데이터마이닝 기법(CHAID)을 이용한 효과적인 데이터베이스 마케팅에 관한 연구

김 신 곤*

A Study on the Effective Database Marketing using Data Mining Technique(CHAID)

Sin-Kon Kim*

Abstract

Increasing number of companies recognize that the understanding of customers and their markets is indispensable for their survival and business success. The companies are rapidly increasing the amount of investments to develop customer databases which is the basis for the database marketing activities. Database marketing is closely related to data mining. Data mining is the non-trivial extraction of implicit, previously unknown and potentially useful knowledge or patterns from large data. Data mining applied to database marketing can make a great contribution to reinforce the company's competitiveness and sustainable competitive advantages.

This paper develops the classification model to select the most responsible customers from the customer databases for telemarketing system and evaluates the performance of the developed model using LIFT measure. The model employs the decision tree algorithm, i.e., CHAID which is one of the well-known data mining techniques. This paper also represents the effective database marketing strategy by applying the data mining technique to a credit card company's telemarketing system.

※ 이 논문은 1998년도 광운대학교 교내학술연구비에 의하여 연구되었음.
* 광운대학교 경영정보학과

1. 서 론

데이터베이스 마케팅은 고객의 소비습관이나 행태를 데이터베이스로 축적하고 이를 판매촉진에 활용하는 기법 가운데 하나를 의미한다. 데이터베이스 마케팅은 초기에는 미국, 영국 등의 다국적 유통업체와 대형 수퍼마켓에 의해 도입되었으나 최근에는 국내에도 대형 유통업체와 백화점을 중심으로 빠르게 확산되고 있다. 국내 기업들도 효과적인 직접 마케팅을 하기 위해서는 소비자에 대한 정확한 정보가 필요하다는 점을 인식하고 고객 데이터베이스를 구축하고 이를 활용한 데이터베이스 마케팅이나 텔레 마케팅에 많은 자원을 할당하고 있다.

한 조직이 데이터를 아무리 잘 수집하고 조직화하여 고객 데이터베이스나 데이터웨어하우스를 구축하였다 하더라도 단순히 이를 저장하는 수준으론 조직의 경쟁력 강화와 이익 창출에 아무런 도움이 되지 않는다. 구축되어 있는 고객 데이터베이스나 데이터웨어하우스에서 유용한 정보나 새로운 지식을 발굴하여 활용할 수 있는 수단이 제공되어야만 다양한 이익을 창출하는 완벽한 고객 데이터베이스라 할 수 있다. 이런 상황에서 데이터마이닝은 방대한 규모의 데이터베이스로부터 숨겨진 지식, 예상치 않았던 패턴, 및 새로운 규칙 등을 추출하는 가장 좋은 수단 가운데 하나로 인식되고 있다[김신곤, 1997].

데이터마이닝은 방대한 양의 데이터로부터 의미있는 패턴, 규칙들을 발견하기 위하여 자동적인 혹은 반자동적인 방법으로 데이터를 분석하고 탐색하는 것을 말한다[Adriaans, 1996]. 또한, 1996년 가트너 그룹(Gartrner Group)의 E.Brethenoux, H. Drenser, K.Strange[1996] 그리고 J. Block의 리포트 "Strategic Data Management : Data Warehouse, Data Mining and Business Intelligence : The Hype Stops Here"에서 향상된 데이터 분석의 궁극적인 목적은 가능한 한 직접적인 이익을 얻을 수 있는 의사결정을 하는 것이라고 주장하고 가치사슬을 "데이터에서 정보로 정보에서 지식으로 지식에서 의사결정"에 이르는 경로를 설명함과 동시에 데이터에서 의사결정까지 이르는 프로세스의 주요 정보기술로서 데이터베이스, 인프라구조 기술, 데이터마이닝을 지적하였다.

본 논문은 고객 데이터베이스로부터 텔레마케팅을 수행하였을 경우 높은 응답율이 예상되는 고객의 패턴을 찾아내고, 발견된 패턴을 이용하여 텔레마케팅의 수행 대상이 되는 고객을 선별하는 모델의 개발에 관한 것이다. 텔레마케터에게 이 모델에 의하여 선별된 고객의 리스트를 제공함으로써 텔레마케팅의 성공율을 높이는 효과적인 데이터베이스 마케팅에 관한 연구이다. 즉, 고객 데이터베이스로부터 텔레마케팅 대상 고객을 무작위로 추출하는 것이 아니라 데이터마이닝을 통해 발견된 패턴에 따라 추출된 대상 고객을 텔레마케터에게 할당함으로써 텔레마케터의 업무성과를 향상시키는 것을 목표로 하고 있다.

본 연구를 위하여 A 카드회사의 고객 데이터베이스에 의사결정트리 알고리즘(CHAID)을 적용하여 모델을 개발하였고 개발된 모델은 교차타당성(Cross Validation) 평가를 통하여 검증하였다. 모델의 성과는 리프트(LIFT)를 사용하여 평가하였다[Berry and Linoff, 1997, Deng, 1993, SAS, 1998].

2. 데이터마이닝과 의사결정트리 알고리즘(CHAID)

데이터마이닝은 대용량의 데이터로부터 기업의 경쟁력 확보를 위한 의사결정을 돕는 유용한 정보를 찾아내는 일련의 분석과정이라고 할 수 있는데, 보통 평균값이나 이상치, 결측치 등을 발견하는 탐색과정(Exploration), 자료의 변환을 위한 변환과정(Modification), 모형화 과정(Modeling), 평가 과정(Assessment)의 단계를 거치게 된다.

특정 문제에 적용하는 데이터마이닝 기법이 정해져 있는 것은 아니다. 얻고자 하는 결과나 데이터의 상태 등에 따라 적용할 수 있는 기법은 다를 수 있다[Fayyad 외 2인, 1996]. 데이터마이닝의 기

법에는 일반적으로 통계학에서 사용되는 여러 분석 기법들을 포함하며, 연관규칙(Associations), 클러스터링(Clustering), 의사결정트리(Decision Tree), 그리고 신경망 (Neural netwroks)과 같은 기법들이 있다[Brachman and Anand, 1994].

본 논문에서 사용하고 있는 알고리즘인 의사결정트리는 많은 요인들을 토대로 의사결정을 내릴 필요가 있을 때, 어떤 요인이 고려 대상이 되는지를 구별하는데 도움을 준다[Mehta 외 2인, 1996].

분류에 관한 연구는 과거 통계(Statistics), 신경망(Neural Network), 의사결정트리(Decision Tree) 등의 분야에서 연구되어 왔다. 의사결정트리는 다른 분류기법과 비교해 볼 때 상대적으로 빠르고 간단하며, 이해하기 쉬운 규칙으로 전환될 수 있기 때문에 [Imielinski and Mannila, 1996] 본 논문은 데이터마이닝 기법으로서 의사결정트리 가운데 하나인 CHAID 알고리즘을 사용하고 있다.

1975년 J.A. Hatigan에 의해 처음 발표된 CHAID (Chi-Square Automatic Interaction Detection) 알고리즘은 카이제곱-검정(이산형 목표변수), 또는 F-검정(연속형 목표변수)을 이용하여 다지분리(Multiway Split)를 수행하는 알고리즘으로 1963년 J.A.Morgan과 J.N. Sonquist이 발표한 AID (Automatic Interaction Detection) 시스템에서 유래되었다. AID에서 암시하고 있는 것과 같이 CHAID는 원래 변수들 간의 통계적 관계를 찾는 것이 그 목적이었다. 변수들간의 통계적인 관계는 다시 의사결정트리를 통해 표현될 수 있었으므로 이 방법은 분류기법(Classification Technique)으로써 사용할 수 있다[Thearling, 1995].

CHAID는 변수의 성격이 범주형 데이터이고 예측 변수(Predictor Variable)와 결과 변수간의 관계를 찾아야 할 때 가장 유용하다[Pyle, 1998]. 다른 의사결정트리와 마찬가지로 CHAID 알고리즘은 두개 이상의 자식노드(Child Node)로 트레이닝 데이터를 쪼개기 위한 입력변수(Input Variables)를 찾는다. 즉, CHAID는 분리기준(Split)를 찾는 것을 시발점으로 하여 자식노드는 특정 변수가 갖고 있는 결과변수의 확률이 각 노드마다 다르게 하는 방식으로 선택된다. CHAID는 데이터의 집합을 검색하여 예측변수의 예측치로서 가장 유의성이 높은 변수를 결정한다.

고객 데이터베이스에서 어떤 고객이 직접 우편(Direct Mail)에 가장 응답할 가능성이 높은가를 예측하려 한다면, CHAID 알고리즘은 최상의 예측변수로서 결정된 변수를 이용하여 응답률에서 가장 큰 차이를 갖는 두개 이상의 구분된 집단으로 나누고 그 결과를 트리로 나타낸다[Deere, 1996].

CHAID 알고리즘은 카이제곱 통계량을 통해 비율이 유지되는 정도를 파악하는데, 여러 변수 중 비율을 가장 많이 깨뜨리는 변수가 결국 결과변수에 영향을 가장 많이 미치는 변수가 된다. 비율이 깨진 정도는 카이제곱에서 r x c 분할표(Contingency Table)로부터 계산된다. 이 때, Pearson의 카이제곱 통계량은

$$x^2 = \sum \frac{(fo - fe)^2}{fe} \quad fo : 관찰치 \quad fe : 예측치$$

과 같이 정의되며, 이 통계량은 자유도가 (r - 1)(c - 1)인 카이제곱 분포를 따른다. 카이제곱 통계량이 자유도에 비해 매우 작다는 것은 입력변수의 각 범주에 따른 결과변수의 분포가 동질적이라는 것을 의미하며, 입력변수가 결과변수의 분류에 영향을 주지 않는다고 말할 수 있다. 자유도에 대한 카이제곱 통계량의 크고 작음은 p-값으로 표현될 수 있는데, 카이제곱 통계량이 자유도에 비해서 작으면 p-값은 커지게 된다. 결국 노드는 p-값이 가장 작은 변수를 기준으로 가지가 형성되는 것이다[이성근 외 2인, 1996].

3. 분류모델의 성과 측정

분류모델의 성과를 비교 평가하는 가장 일반적인 방법은 리프트(LIFT) 측정치를 이용하는 것이다. 리프트는 분류모델을 사용하여 모집단으로부터 표본을 추출할 때 특정 클래스가 모집단과 표본에 포함되어 있는 비율의 변화를 측정하는 것이다[Berry and Linoff, 1997].

LIFT = P (클래스t 1 표본) / P (클래스t 1 모집단)

리프트(LIFT)는 직접 마케팅(Direct Marketing) 산업에서 유래된 것으로 마케팅 반응 모델(Marketing Response Model)에 쉽게 적용될 수 있다. 예를 들어, 고객에게 직접 우편(Direct Mailing)을 보냈을 때 누가 반응할 가능성이 가장 높은가를 예측하는 분류모델을 개발한다면 개발된 분류모델은 각각의 대상고객에 대하여 응답 또는 무응답으로 예측을 한다. 물론, 이러한 예측이 실제 결과와 항상 일치하는 것은 아니다. 그러나 이 분류모델이 좋은 모델이라면 이 모델에 의하여 추출한 표본(Biased Sample)에 포함되어 있는 응답 건수의 비율은 전체 평가 모집단에 포함되어 있는 실제 응답 건수의 비율보다 높을 것이다. 만약 평가 모집단이 5%의 응답비율을 보이고 있는데 반하여 분류모델에 의하여 선택한 표본은 50% 응답비율을 나타내고 있다면 그 모델의 리프트는 10이다 (50/5 = 10)[Agrawal and Psaila, 1995].

4. 데이터마이닝 기법을 이용한 분류모델의 개발

4.1 A 카드 회사의 텔레마케팅 시스템

현재 14명으로 구성된 A 카드 회사의 텔레마케터의 주 업무는 카드 발급 후 카드를 사용하지 않았거나 연회비를 납부하지 않은 고객을 대상으로, 회원 유지를 목표로 텔레마케팅을 실시하는 것이다. 현재 거의 모든 텔레마케터의 작업들은 과거의 영업방식을 그대로 답습하거나, 특정 몇몇 텔레마케터들이 가지고 있는 경험적 직관에 의해서 진행되고 있는 것이 현실이다.

현재의 텔레마케팅 시스템이 가지고 있는 문제점은 개개인의 텔레마케터에게 할당된 데이터 처리양이 많기 때문에 특히 업무가 몰리는 월말 결산과 같은 때는 전체적으로 텔레마케팅을 수행하는데 있어서 집중력이 떨어져 마케팅 효과가 별로 없을 뿐만 아니라 텔레마케팅에 응답할 가능성이 높은 고객도 이로 인하여 간과되는 것이 보통이다.

이러한 이유로 A 카드 회사의 텔레마케팅 부서는 보다 효율적인 업무 처리와 효과적인 텔레마케팅을 위해 체계적이고 과학적인 텔레마케팅 방안을 강구하게 되었다.

4.2 데이터의 생성 및 사전처리

본 연구를 위하여 A 카드 회사의 고객 데이터베이스를 활용하였다. 고객 데이터베이스에 포함되어 있는 레코드 수는 충분히 확보할 수 있었으나 연구에 유용하리라 생각되는 변수 또는 데이터 속성을 가지고 있는 레코드를 충분히 확보하지는 못 하였다. 그나마 확보한 변수마저도 그 값이 없는 경우가 종종 발생하였다. 즉, A 카드 회사의 시스템 구조로 인한 문제와 고객의 세부 정보를 추출할 수 없다는 점 때문에 데이터를 확보하였다고 하더라도 업무 특성상 그 값을 알 수 없는 경우가 많았다.

고객 데이터베이스로부터 과거 일년 동안 텔레마케팅을 실시한 고객 중 카드발급 후 14개월간 미사용 고객을 추출하였다.

고객에 대한 변수는 <표 1>에서 보는 바와 같이 인구통계적 정보, 직접우편 (Direct Mailing) 발송 여부, 쿠폰발송 여부, 자동이체 신청 여부와 카드유치에 관한 정보 등 15개 필드를 포함하고 있다. 그러나, 필드에 포함되어 있는 데이터 중 20% 이상의 Null 데이터 값을 가지고 있는 레코드는 대상 변수 항목에서 제외시켰다[Famili 외 2인, 1996].

이렇게 추출된 데이터는 총 33,675건이며, 이 가운데 텔레마케팅을 수행한 후 실제로 응답한 고객, 즉 1년 이내에 카드를 사용한 고객의 데이터는 1,610건으로 전체 데이터 셋의 반응 비율은 4.78% (1,610 / 33,675 × 100 = 4.78%) 이다. 모델을 개발하고 검증하기 위하여 33,675 건의 전체 데이터 셋을 모델을 만들기 위한 트레이닝 셋(Training Set)과 만들어진 모델을 테스트하기 위한 테스트 셋(Test Set)으로 나누고 교차 타당성 평가(Cross

〈표 1〉 입력 데이터 셋의 속성 및 형식

데이터 속성 이름	데이터 코딩 및 형식	비 고
통화일자	Date	본인과의 통화일자
통화약속여부	Y : 1, N : 0	통화 약속일
데이터 구분	14 : 1, 90 : 2, 60 : 3, 휴면 : 4	카드 기간 구분
나이	Number	나이
생월	Date	고객 태어난 월
우편번호	Char(6)	서울지역은 區까지 구분. 지방은 동, 구 구분하지 않음
자동이체	신청 : 1, 취소 : 2, 미신청 : 3	자동이체 신청 여부
카드종류	그린 : 1, 골드 : 3	카드구분
유치경로	유치경로	유치경로
DM 발송	Y : 1, N : 0	DM 발송여부
재발급	Y : 1, N : 0	카드 재발급(분실 및 카드 사용 기간 연장)
연회비 면제	Y : 1, N : 0	2년 연회비 미납회원 중 카드 1년 연회비 면제
쿠폰발송	Y : 1, N : 0	각종 쿠폰 발송 여부
성별	M : 1, F : 2	성별
TM 후 사용 여부	Y : 1, N : 0	카드 사용 여부(타겟 필드)
발급월	Date	최초 카드 발급월
국적	N : 내국인, F : 외국인	국적

Validation) 방법을 사용하였다.

14개월 미사용 고객 데이터의 트레이닝 셋은 전체 데이터 셋의 70%로 총 23,572건(33,675 × 0.7= 23,572)이다. 이 가운데 텔레마케팅 수행 후 실제 반응을 보인 데이터 건은 1,133건, 무반응 데이터 건은 22,439건으로, 트레이닝 셋의 반응 비율은 4.81%(1,133 / 23,572 × 100 = 4.81%)이다. 테스트 셋은 전체 데이터 셋의 30%를 차지하는 10,103건(33,675 × 0.3= 10103)이다. 이 가운데 텔레마케팅 실행 후 실제 반응을 보인 데이터 건은 477건, 무반응 데이터 건은 9,626건으로, 반응 비율은 4.72%(477 / 10,103 × 100 = 4.72 %)이다.

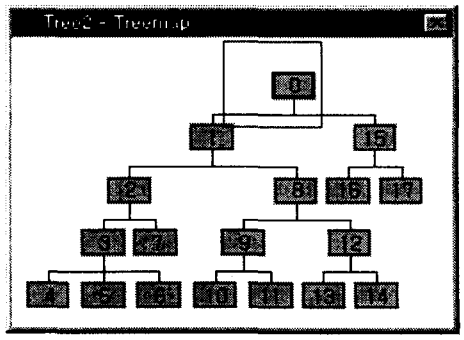
〈표 2〉 데이터 셋 요약표

	반응 건수	반응율 (%)	무반응 건수	무반응 율(%)	합계	전체 비율(%)
전체 데이터 셋	1,610	4.78	32065	95.22	33,675	100
트레이닝 셋	1,133	4.81	22,439	95.19	23,572	70
테스트 셋	477	4.72	9,626	95.28	10,103	30

모델 개발을 위한 데이터마이닝을 수행하기 위해 필터링, 중복 데이터 처리, 비대칭 분포 처리, 그리고 결측치 처리와 같은 사전처리(preprocessing) 작업을 수행하였다.

4.3 분류 모델의 개발

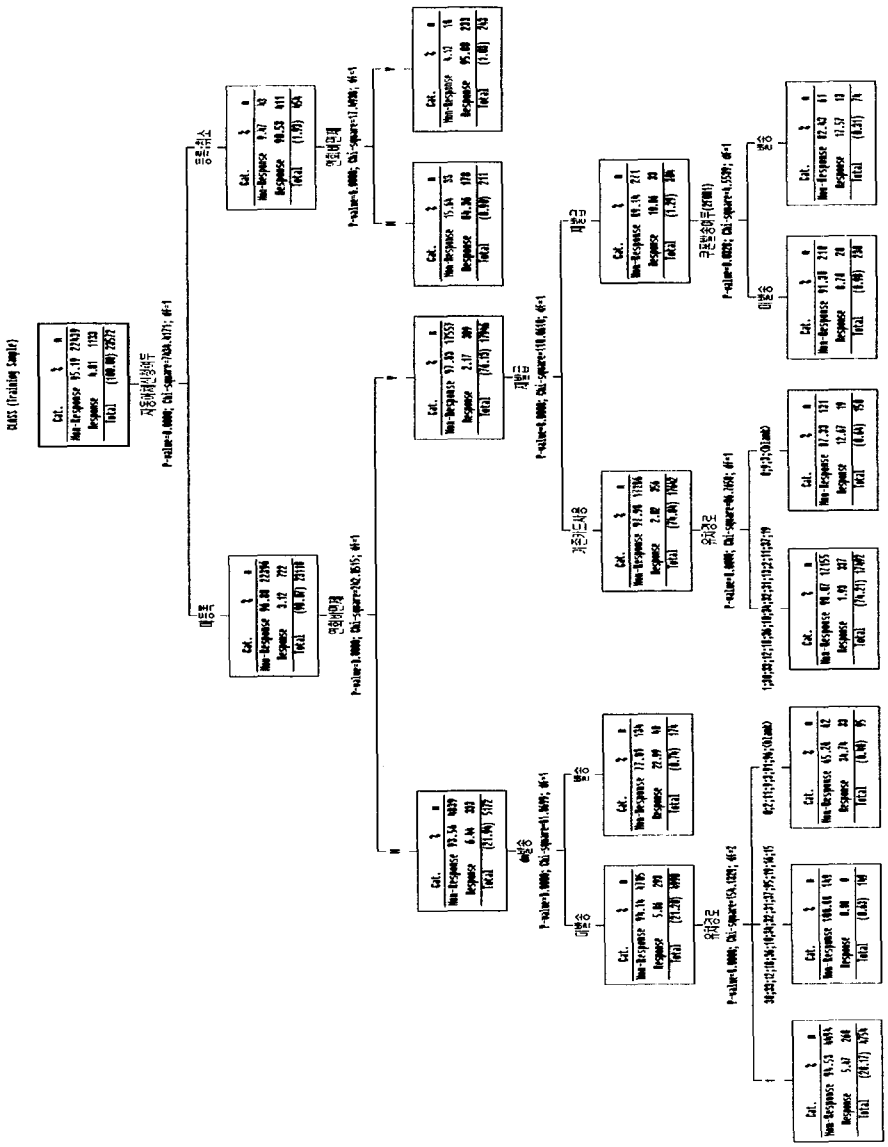
텔레마케팅을 수행할 경우 반응 가능성이 높은 고객을 선별하기 위한 분류모델(Classification Model)을 개발하기 위하여 데이터마이닝 기법 가



(그림 1) 트레이닝 셋의 트리 구조도(Tree Map)

운데 하나인 CHAID 트리 생성 알고리즘을 사용하였다. CHAID 알고리즘을 트레이닝 셋에 적용한 결과 (그림 1)과 같은 트리 구조를 얻을 수 있다. 트리 구조도를 해석하고 설명의 편의를 위하여 각각의 노드에 번호를 부여하였다.

일반적으로 트리 생성 알고리즘에서는 언제 노드의 확장을 멈춰야 하는가를 결정하는 정지 법칙(Stopping Rule)이 필요하다[42,43]. 트레이닝 셋으로부터 트리를 생성하기 위하여 CHAID 알고리즘의 정지 규칙을 적용할 때 최대 노드 깊이(Node Depth)는 4로 설정하였다. 부모 노드(Parent Node)와 자식 노드를 형성하기 위한 필요조건으로 최소 데이터 수를 부모 노드는 100건, 자식 노드는 50건으로 각각 설정하였다. 또한 의사결정 트리 노드



(그림 2) 트레이닝 셋의 의사결정트리 분류모델

의 최적화를 위하여 입력 변수(데이터 속성)와 결과 변수(반응 여부)간의 상관관계가 높은 변수를 뿌리노드에 가장 가까운 입력 변수로 선정하였다. 의사결정 트리를 형성할 때 노드의 쪼개짐(Split)과 결합(Merge)을 결정하기 위한 유의수준은 0.05로 하였다.

개발된 분류모델은 (그림 2)에서 보는 바와 같이 뿌리 노드를 포함하여 18개의 노드로 표현되었으며 트리의 모든 노드는 텔레마케팅을 수행한 후 카드를 다시 사용하거나, 연회비를 입금한 반응 고객과 그렇지 않은 무반응 고객으로 구분하여 각각의 분포를 보여주고 있다.

(그림 2)의 분류모델을 살펴보면 뿌리노드로부터 전체 23,572명 중 텔레마케팅의 수행후 보인 반응율은 4.81%임을 알 수 있다. 다음으로 오른쪽의 두개 마지막 노드의 반응율은 각각 95.88%와 84.36%로서 매우 높은 것을 알 수 있다. 이 노드는 전체집단 가운데 자동이체신청을 하였거나 나중에 자동이체신청을 취소한 집단으로서 연회비를 면제받은 집단과 받지 못한 집단으로 분류되어 있다.

<표 3>은 트레이닝 셋의 정보 이익 요약표(Information Gain Summary)이다. 정보이익 요약표는 목표변수의 각 개체들이 각 마디에서 어떻게 분포되고 있는지를 알려주며, 이를 통해 기존 마디의 병합과 새로운 마디의 쪼개짐에 대한 정보를 제공하여 주고 있다.

- Node : 트리 구조도에 나타난 노드의 번호
- Node:n : 노드에 속하는 개체의 건수
- Node:% : 노드에 속하는 개체의 건수 / 전체 개체의 건수
- Resp:n : 노드에 속하는 목표 변수(응답 변수)의 건수
- Resp:% : 노드에 속하는 목표 변수의 건수 / 전체의 목표변수의 건수
- Gain(%) : 노드에 속하는 목표변수의 건수 / 노드에 속하는 개체의 건수
- Lift : 노드에서의 목표변수의 비율 / 전체에서의 목표변수의 비율

4.4 분류모델의 검증

의사결정트리 알고리즘을 이용하여 개발된 분류모델이 얼마나 타당성을 가지고 있는지를 평가하는 것은 매우 중요하다. 개발된 분류모델을 테스트 셋에 적용하여 봄으로써 모델의 적합도, 타당성 및 그 성능을 평가할 수 있다. 트레이닝 셋에서 만들어진 모델을 테스트 셋에 적용하여 동일하거나 거의 유사한 결과를 나타낸다면, 이것은 모델이 데이터를 잘 표현하고 있다는 것을 의미하며 현재 데이터와 유사한 성격을 가진 다른 데이터에 대해서도 유사한 결과를 나타낼 것이라고 유추할 수 있다[Murthy, 1995].

〈표 3〉 트레이닝 셋 정보 이익 요약표(Information Gain Summary)

Node	Node n	Node %	Cumul. N	Cumul. N %	Resp:n	Resp:%	Cumul. Resp:n	Cumul. Resp:%	Gain %	Cumul. Gain %	Node Lift	Cumul. Lift
17	243	1.03	243	1.03	233	20.56	233	20.56	95.88	95.88	19.93	19.93
16	211	0.90	454	1.93	178	15.71	411	36.28	84.36	90.53	17.54	18.82
6	95	0.40	549	2.33	33	2.91	444	39.19	34.74	80.87	7.22	16.81
7	174	0.74	723	3.07	40	3.53	484	42.72	22.99	66.94	4.78	13.92
14	74	0.31	797	3.38	13	1.15	497	43.87	17.57	62.36	3.65	12.96
11	150	0.64	947	4.02	19	1.68	516	45.54	12.67	54.49	2.63	11.33
13	230	0.98	1177	4.99	20	1.77	536	47.31	8.70	45.54	1.81	9.47
4	4754	20.17	5931	25.16	260	22.95	796	70.26	5.47	13.42	1.14	2.79
10	17492	74.21	23423	99.37	337	29.74	1133	100.00	1.93	4.84	0.40	1.01
5	149	0.63	23572	100.00	0	0.00	1133	100.00	0.00	4.81	0.00	1.00
합계	23572				1133							

정보이익 요약표(Information Gain Summary)는 분류모델이 형성한 각 노드의 타당성을 평가하기 위해 사용될 수 있다.

트레이닝 셋의 정보 이익 요약표 <표 3>과 테스트 셋의 정보 이익 요약표 <표 4>에 나타난 Resp:% 열과 Gain(%) 열의 수치를 비교하여 보면 매우 유사한 결과를 보여주고 있다.

분류모델의 적합도를 판단하고 예측력을 쉽게 파악할 수 있는 다른 방안으로 오차 분석표(Misclassification Matrix)가 사용될 수 있다. 오차 분석표는 트레이닝 셋과 테스트 셋의 각각에 대하여 실제 데이터 수와 예측된 데이터 수의 관계를 보여주고 있다.

분류모델의 오차분석표에서 대각(Diagnol)에 존재하는 도수(Frequency)는 실제 범주와 예측범주가 일치하는 즉, 제대로 예측한 개체의 수이고 비대각(Off-Diagnol)에 존재하는 도수는 예측이 어긋난 개체의 수라고 할 수 있다.

<표 5>의 오차 분석표에서 보는 것과 같이 트레이닝 셋에 대한 위험 추정치(Risk Estimate)는 실제로 무응답한 건수를 응답으로 잘못 예측한 43건의 데이터와 실제 응답한 데이터를 무응답으로 잘못 분류한 722건의 합계 765건을 트레이닝 셋의 총 데이터 건수인 23,572로 나눈 0.0324538이고, 위험 추정치의 표준 오차(Standard Error)는 0.00115417이다.

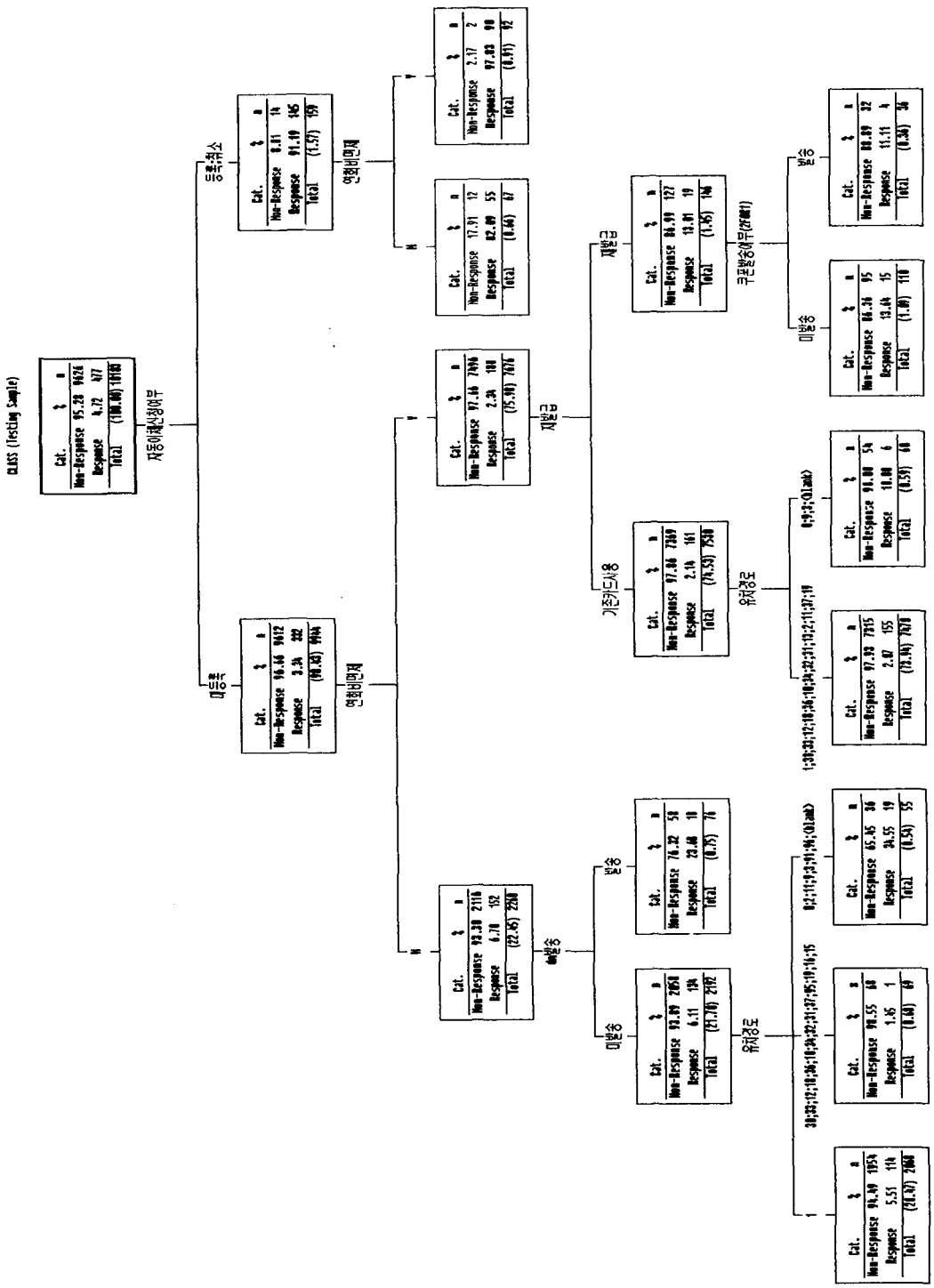
트레이닝 셋의 총 23,572건 중 올바르게 분류된 데이터는 22,807건(22,396 + 411 = 22,807)으로 트레이닝 셋에 대한 분류 정확성은 96.76%(22,807 / 23,572 × 100 = 96.76%)를 나타내고 있다. 즉, 14개월 미사용 트레이닝 셋의 오차 분석표는 이 분류모델이 트레이닝 셋의 96.76%를 올바르게 분류하고 있는 것을 보여주고 있다. 따라서 이 분류모델에 의하면 텔레마케팅에 대하여 응답한 고객을 무응답으로 분류하거나 무응답한 고객을 응답으로 분류할 가능성은 매우 적은 것으로 나타나 있다.

〈표 4〉 테스트 셋 정보 이익 요약표(Information Gain Summary)

Node	Node n	Node %	Cumul. N	Cumul.n %	Resp:n	Resp:%	Cumul. Resp:n	Cumul. Resp:%	Gain %	Cumul. Gain %	Node Lift	Cumul. Lift
17	92	0.91	92	0.91	90	18.87	90	18.87	97.83	97.83	20.73	20.73
16	67	0.66	159	1.57	55	11.53	145	30.40	82.09	91.19	17.39	19.32
6	55	0.54	214	2.12	19	3.98	164	34.38	34.55	76.64	7.32	16.24
7	76	0.75	290	2.87	18	3.77	182	38.16	23.68	62.76	5.02	13.30
14	110	1.09	400	3.96	15	3.14	197	41.30	13.64	49.25	2.89	10.43
11	36	0.36	436	4.32	4	0.84	201	42.14	11.11	46.10	2.35	9.77
13	60	0.59	496	4.91	6	1.26	207	43.40	10.00	41.73	2.12	8.84
4	2068	20.47	2564	25.38	114	23.90	321	67.30	5.51	12.52	1.17	2.65
10	7470	73.94	10034	99.32	155	32.49	476	99.79	2.07	4.74	0.44	1.01
5	69	0.68	10103	100.00	1	0.21	477	100.00	1.45	4.72	0.31	1.00
합계	10103				477							

〈표 5〉 오차 분석표(Misclassification Matrix)

Misclassification Matrix							
		Actual Category			Actual Category		
		Non-Response	Response	Total	Non-Response	Response	total
Predicted Category	Non-Response	22396	722	23118	9612	332	9944
	Response	43	411	454	14	145	159
	Total	22439	1133	23572	9626	477	10103
		Learning Sample			Testing Sample		
Risk Estimate		0.0324538			0.0342473		
SE of Risk Estimate		0.00115417			0.00180934		



(그림 3) 테스트 셋의 의사결정트리 분류모델

테스트 셋에 대한 위험 추정치(Risk Estimate)는 실제로 무응답한 건수를 응답으로 잘못 예측한 14건의 데이터와 실제 응답한 데이터를 무응답으로 잘못 분류한 332건의 합계 346건을 테스트 셋의 총 데이터 건수인 10,103로 나눈 0.0342473이고, 위험 추정치의 표준 오차(Standard Error)는 0.00180934로서 트레이닝 셋의 위험 추정치와 표준오차가 큰 차이가 없다.

또한 테스트 셋 10,103건 중 올바르게 분류된 데이터는 9,757건(9,612 + 145 = 9,757)으로 테스트 셋에 대한 분류 정확성은 96.57%(9,757 / 10,103 × 100 = 96.57%)이므로 트레이닝 셋에 대한 분류 정확성 96.76%와 거의 차이를 보이지 않고 있다. 즉, 14개월 미사용 테스트 셋의 오차 분석표는 이 모델이 테스트 셋에 대하여 96.57%의 높은 정확성을 나타내고 있으며, 정확도가 트레이닝 셋의 정확도 96.76%와 비슷하여 매우 안정적임을 보여주고 있다.

4.5 분류모델의 평가

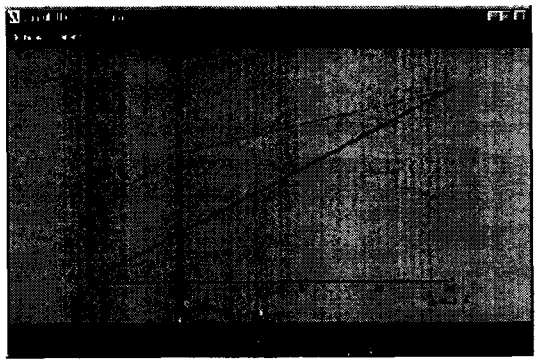
(그림 3)은 트레이닝 셋에서 개발된 모델을 테스트 셋에 적용시킨 결과, 형성된 의사결정트리 분류모델이다. <표 4>는 테스트 셋의 분류모델의 마지막 노드(Leaf Node) 가운데 4, 6, 7, 11, 13, 14, 16, 17번의 8개 노드가 전체 테스트 셋의 반응율 즉, 뿌리 노드의 고객 반응율인 4.72% 이상의 반응율을 나타내고 있음을 보여주고 있다.

개발된 분류모델에 의하여 텔레마케팅을 위한 대상고객을 선별할 경우, 정보 이익(Gain %)이 가장 큰 노드의 고객부터 우선적으로 텔레마케팅의 대상고객에 포함시켜야 성공 가능성이 가장 높으며 이때의 반응률은 97.83%에 이르고 있다. 텔레마케팅 대상고객의 수가 늘어남에 따라 정보이익이 그 다음으로 큰 노드가 차례로 그 대상에 포함되어야 한다.

<표 4>에 의하면 텔레마케팅의 대상고객으로는 정보이익이 97.83%로 가장 높은 노드 17이 가장 먼저 포함되어야 하며, 이 노드의 92건의 대상고객에는 텔레마케팅을 실시하여 반응을 보인 고객이 90건이 포함되어 있어 노드 17의 반응률은 97.83%임을 의미한다. 또한 노드 17의 리프트는 20.73(97.83 / 4.72 = 20.73)임을 나타내고 있다. 이것은 이 분류모델에 의하여 선택된 노드 17의 고객에게 텔레마케팅을 실시할 경우 대상고객을 무작위로 추출하여 실시하였을 때 보다 20배 이상의 성공 가능성이 높다는 것을 의미한다.

노드 17에 포함되어 있는 92건 보다 많은 텔레마케팅 대상 고객수가 필요하다면 당연히 그 다음으로 정보이익이 큰 노드 16에 속해 있는 고객 67건을 추가적으로 포함시켜야 할 것이다. 이때 대상고객의 수는 159건(92 + 67 = 159)으로서 그 가운데 반응을 보인 고객의 수는 145건(90 + 55 = 145)이 포함되어 있어 반응률은 91.19%(145 / 159 * 100= 91.19%)이다. 또한 리프트는 19.32(91.19 / 4.72 = 19.32)로서 노드 17과 노드 16의 고객에게 텔레마케팅을 실시할 경우 대상고객을 무작위로 추출하여 실시하였을 때 보다 19.32배 이상의 성공 가능성이 높다는 것을 의미한다.

이와 같이 대상고객의 수를 증가시킬 필요가 있을 때 추가 대상이 되는 노드는 <표 4>에서 알 수 있는 바와 같이 반응율이 4.72% 이상을 나타내는 8개 노드 즉, 4, 6, 7, 11, 13, 14, 16, 17번의 8개 노드만이 의미 있는 정보로 해석할 수 있다. 8개 노드의 전체 데이터 수는 2,564건(8개 노드의 데이터 누적 건수)이고, 이 가운데 텔레마케팅 실행에 응답한 건수는 321건(8개 노드의 반응 데이터 누적 건수)이므로 8개 노드의 총 응답률은 12.52%



(그림 4) 분류모델에 의한 타겟 텔레마케팅과 무작위 추출에 의한 텔레마케팅의 효율성 비교

(321 / 2564 × 100 =12.52%), 리프트는 2.65이다.

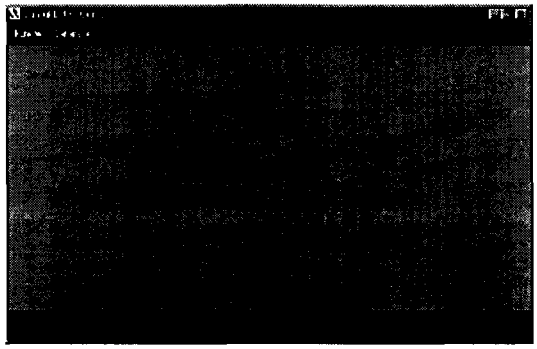
테스트 셋에 대하여 분류모델에 의한 타겟 마케팅(Target Marketing)을 수행하였을 경우와 무작위 추출에 의한 매스 마케팅(Mass Marketing)을 수행하였을 경우의 텔레마케팅의 효율성을 그래프로 나타낸 것이 (그림 4)이다.

(그림 4)의 45도 대각선은 무작위로 테스트 셋에서 추출하여 텔레마케팅을 실시하였을 경우 예상되는 반응율을 나타내는 것이고, 그 위쪽의 선은 분류모델을 사용하였을 경우의 효율성을 나타내 주고 있다. 즉, 45도 대각선과 그 위쪽선의 차이가 나는 부분은 분류모델에 의한 타겟 텔레마케팅을 실시하므로서 얻어지는 정보 이익, 또는 효율성의 차이라고 볼 수 있다. 따라서 텔레마케팅 시스템 운영자는 (그림 4)로부터 얻을 수 있는 정보 이익에 관한 정보를 대상고객의 크기를 결정하는 한가지의 요소로 고려할 수 있다.

테스트 셋에 대한 분류모델의 효용성은 <표 4>의 정보이익 요약표와 (그림 4)를 통하여 실제적인 의미를 살펴볼 수 있다. 예를 들어, <표 4>의 결과로부터 정보이익(Gain %)이 높은 상위 25.38%에 해당하는 고객을 관리하는 것이 고객 전체를 관리하는 것에 비해서 2.65배의 효율을 얻을 수 있다는 것을 알 수 있으며, 이때 상위 25.38%인 2,564건의 텔레마케팅을 실시하여 전체 반응을 보인 477건의 67.3%인 321건으로부터 반응을 얻어 낸 것을 알 수 있다.

<표 4>에 의하면 정보이익이 높은 노드의 순서대로 대상고객에 포함되어 전체 데이터의 반응률인 4.78%보다 높은 노드 4 까지 대상고객에 포함될 경우, 총 대상고객 수는 2,564건으로 전체 테스트 셋 10,103 건의 25.38%에 불과하다. 그러나 이들을 대상으로 텔레마케팅을 실시할 경우 테스트 셋에 포함되어 있는 477건의 총 반응 건수 가운데 67.30%에 해당하는 321건의 반응을 기대할 수 있다.

리프트는 표본수의 함수이다. 즉, 전체 모집단에서 대상고객으로 선택하는 표본수가 늘어남에 따라 리프트 값은 떨어질 수밖에 없다[28]. (그림 5)는 분류모델이 테스트 셋으로부터 대상고객의 추출 표본수가 증가함에 따라 리프트의 값이 적어지는 현상을 나타내고 있다. 따라서 노드 16의 대상고객이 노드 17에 추가적으로 포함되면 대상고객의 표본수가 늘어나게 되고, 이때 리프트는 20.73에서 19.32로 떨어진다.



(그림 5) 테스트 셋의 리프트 차트(Lift Chart)

5. 결 론

A 카드회사의 텔레마케팅 시스템에서 사용하고 있는 고객 데이터베이스에 데이터 마이닝 기법 가운데 하나인 CHAID 알고리즘을 이용하여 텔레마케팅의 대상고객을 선별하는 분류 모델을 개발하고 모델의 적합도와 성능을 교차검증(Cross Validation) 방법을 통하여 검증하였으며, 그 성과를 평가하였다.

모델은 96% 이상의 높은 정확성을 보였다. 대상고객 수의 결정은 텔레마케팅에 따른 비용과 수익, 리프트를 이용하여 결정할 수 있다. 또한 정보이익(Gain %)과 같은 지표를 통하여 고객 데이터베이스에서 우선적으로 대상고객에 포함되어야 할 고객을 결정할 수 있었다. 모델의 구조가 트리형식을 가지고 있어 모델을 SQL문이나 IF문과 같은 상위단계 언어로 쉽게 표현할 수 있어 이를 토대로 고객 데이터베이스에서 대상고객을 추출하는 것이 매우 용이하였다.

모델의 성과를 측정하기 위하여 리프트(Lift)를 사용하였다. 이 분류모델을 사용할 경우, 추출 대상고객의 수에 제한이 없다면 무작위 추출에 의한 경우와 비교하여 최고 20배 이상, 최소 3배 이상의 텔

레마케팅의 효율성을 높일 수 있음을 확인하였다.

참 고 문 헌

[1] 김신곤, "데이터 마이닝과 지식발견", 한국 전문가시스템 학회, 춘계학술대회 논문집, 1997.

[2] 나민영, "데이터 마이닝' 대규모 지식데이터베이스에서 유용한 지식 추출하는 기법", 1998.

[3] Adriaans, Pieter, Dolf Zantinge, *Data Mining*, Addison Wesley, 1996.

[4] Agrawal, R., and Psaila, G. "Active Data Mining," In Proceedings of the first Internaltional Conference On Knowledge Discovery and Data Mining(KDD-95), 3-8, 1995.

[5] Agrawal, Rakesh, Ashish Gupta, Sunita Sarawagi, "Research Report : Modeling Multidimensional Databases," IBM Almaden Research Center.

[6] B. de la Iglesia, J.C.W. Debuse and V.J. Rayward-Smith, "Discovering Knowledge in Commercial Databases Using Modern Heuristic Techniques," KDD-96, 1996.

[7] Berry, Michael J. A., Gordon Linoff, *Data Mining Techniques for Marketing, Sales, and Customer Support*, Wiley Computer Publishing-John Wiley & Sons, Inc, 1997.

[8] Brachman, Ronald J. Tom Khabaza, Willi Kloesgen, Gregory Piatetsky-Shapiro, and Evangelos Simoudis, "Mining Business Databases," *Communications of the ACM*, November, Vol 39, No.11. 1996.

[9] Brachman, Ronald J., Tej Anand,"The Process of Knowledge Doscovery in Databases : A First Sketch," AAAI-94 Workshop on Knowledge Discovery in Databases, KDD-94, 1994.

[10] Brethenoux, E., H. Drenser, K.Strange, J. Block, "Strategic Data Management: Data Warehouse, Data Mining and Business Intelligence : The Hype Stops Here," GartnerGroup Report, SDM:R-300-105, October 28, 1996.

[11] Deng, Stephen, "Better segmentation using SPSS CHAID," SPSS Inc.

[12] Evan, Robert B., "Driving Down Coasts : A Case Study in Data Mining," *Database Programming & Design*, April 1997.

[13] Famili, A., Wei-Min Shen, Richard Weber, Evangelos Simoudis, "Data Preprocessing and Intelligent Data Analysis," *Intelligent Data Analysis*:Elsevier Science Inc., 1996.

[14] Fayyad, Usama, "Diving into Databases." *Database Programming & Design*, March, 1998.

[15] Fayyad, Usama, Gregory Piatetsky,-Shapiro, and Padhraic Smyth. "From Data Mining to Knowledge Discovery in Databases." American Association for Artificial Intelligence. *AI magazine*. Fall. 1996.

[16] Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth, "The KDD Process for Extracting Useful Knowledge from Volumes of Data," *COMMUNICATIONS OF THE ACM*, November, Vol.39, No.11.,1996.

[17] Gendelev, Boris, "Closing the OLAP Gap," *Database Programming & Design*, April 1998.

[18] Glymour, Clark, David Madigan, Daryl Pregibon, and Padhraic Smyth, "Statistical Inference and Data Mining," *COMMUNICATIONS OF THE ACM*, Vol.39, No.11, November 1996.

[19] Imielinski, Tomasz, and Heikki Mannila. "A Database Perspective on Knowledge Discovery," Vol.39, No.11, *Communication of ACM*, Novemver 1996.

[20] Information Discovery, Inc white paper, "Rules are Much More then Decision Trees," 1996.

[21] Jensen, David, Tim Oates, and Paul R. Cohen. "Building Simple Models : A Case Study with Decision Trees." To appear in Proceedings of the Second International Symposium on Intelligent Data Analysis. July 1997.

[22] Kamber, Micheline, Lara Winstone, Wan Gong, Shan Cheng, Jiawei Han. "Generalization and Decision Tree Induction : Efficient Classification in Data Mining." Database Systems Research Laboratory. Simon Fraser University, BC,., Canada V5A 1S6. kamber, winstone, wgong, shanc,

[23] Mannila, Heikki, Department of Computer Science University of Helsinki. "Methods and Problems in data mining".

[24] Mehta, Manish, Jorma Rissanen, Rakesh Agrawal, "MDL-based Decision Tree Pruning," IBM, Almaden Research Center, mmehta, rssanen, agrawal@almaden.ibm.com

[25] Parsaye, Kamran, "New Realms of Analysis," *Database Programming & Design*, April 1996.

[26] Parsaye, Kamran, "OLAP & Data Mining : Bridging the Gap." *Database Programming & Design*. February 1997.

[27] Pilot Software Corp., "An Introduction to Data Mining : Discovering hidden value in your data warehouse," Pilot Software white paper.

[28] Pyle, Dorian, "Putting Data Mining In Its Place". Database Programming & Design. March 1998.

[29] Russel, Stuart, Peter Norvig, *Artificial Intelligence : A Mordern Approach*, Prentice Hall International Editions, 1995.

[30] SAS Corp. "데이터 마이닝 솔루션" 백서. 1998.

[31] Shager, John, Rakesh Agrawal, Manish Mehta, "SPRINT : A Sacalable Parallel Classifier for Data Mining," IBM Almaden Research Center, Proceedings of the 22nd VLDB Conference Mumbai(Bombay), India, 1996.

[32] Tendem Computer Incorporated, "Knowledge Discovery through Data Mining," White Paper.

[33] Tendem Computer Incorporated, "Objects Relational Data Mining Solutions for Card Issuers and Acquirers," Decision Support Solution White paper.

[34] Thearling, Kurt. "From Data Mining to Database Marketing." DIG White Papegr. October 1995.

[35] Ullman, Jeffrey D., "Efficient Implementation of Data Cubes Via Materialized Views," Department of Computer Science, Stanford University.

[36] Venkata, Kolluru, Sreerama Murthy," On Growing Better Decision Trees from Data," The JohnsHopkins University, dissertation for the degree of Doctor of Philosophy, 1995.

■ 저자소개

김 신 곤

연세대학교 경영학과를 졸업하고, 서울대학교에서 재무관리 석사, 죠지아 주립대학에서 Computer Information System 석사, 경영정보학 박사학위를 취득하였으며, ㈜ KLSC에서 즉석복권과 전자복권의 게임 시스템을 개발하는 업무를 담당하였다. 현재 광운대학교 경영대학 경영정보학과 부교수로 재직하고 있으며, 주요 관심분야는 데이터마이닝, 정보기술 아웃소싱 분야이다.