점은 주말건수비율, 거래기간일수, 신규캠페인소요기간일수, 점포구매주기, 캠페인성공수, 캠페인대상선정수, 신규캠페인성공여부 등으로 나타났다. 그 외에도 제3지점은 전체할인금액, 전체할인건수, 할인금액비율, 전체반품구매금액,

전체주중구매건수, 구매일수, 신규캠페인성공여부 등인 것으로 나타났으며, 제4지점은 식품건수비율, 전체주말구매금액, 주말금액비율, 구매일수, 구매개월수, 점포구매주기, 신규캠페인성공여부 등인 것으로 나타났다. 이 결과는 추후 살펴보게 될 의사결정나무 분석결과가 주로 전체매출금액, 또는 구매일수 등이 주요한 변수로 선택되는 것과는 상이한 결과임을 알 수 있다.

이와 같이 각 변수의 설명력과 중요도를 고려하여 최종 선별된 변수의 수는 제1지점이 32개, 제2지점이 27개, 제3지점이 27, 제4지점이 26개인 것으로 나타났다. <표 2>를 통해 살펴볼 수 있듯이, R2에 의한 결과와 인공신경망에 의해 선별된 변수의 중복이 가장 많은 지점은 제2, 제3지점이며, 중복이 가장 적은 지점은 제4지점으로 나타났으며, 최종 선별된 변수를 이용하여 의사결정나무 분석을 실시하였다.

<표 2> 단계별 선별된 변수의 수

지점명	R^2 선별	인공신경망 선별	최종선별
제1지점	28	7	32
제2지점	26	7	27
제3지점	26	7	27
제4지점	19	7	26

2. 의사결정나무 분석결과

최종 선별된 변수를 기반으로 도출된 각 지점별 의사결정나무 분석결과를 살펴보면 <표 3>~<표 6>과 같으며, 그 중 제3지점의 분석결과를 이해하기 쉽게 그림으로 나타내면 <그림 3>과 같다.

<표 3> 제1지점의 의사결정나무 분석결과

실제 총 고객수	1년 후 실제 우수 고객수	우수고객 비중	Rule	변수명	변수값	정분류 고객수	정분류율
14,477	2,577	17.80%	A	전체 매출액	[1600000, 2200000)	152	54.87%
				구매 개월수	>= 3		
				신규캠페인 소요기간일수	< 60		
			B	전체 매출액	[1600000, 2200000)	274	63.87%
				구매 개월수	>= 4		
			C	전체 매출액	[2200000, 3100000)	518	83.15%
				거래 기간일수	>= 80		
			D	전체 매출액	>= 3100000	629	96.62%
합계						1,573	79.44%

<표 3>은 제1지점의 의사결정나무 분석결과로써, 해당 지점의 총 고객수는 14,477명이며, 분석결과 4개의 규칙(Rule)이 발생되었다. 첫 번째 규칙, Rule A에 의하면 전체 매출액이 160만원 이상~220만원 미만이고, 구매 개월 수가 3회, 신규캠페인 소요기간일수가 60일 미만인 고객의 55% 가량인 152명이 우수고객인 것으로 나타났다. 두 번째 규칙, Rule B에 의하면 전체 매출액이 160만원 이상~220만원 미만, 구매 개월 수가 4회 이상인 고객의 64% 가량인 274명이 우수고객인 것으로 나타났다. 또한 세 번째 규칙, Rule C에 의하면 전체 매출액이 220만원 이상~310만원 미만, 거래 기간일수가 80일 이상 고객의 83% 가량인 518명이 우수고객인 것으로 나타났다. 마지막 규칙, Rule D에 의하면 전체 매출액이 310만원 이상 고객의 97% 가량인 629명이 우수고객인 것으로 나타났다.

<표 4> 제2지점의 의사결정나무 분석결과

실제 총 고객수	1년 후 실제 우수 고객수	우수고객 비중	Rule	변수명	변수값	Rule에 의한 우수 타겟	정분류율
8,757	1,462	16.70%	A	전체 매출액	[1500000, 2500000)	155	63.87%
				거래 기간일수	[30, 130)		
				할인 구매건수	>= 11		
			B	전체 매출액	[1500000, 2500000)	223	64.13%
				거래 기간일수	>= 130		
			C	전체 매출액	[2500000, 3500000)	267	84.27%
				점포 구매주기	< 30		
			D	전체 매출액	>= 3500000	383	97.65%
합계						1,028	81.81%

<표 4>는 제2지점의 의사결정분석 결과이며, 해당 지점의 총 고객수는 8,757명으로써 총 4개의 규칙이 발생되었다. 첫 번째 규칙, Rule A에 의하면 전체 매출액이 150만원 이상~250만원 미만, 거래 기간일수가 30일 이상~130일 미만, 정상 구매건수가 11회 이상인 고객의 64% 가량인 155명이 우수고객인 것으로 나타났다. 두 번째 규칙, Rule B에 의하면 전체 매출액이 150만원 이상~250만원 미만, 거래 기간일수가 130일 이상 고객의 64% 가량인 223명이 우수고객인 것으로 나타났다. 또한 세 번째 규칙, Rule C에 의하면 전체 매출액이 250만원 이상~350만원 미만, 점포 구매주기가 30일 미만 고객의 84% 가량인 267명이 우수고객인 것으로 나타났다. 마지막 규칙, Rule D에 의하면 전체 매출액이 350만원 이상 고객의 98% 가량인 383명이 우수고객인 것으로 나타났다.

<표 5> 제3지점의 의사결정나무 분석결과

실제 총 고객수	1년 후 실제 우수 고객수	우수고객 비중	Rule	변수명	변수값	Rule에 의한 우수 타겟	정분류율
6,275	1,040	16.57%	A	전체 매출액	[1300000, 1700000)	26	65.38%
				구매 개월수	>= 3		
				캠페인 대상 선정수	>= 10		
			B	전체 매출액	[1700000, 2000000)	9	66.67%
				구매 개월수	< 3		
				반품 비율	> 20		
			C	전체 매출액	[1700000, 2000000)	140	55.00%
				구매 개월수	>= 3		
			D	전체 매출액	[2000000, 2700000)	220	67.73%
				거래 기간일수	>= 90		
			E	전체 매출액	[2700000, 3900000)	218	83.03%
				거래 기간일수	> 60		
			F	전체 매출액	>= 3900000	255	97.25%
합계						868	78.11%

<표 5>는 제3지점의 의사결정분석 결과이며, 해당 지점의 총 고객수는 6,275명으로써 총 6개의 규칙이 발생되었다. 첫 번째 규칙, Rule A에 의하면 전체 매출액이 90만원 이상~130만원 미만, 정상 구매건수가 13회 이상, 주말 구매 비율이 32% 이상인 고객의 65% 가량인 26명이 우수고객인 것으로 나타났다. 두 번째 규칙, Rule B에 의하면 전체 매출액이 130만원 이상~180만원 미만, 캠페인 성공률이 5이상, 일시불 구매금액이 130만원 이상 고객의 67% 가량인 9명이 우수고객인 것으로 나타났다. 또한 세 번째 규칙, Rule C에 의하면 전체 매출액이 180만원 이상 고객~250만원 미만, 거래 기간일수가 90일 이상 고객의 55% 가량인 140명이 우수고객인 것으로 나타났다. 네 번째 규칙, Rule D에 의하면 전체 매출액이 250만원 이상~350만원 미만, 정상 구매건수가 12회 미만 고객의 68% 가량인 220명이 우수고객인 것으로 나타났다. 다섯 번째 규칙, Rule E에 의하면 전체 매출액이 250만원 이상~350만원 미만, 정상 구매건수가 12회 이상, 점포 구매주기가 25일 미만 고

객의 83% 가량인 218명이 우수고객인 것으로 나타났다. 마지막 규칙, Rule F에 의하면 전체 매출액이 350만원 이상 고객의 97% 가량인 255명이 우수고객인 것으로 나타났다.

<표 6> 제4지점의 의사결정나무 분석결과

실제 총 고객수	1년 후 실제 우수 고객수	우수고객 비중	Rule	변수명	변수값	Rule에 의한 우수 타겟	정분류율
5,761	829	14.39%	A	전체 매출액	[900000, 1300000)	18	66.67%
				할인 구매건수	>= 13		
				주말 구매건수비율	>= 32		
			B	전체 매출액	[1300000, 1800000)	29	65.52%
				캠페인 성공율	>= 5		
				일시불 구매금액	>= 1300000		
			C	전체 매출액	[1800000, 2500000)	238	64.29%
				거래 기간일수	>= 90		
			D	전체 매출액	[2500000, 3500000)	118	63.56%
				할인 구매건수	< 12		
			E	전체 매출액	[2500000, 3500000)	81	91.36%
				할인 구매건수	>= 12		
				점포 구매주기	< 25		
			F	전체 매출액	>= 3500000	183	96.17%
합계						667	76.31%

<표 6>은 제4지점의 의사결정분석 결과이며, 해당 지점의 총 고객수는 5,761명으로써 총 6개의 규칙이 발생되었다. 첫 번째 규칙, Rule A에 의하면 전체 매출액이 90만원 이상~130만원 미만, 정상 구매건수가 13회 이상, 주말 구매 비율이 32% 이상 고객의 67% 가량인 18명이 우수고객인 것으로 나타났다. 두 번째 규칙, Rule B에 의하면 전체 매출액이 130만원 이상~180만원 미만, 캠페인 성공률이 5이상, 일시불 구매금액이 130만원 이상 고객의 66% 가량인 29명이 우수고객인 것으로 나타났다. 또한 세 번째 규칙, Rule C에 의하면 전체 매출액이 180만원 이상 고객~250만원 미만, 거래 기간일수가 90일 이상 고객의 64% 가량인 238명이 우수고객인 것으로 나타났다. 네 번째 규칙, Rule D에 의하면 전체 매출액이 250만원 이상~350만원 미만 정상 구매건수가 12회 미만 고객의 64% 가량인 118명이 우수고객인 것으로 나타났다. 다섯 번째 규칙, Rule E에 의하면 전체 매출액이 250만원 이상~350만원 미만, 정상 구매건수가 12회 이상, 점포 구매주기가 25일 미만 고객의 91% 가량인 81명이 우수고객인 것으로 나타났다. 마지막 규칙, Rule F에 의하면 전체 매출액이 350만원 이상 고객의 96% 가량인 183명이 우수고객인 것으로 나타났다.

<표 3>~<표 6>까지의 분석결과를 살펴보면 제1지점과 제2지점의 특성이 유사하고, 제3지점과 제4지점의 특성이 유사한 것을 알 수 있다. 또한 전체 지점의 평균 정분류율이 80%가량이며, Rule에 의해 선별되는 우수고객의 비중도 60% 가량으로, 정분류율과 고객 비중 모두 높은 편이었다. 앞서 언급한 바와 같이 모든 지점에서 전체매출액이 가장 우선적인 분류기준으로 나타남으로써, 가장 중요한 변수가 전체매출액임을 알 수 있다. 이는 인공신경망 분석 결과와는 다소 상이하다. 의사결정나무분석 결과에서는 전체매출액 외에도 거래기간일수 및 구매일수 등 방문빈도와 관련된 변수들이 주요한 변수로 나타났다.

3. 의사결정나무 세분화 방법과 기존의 세분화 방법 비교

A백화점은 전체매출액과 내점일수를 기반으로 고객을 세분화하여 특정 금액 및 일수를 만족하는 고객들을 대상으로 재방문 및 재구매 유도 캠페인을 실시해왔다. 본 절에서는 의사결정나무 분석 방법의 효과성 및 활용가능성을 평가하기 위하여 <표 7>과 같이 기존의 방법과 비교해 보았다. 정분류율은 전체 고객 대비 실제값과 예측값이 일치한 고객의 수를 의미한다. 즉, 오분류표에서 정분류의 비율을 의미한다. 기존 방법으로 고객을 세분한 결과 정분류율은 평균 41.3%로써 매우 낮았으나, 의사결정나무 분석 방법으로 세분한 결과 정분류율이 지점 평균 78.9%로 매우 높게 나타났다. 즉, 의사결정나무 분석을 기반으로 고객을 세분하였을 때, 2배 가량 정확한 세분화가 가능하다는 것을 의미하며, 이는 무의미하게 버려지는 캠페인 비용이 줄어드는 것을 의미한다.

비록 캠페인 실시 대상인 전체 타겟 고객수가 2배 이상 늘어남으로써, 마케팅 비용에 부담이 가중되기는 하였지만, 비용대비 효율 측면에서 살펴보면 의사결정나무분석에 의해 고객을 세분화하는 것이 비용 대비 효과성이 높음을 알 수 있다. 마케팅 비용을 줄이기 위한 방안으로는 의사결정나무분석 결과에 의해 도출된 4개 또는 6개 규칙들 중에 정분류율이 높은 규칙을 선별하여 캠페인에 우선 적용하게 되면, 비용 부담 문제도 해결할 수 있다.

<표 7> 기존 방법과 의사결정나무의 고객 세분 결과 비교

	기존방법	의사결정나무분석
타겟고객수	2,207(명)	5,117(명)
비타겟고객수	38,145(명)	35,055(명)
정분류율	41%	79%

IV. 결론

본 논문에서는 실제 백화점 사례를 기반으로 의사결정나무 분석을 활용하여 고객을 세분화함으로써 기존에 이루어지던 규칙 시스템 기반의 세분화 및 타깃팅 방법의 문제점을 극복하고자 하였다. 기존의 방법에 비해 의사결정나무분석 방법에 의한 고객 분류의 정확도가 두 배가량 높게 나타남으로써, 보다 효과적임을 알 수 있었다. 또한 의사결정나무분석 결과는 분류 기준이 이해하기 쉬운 규칙의 형태로 나타남으로써, 분석전문가가 아니더라도 일반적인 마케팅 담당자가 실무에 적용 및 활용하는 것이 용이하다.

그러나 분석결과가 아무리 정확하다 하더라도 캠페인에 대하여 고객이 느끼는 매력도가 낮으면 좋은 효과를 기대하기 어렵다. 정확한 고객 세분화와 함께, 목표 고객에게 적절한 캠페인을 제공하기 위한 캠페인 분석이 연계되어야 할 것이다.

본 논문을 통해 기존의 방법 보다 정확도가 높은 고객 세분화 방법을 제시하기는 했지만, 시간의 흐름에도 지속적으로 정확도를 높게 유지할 수 있는지 확인하는 안정성 검증이 이루어지지 못했다. 향후 의사결정나무분석 결과에 의해 실제로 캠페인이 이루어지고, 그에 대한 데이터가 누적되면, 규칙의 안정성에 대해서도 검증해 보고자 한다.

논문접수: 2009. 12. 01
수정보완: 2010. 01. 28
게재확정: 2010. 02. 08

참고문헌

구자용, 박헌진, 최대우(2000), “데이터마이닝에서의 폴리클라스”, *응용통계연구*, 제13권 제2호, 489-503.

김근형, 오성열(2009), "온라인 고객리뷰 분석을 통한 시장세분화에 텍스트마이닝 기술을 적용하기 위한 방법론", *한국콘텐츠학회논문지*, 제9권 제8호, 272-284.

방정혜, Lutz Hamel, Brian Ioerger(2007), "고객관계관리의 시장세분화를 위한 Self-Organizing Maps 재고찰", *한국지능정보시스템학회논문지*, 제13권 제4호, 17-34.

이건창, 권순재, 신경식(2001), "은행고객 세분화를 통한 이탈 고객 관리분석-가계성 예금을 중심으로", *한국지능정보시스템학회논문지*, 제7권 제1호, 177-196.

조홍규(2003), "**인공지능 방법을 이용한 신용평가 모형에 대한 개관**", 나이스채권평가 금융공학연구소.

박명호, 한장희, 김상우, 백운배(2005), **인터넷마케팅**, 명경사.

이두희(2006), **통합적 인터넷 마케팅**, 박영사.

Breiman, L., Friedman, J., Olshen, R. and Stone, C.(1984) "*Classification and Regression Tree*", New York: Chapman and Hall.

Elmer, P. and Borowski, D.(1988), "An expert system approach to financial analysis, The case of S&L bankruptcy", *Financial Management*, 66-75.

Kass, G. V.(1980), "An Exploratory Technique for Investigating Large Quantities of Categorical Data", *Journal of Applied Statistics*, 29, 119-127.

Quinlan, J. R.(1986), "Induction of Decision Trees", *Machine Learning*, 1, 81-106.

Quinlan, J. R.(1993), "*C4.5: Programs for Machine Learning*", Morgan Kaufmann.

Han, J. and Kamber, M.(2001), *Data Mining Concepts and Techniques*, Morgan Kaufmann Publishers.

Abstract

A Case Study on segmentation of Department Store using Decision Tree Analysis

Chae, Kyung-Hee*, Kim, Sang-Cheol**

Segmentation, targeting, and positioning are marketing tools used by a company to gain competitive advantage in the market. For an accurate segmentation, various statistics models or datamining techniques are used. Especially, datamining techniques are introduced in the beginning of the 1980s and solved several marketing problems effectively.

In this paper, we research about datamining technique for segmentation and analyze customer's transaction data of Department Store using Decision Tree Analysis, one of the dataming technique. After that, we discuss effects and advantages of segmentation using Decision Tree.

Key words : Segmentation, Datamining, Decision Tree, STP strategy

* Consultant, Lris consulting co

** Associate Professor, Dept. of Distribution Management, Yuhan University