분류모델의 성과 비교에 관한 연구

김 신곤* • 박 성용**

A Study on the Comparison of Classification Models' Performance

Shinkon Kim*, Sungyong Park**

요 약

본 연구는 A카드 회사에서 현재 실시하고 텔레마케팅 시스템에 데이터마이닝 기법 가운데 하나인 CHAID, CART 알고리즘 및 신경망 기법을 적용하여 모델을 개발하고 개발된 모델들의 성과를 분석한다. 이를 통하여 어떻게 기업이 데이터베이스와 데이터마이닝 기법을 마케팅에 효과적으로 사용할 수 있는가에 대한 방안을 제시하고 여러 모델들의 성과를 비교 분석하는 방안을 함께 제시한다.

Key Words: 데이터마이닝, 데이터베이스 마케팅, CHAID, CART, 인공신경망, 리프트

1. 서론

본 연구는 A카드 회사의 텔레마케팅 시스템에 데이터마이닝 기법의 적용을 통해 투자대비 효과와 효율성을 극대화시킬 수 있는 방안을 제시하고 이와 함께 적용한 여러 알고리즘 기법들의 결과에 대한 성과를 비교 분석하는 방안을 제시하는 것을 목적으로 하고 있다.

데이터마이닝 기법을 이용한 모델을 평가하기 위해 사용되는 오차율, 컨피던스 측정치, 그리고 표준편차 등은 여러가지 문제점을 내포하고 있다. 특히, 상이한 종류의 모델들을 비교하고자 할 경우에 모델의 평가는 더 더욱 쉽지 않다. 본 연구에서는 A 카드회사의 텔레마케팅 시스템에서 사용하고 있는 실제 데이터에 CHAID, CART 알고리즘 및 신경망 기법을 적용하여 분류 모델을 개발하고 모델의 성과를 리프트 (LIFT) 개념을 이용하여 분석한다.

본 논문은 의사결정 트리 알고리즘인 CHAID, CART 알고리즘 및 신경망 기법을 이용하여 텔레마케팅 실행에 대해 응답율이 높은 고객 패턴 또는 고객의 특성을 찾아내고 세개의 모델간의 투자 대비 효과를 분석하는 방안을 제시한다. 이를 위하여 A 카드 회사의 과거 1년 동안 텔레마케팅을 실시한 총 33,675 건을 추출하였으며 추출된 데이터는 사전처리 과정 (preprocessing)을 거쳐 트레이닝 데이터와 테스트 데이터로 나누어 진다. 트레이닝 데이터을 통해 만들어진 모델을 테스트 데이터에

*광운대학교 경영정보학과
** 동양시스템하우스

적용하고 그 결과를 리프트를 통하여 비교 분석하였다.

2. 의사결정트리 알고리즘

2.1 CHAID (Chi-Square Automatic Interaction Detection)

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 예측 변수로서 결정된 변수를 이용하여 응답률에서 가장 큰 차이를 갖는 두개 이상의 구분된 집단으로 나누고 그 결과를 트리로 나타낸다 [Deng, 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

$$\chi^2 = \sum \frac{(fo - fe)^2}{fe}$$

fo : 관찰치 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.

2.2 CART (Classification And Regression Tree)

CART 알고리즘은 의사결정 트리 방법론 중 가장 잘 알려진 방법론 가운데 하나이다. 1984 년 Brieman et. al 이 CART (Classification and Regression Tree) 기법을 발표한 이래로 기계학습 실험의 필수 기법이 되어왔다 [Berry and Linoff, 1997]. CART 기법은 전체 데이터 셋을 갖고 시작하여 반복해서 두 개의 자식 노드 (Child node)를 생성하기 위해 모든 예측 변수 (predictor variables)를 사용하여 데이터 셋의 부분집합을 쪼갬으로써 의사결정 트리를 생성한다 [Berry and Linoff, 1997, SPSS, 1998]

그 순수성은 다음과 같은 원리에 의하여 계산된다. 남자 500 명중, 응답자와 비응답자가 각각 100, 400 이라 하자. 2 번에 걸쳐 고객을 뽑을 때, 2 번 모두 응답자일 확률은 $(100 / 500)^2 = (1 / 5)^2$ 이고 마찬가지로 비응답자일 확률은 $(400/500)^2 = (4 / 5)^2$ 이다. Gini Index 는 최종적으로 다음과 같이 계산이 된다 [SAS, 1998].

$$\text{Gini Index} = 1- (1/5)^2 - (4/5)^2$$

위에서 계산된 Gini Index 는 '모든 카타고리 (응답/비응답)에 대하여 임의로 두개 (2)의 원소 (고객)를 뽑을 때, 두개의 원소가 각각 다른 카타고리에서 뽑힐 확률'로 해석할 수 있다. 의사결정나무는 Gini Index 가 작아지는 방향으로 움직이며 Gini Index 값을 가장 많이 감소시켜 주는 변수가 영향을 가장 많이 끼치는 변수가 된다. 그리고 이 변수를 기준으로 의사결정나무의 가지가 만들어 진다 [SAS, 1998] .

2.2.1 CART 알고리즘 처리 절차

2.2.1.1. 초기 구분자 발견 (Finding the Initial Split)

프로세스를 시작할 때, 사전에 미리 분리되어 있는 레코드로 구성된 트레이닝 셋을 갖고 시작한다. 독립변수 또는 다른 변수의 값을 토대로 새로운 레코드의 타겟 필드를 클래스에 할당하게 함으로써 의사결정 트리를 만드는 것이 목적이다. CART 알고리즘은 함수에 따라 각 노드에서 레코드를 쪼갬으로써 이진 트리를 만든다. 그러므로 첫 번째 작업은 어떤 독립 필드가 가장 좋은 구분자 (splitter)인지를 결정하는 것이다. 잠재적 구분자의 능력을 측정하는 방법은 다양성 (diversity)이다. 가장 좋은 구분자는 레코드 셋의 다양성 (diversity)을 줄이는 것이다.

Diversity (before split) - (diversity (left child) + diversity (right child))

2.2.1.2. 전체 의사결정 트리 생성 (Growing The full Tree)

초기 분리는 두 개의 노드를 만든다. 다시 한 번 초기 작업과 같이 대상 구분자를 찾기 위해 입력 필드 모두를 검색한다. 하나의 값만을 갖는 필드가 있다면 더 이상 쪼갤 수 없기 때문에 고려 대상에서 제외시킨다. 대상 노드에서 다양성의 유의적인 감소가 발견되지 않을 때 이것을 리프 노드 (leaf node)라 한다.

2.2.1.3 각 노드의 에러율 측정 (Measuring Error Rate at Each Node)

트리 확장 프로세스의 마지막에 트레이닝 셋의 모든 레코드는 전체 의사결정 트리의 몇 개의 리프 노드(leaf node)를 갖게 된다. 각 리프 노드(leaf node)는 클래스와 에러율이 할당된다. 리프 노드는 더 이상 다양성을 유의적으로 감소시킬 구분자가 발견되지 않는다는 것을 말한다. 그러나, 모든 리프 노드 (leaf node)에 도달한 모든 레코드가 동일한 클래스를 갖는 것을 의미하는 것은 아니다.

2.2.1.4. 트리 가지 치기 (Pruning the tree)

CART 알고리즘은 새로운 분리점이 발견되지 않는 한 의사결정 트리는 계속해서 확장될 것이다. 트레이닝 데이터가 평가를 위해 이용되어 진다면, 트리의 어떠한 가지 치기(pruning)도 에러율을 증가시킬 것이다. 그렇다고 해서 전체 트리가 새로운 데이터 셋을 분류하는 데 가장 최적의 결과를 가져온다는 것을 의미하는 것은 아니다.

2.2.1.5. 하위 트리 후보 구별 (Identifying Candidate subtree)

얼마만큼 트리를 가지 치기를 할 것인가를 결정하기 위해 첫 번째로 반복된 가지 치기 처리 과정을 통해 후보 하위 트리 (candidate subtree)를 구별해야 한다. 목적은 리프 노드 당 최소 추가 예측력을 제공하는 브랜치를 가지 치기 하는 것이다. 가장 유용성이 적은 브랜치를 구별하기 위해, 트리의 조정 에러률 (adjusted error rate)의 개념이 필요하다.

$$AE(T) = E(T) + \alpha \, leaf\_count(T)$$

α 가 0 일 때, 조정 에러률은 에러율과 동일하다. 첫 번째 하위 트리를 발견하기 위해, 뿌리 노드를 포함하는 모든 가능한 하위 트리를 위한 조정 에러률은 α가 점차적으로 증가하는 것으로서 평가된다. 몇몇 하위 트리의 조정 에러률이 완성된 트리의 조정 에러률과 같거나 작을 때, 첫 번째 하위 트리 후보를 발견한 것이 된다 ($\alpha 1$).

2.2 신경망의 구조 및 자료처리 과정

인공 신경망 (Artificial Neural Network)은 인간 두뇌구조의 신경망 구조를 컴퓨터화하여 인간사고의 장점인 비선형성 사고방식의 특성을 모방하고자 개발되었다. 인공 신경망은 순차적 명령구조로 이루어진 종래의 컴퓨터 프로그래밍에 비하여 많은 예의 처리를 통하여 인식된 비 선형적 패턴인식을 이용하여 분석모형을 도출하게 된다.

인공 신경망은 노드(node)의 배열, 층(layer) 사이 또는 층 내의 노드의 연결방법, 뉴런이 정보를 받고 처리하여 결과를 출력하는 방법 (전이함수) 및 상호연결관계의 강도 결정 방법 (학습알고리듬)등에 의하여 인공 신경망의 종류가 결정되는데, 가장 일반적으로 이용되고 있는 "역전파 (Back-Propagation)알고리듬"은 입력층의 각 유니트에 학습대상 패턴을 입력하면, 이 신호는 입력 층의 각 처리요소에서 전이함수를 통하여 변환되어 출력 층에 전달된다. 이 때 출력 층에 전달된 값은 은닉층에서와 같은 절차를 거쳐 출력 값을 산출하여 실제 값과 비교하게 된다. 이 때, 실제 값과의 차이를 최소화하기 위해 역으로 각 처리요소와 연결되어 있는 유니트의 연결강도를 조정하고, 다시 순방향으로의 계산과 역 방향으로의

연결강도 조정을 계속하여 모든 패턴의 값에 만족하도록 오차를 줄여나가는 방향으로 학습하게 된다. 학습된 프로그램은 데이터의 패턴을 판별하거나 예측하는 데 이용하게 된다.

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].

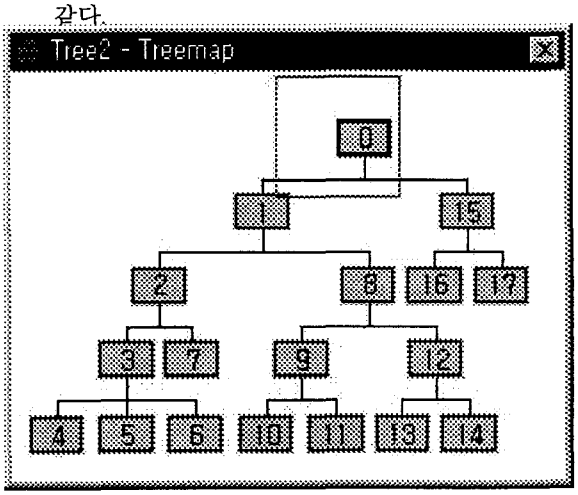
LIFT = P (클래스t 1 표본) / P (클래스t 1 모집단)

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% 응답 비율을 나타내고 있다면 그 모델의 리프트는 10 이다 (50/5 = 10) [Agrawal and Psaila, 1995].

3. 의사결정트리와 분류모델

4.1 CHAID 알고리즘을 이용한 분류모델

CHAID 알고리즘을 트레이닝 셋에 적용하여 개발한 분류모델의 트리 구조도는 [그림 1]과 같다.



[그림 1] CHAID: 트레이닝 셋의 트리 구조도 (Tree Map)

트레이닝 셋에서 개발한 모델을 테스트셋에 적용한 결과는 [그림 2]와 같다.

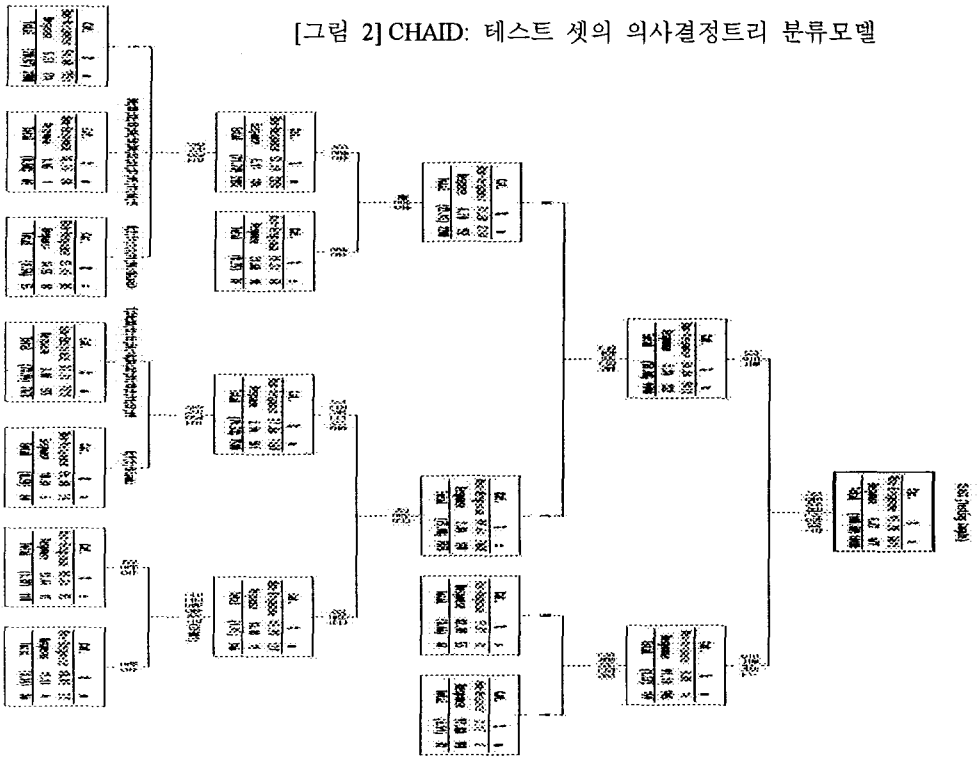
개발된 분류모델에 의하여 텔레마케팅을 위한 대상고객을 선별할 경우, 정보 이익 (Gain %)이 가장 큰 노드의 고객부터 우선적으로 텔레마케팅의 대상고객에 포함시켜야 성공 가능성이 가장 높으며 이때의 반응률은 97.83%에 이르고 있다. 텔레마케팅 대상고객의 수가 늘어남에 따라 정보이익이 그 다음으로 큰 노드가 차례로 그 대상에 포함되어야 한다.

[표1]에의하면 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.

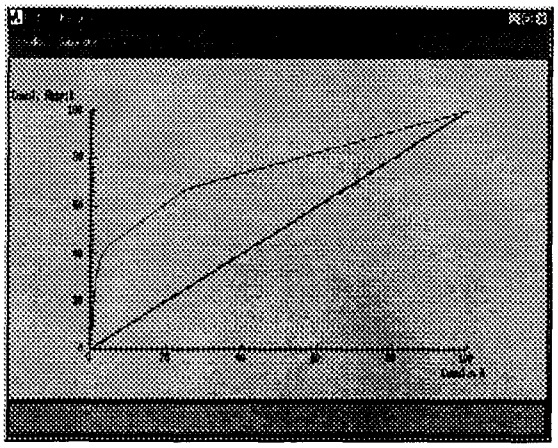
테스트 셋에 대하여 분류모델에 의한 타겟 마케팅 (Target Marketing)을 수행하였을 경우와 무작위 추출에 의한 매스 마케팅 (Mass Marketing)을수행하였을 경우의 텔레마케팅의 효율성을 그래프로 나타낸 것이 [그림 3]이다.

[그림 2] CHAID: 테스트 셋의 의사결정트리 분류모델



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							

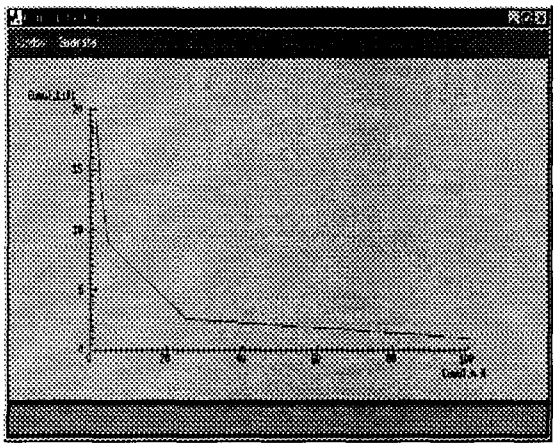
[표 1] CHAID: 테스트 셋 정보 이익 요약표 (Information Gain Summary)



[그림 3] CHAID: 분류모델에 의한 타겟 텔레마케팅과 무작위 추출에 의한 텔레마케팅의 효율성 비교

[그림 3]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 3]로부터 얻을 수 있는 정보 이익에 관한 정보를 대상고객의 크기를 결정하는 한가지의 요소로 고려할 수 있다.

위의 정보 이익 요약표를 바탕으로 리프트 차트를 작성하면 [그림 4]와 같다.



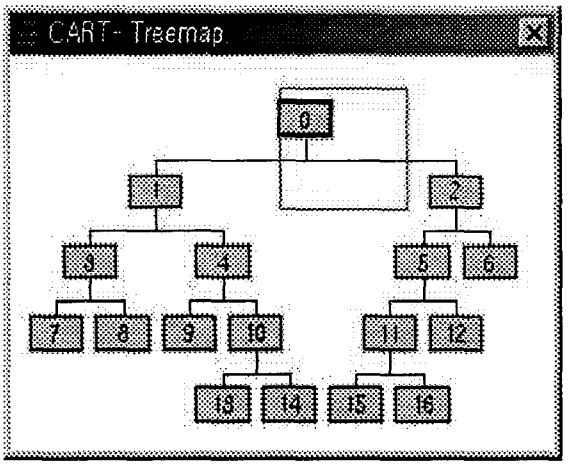
[그림 4] CHAID: 테스트 셋의 리프트 차트 (LIFT Chart)

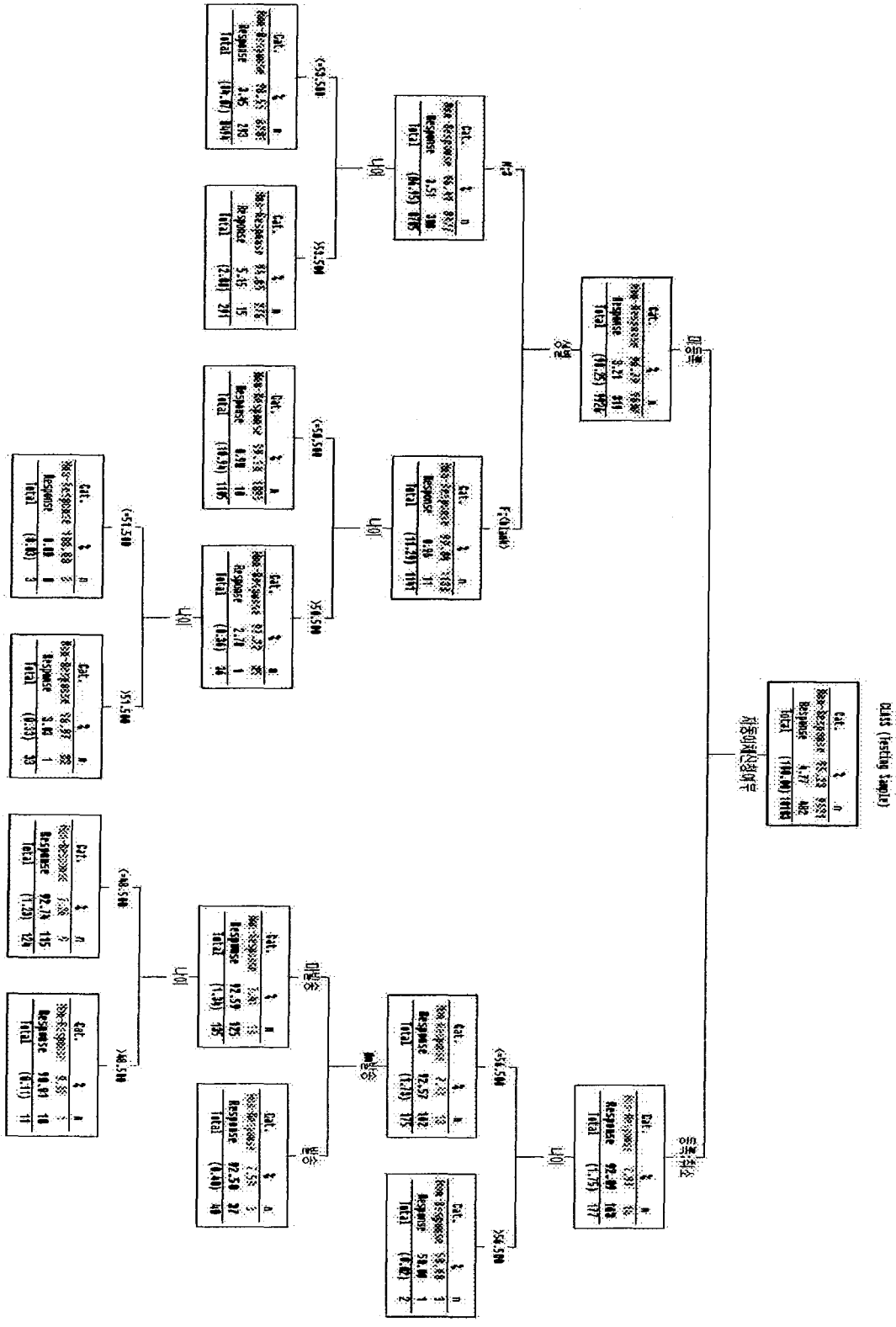
4.2 CART 알고리즘

CART 알고리즘을 트레이닝 셋에 적용하여 개발한 분류모델의 트리 구조도는 [그림 5]과 같다.

[그림 5] CART: 트레이닝 셋의 트리 구조도 (Tree Map)

트레이닝 셋에서 개발한 모델을 테스트 셋에 적용한 결과는 다음과 같다.





[그림 6] CART: 테스트 셋의 의사결정트리 분류모델

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 92.74%에 이르고 있다. 텔레마케팅 대상고객의 수가 늘어남에 따라 정보이익이 그 다음으로 큰 노드가 차례로 그 대상에 포함되어야 한다.

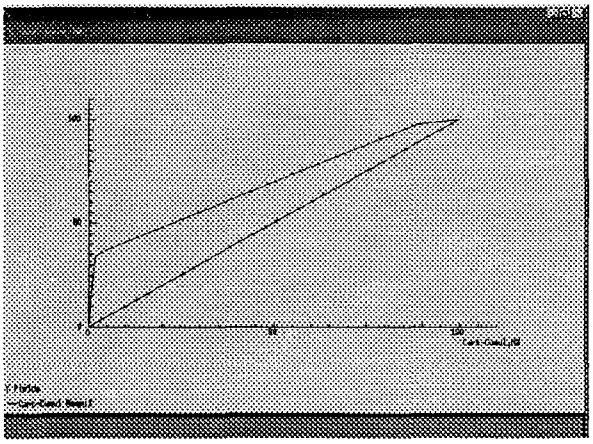
[표 2]에 의하면 텔레마케팅의 대상고객으로는 정보이익이 92.74%로 가장 높은 노드 15가 가장 먼저 포함되어야 하며 이 노드의 124건의 대상고객에는 텔레마케팅을 실시하여 반응을 보인 고객이 115건이 포함되어 있어 노드 15의 반응률은 92.74%임을 의미한다. 또한 노드 15의 리프트는 19.36임을 나타내고 있다. 이것은 이 분류모델에 의하여 선택된 노드 15의 고객에게 텔레마케팅을 실시할 경우 대상고객을 무작위로 추출하여 실시하였을 때보다 19배 이상의 성공 가능성이 높다는 것을 의미한다.

테스트											
Node	Node: n	Node: %	Cumul. N	Cumul.N %	Resp: n	Resp: %	Resp Cumul. N	Cumul Resp: %	Gain (%)	Node Lift	Cumul. Lift
15	124	1.23	124	1.23	115	23.86	115	23.86	92.74	19.36	19.36
12	40	0.4	164	1.62	37	7.68	152	31.54	92.5	19.31	19.34
16	11	0.11	175	1.73	10	2.07	162	33.61	90.90	18.97	19.32
6	2	0.02	177	1.75	1	0.21	163	33.82	50	10.43	19.22
8	291	2.88	468	4.63	15	3.11	178	36.93	5.15	1.07	7.94
7	8494	84.07	8962	88.71	293	60.79	471	97.72	3.44	0.72	1.09
14	33	0.33	8995	89.03	1	0.21	472	97.93	3.03	0.63	1.09
9	1105	10.94	10100	99.97	10	2.07	482	100	0.90	0.18	0.99
13	3	0.03	10103	100	0	0	482	100	0	0	0.99
합계	10103				482						

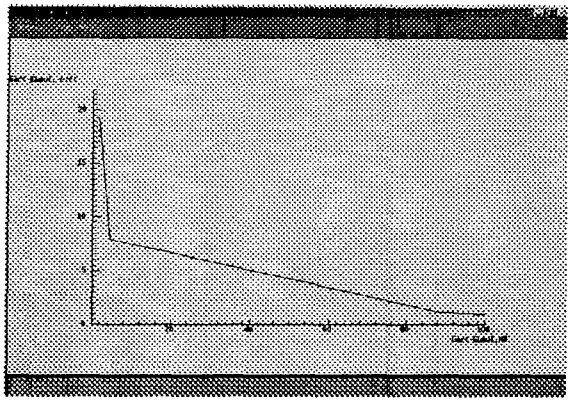
[표 2] CART: 테스트 셋 정보 이익 요약표 (Information Gain Summary)

테스트 셋에 대하여 CART의 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 7]이다.

위의 정보 이익 요약표 2를 바탕으로 CART의 리프트 차트를 작성하면 [그림 8]와 같다.



[그림 7] CART: 분류모델에 의한 타겟 텔레마케팅과 무작위 추출에 의한 텔레마케팅의 효율성 비교 (CART : Cumulative Response Chart (Test Set))



[그림 8] CART Lift Chart (Test Set)

3. 인공 신경망

트레이닝 데이터 셋의 23572건 가운데 응답 (Response)건 수가 1133건으로 응답 (Response) 값과 무응답 (Non-Response) 간에는 큰 불균형 (Skewdness) 으로 인해 신경망을 이용한 모델 형성 작업 과정에서 응답 (Response) 값은 무응답 (Non-Reponse) 값으로 수렴되는 현상이 발생하였다. 따라서, 이러한 현상을 회피하기 위한 방법으로 1133건의 응답 (Response)건수를 중복 (replication) 하여 트레이닝 셋에 포함시켰다. 얼마나 중복하여 트레이닝 셋에 포함시킬 것인가는 여러번의 시행착오를 거쳐 1133건의 8배인 9064건으로 응답 건수를 확장하여 트레이닝 셋에 포함 시켰다.

또한, CHAID와 CART 모델링에서 발견된 입력변수들 중 상관관계가 높은 변수들만 신경망의 입력변수로 선택하였다. 데이터를 가지고 신경망을 학습 혹은 훈련 (training) 시키는 것은 여러 가지 입력변수의 정보로부터 은닉계층 내에서의 다소 복잡한 내부작업을 통해 가장 정확한 결과를 주도록 연결가중치의 값을 찾아가는 것이기 때문에 연관도가 적은 변수들을 제외시켜 신경망의 정확도를 높일 수 있다.

채택된 변수는 우편번호, 자동이체신청여부, 유치경로, 연회비많음 (미사용이유), DM발송여부, 재발급, 쿠폰발송 총 7개 변수를 입력변수로 하였다.

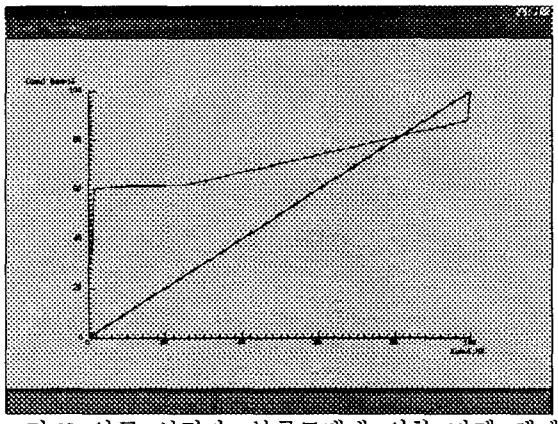
신경망에서 은닉계층의 수는 3계층으로 지정하였으며, 각각의 계층 계수는 Layer 1 : 20, Layer 2 : 15, Layer 3 : 10으로 설정하여 학습시켰다. 물론, 계층의 계수 또한, 여러번의 시행착오를 거쳐 최적값을 선택하였다.

[표3]은 신경망을 이용한 모델의 정보이익 요약표이다. 테스트 셋에 대하여 신경망의 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 9]이다.

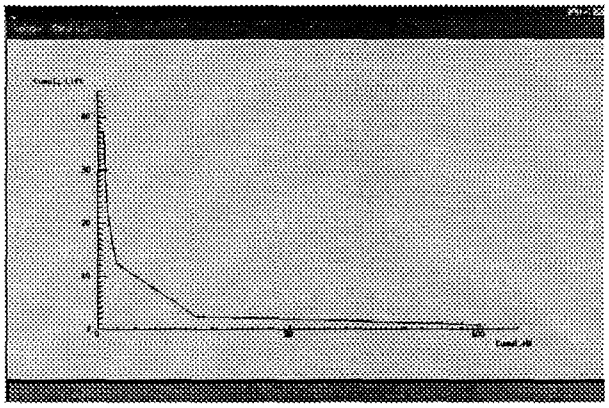
위의 정보 이익 요약표 3 를 바탕으로 신경망 모델의 리프트 차트를 작성하면 [그림 10]와 같다.

e	Node:n	Node: %	Cumul.N	Cumul.N%	Resp:n	Resp: %	Resp Cumul.N	Cumul Resp:%	Gain (%)	Node Lift	Cumul. Lift
1	92	0.910621	92	0.910 1.6206	92	33.82353	92	33.8235	100	37.175	37.175
2	67	0.663169	159	1.57379	67	24.63235	159	58.4559	100	37.175	37.175
3	55	0.544393	214	2.1181827	6	2.205882	165	60.6618	10.90909	4.0554	28.663
4	76	0.752252	290	2.8704345	0	0	165	60.6618	0	0	21.151
5	110	1.088786	400	3.95922	0	0	165	60.6618	0	0	15.335
6	36	0.35633	436	4.3155498	0	0	165	60.6618	0	0	14.068
7	60	0.593883	496	4.9094328	1	0.367647	166	61.0294	1.666667	0.6196	12.442
8	2068	20.46917	2564	25.3786	2	0.735294	168	61.7647	0.096712	0.03595	2.4358
9	7470	73.93843	10034	99.317035	71	26.10294	239	87.8676	0.950469	0.3533	0.8855
10	69	0.682965	10103	100	33	12.13235	272	100	47.82609	17.779	1.00084

[표 3] 신경망 정보이익 요약표 (Test Set)



[그림 9] 인공 신경망: 분류모델에 의한 타겟 텔레마케팅과 무작위 추출에 의한 텔레마케팅의 효율성 비교 (Neural Network : Cumulative Response Chart (Test Set))



[그림 10] 인공 신경망 LIFT Chart (Test Set)]

5. 결론

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

모델의 성과를 분석하기 위하여 리프트 (LIFT)를 사용하였다. 본 논문에서 논의한 분류모델을 사용할 경우, 현실성이 없는 가정이기는 하나 추출 대상고객의 수에 제한이 없다면 무작위 추출에 의한 경우와 비교하여 최고 37배 이상, 텔레마케팅의 효율성을 높일 수 있다고 보여진다.

누적반응표와 리프트의 값을 이용하여 모델들의 성능을 분석하였다. 그러나 어느 모델에서나 표본 집단의 수를 연속적으로 원하는 만큼 늘려갈 수 있다는 가정하에 산정된 것이므로 이 가정은 정보요약표에서 보듯이 현실적이지 못하다. 따라서 모델들 사이의 보다 더 정교한 비교 분석을 위하여는 의사결정트리의 노드와 신경망 모델에 의한 표본수의 선택이 연속적이지 못한 부분에 대한 검증이 요구된다.

참고 문헌

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

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

[3] 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.

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

[5] Deng, Stephen, "Better segmentation using SPSS CHAID", SPSS Inc., http://www.spss.com/cool/papers/chaid1.htm

[6] 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

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

[8] Information Discovery, Inc white paper, "Rules are Much More then Decision Trees", 1996, http://www.datamining.com/datamine/trees.html

[9] 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

[10] 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, han@cs.sfu.ca

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

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

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

[14] SAS Corp. "데이터 마이닝 솔루션" 백서. 1998. http://www.sas.com/offices/asiapacific/korea/solution/mining/wp/mining-wp.html

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

[16] Thearling, Kurt. "From Data Mining to Database Marketing". DIG White Paper. October 1995. www.santafe.edu/~kurt/text/wp9502/wp9502.shtml

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

[18] SPSS Corp, "AnswerTree 1.0 User's Guide", 1998

[19] SAS Corp. "데이터 마이닝 솔루션" 백서. 1998. http://www.sas.com/offices/asiapacific/korea/solution/mining/wp/mining-wp.html