

데이터 마이닝을 이용한 효과적인 데이터베이스 마케팅에 관한 연구

김 신곤

광운대학교 경영정보학과

E-Mail.shinkon@daisy.kangwoon.ac.kr

박 성용

동양 Systemhouse

parksy@tysystemhouse.com

Abstract

본 연구는 A카드社에서 현재 실시하고 텔레마케팅 시스템에 데이터 마이닝을 적용하여 현실에서 어떻게 기업이 데이터마이닝을 이용하여 데이터베이스를 마케팅에 효과적으로 사용할 것인지에 대한 방향을 제시하고 있을 뿐만 아니라 현재와 미래에 대한 연구 방향을 제시하고 있다.

1. 서론

모든 산업에서 수많은 데이터가 놀랄만한 비율로 증가하고 있다. 금융업에서 통신업체에 이르기 까지 기업들은 경쟁을 위해 엄청나게 쌓여있는 데이터의 분석에 점차로 의존하고 있다. 급속도로 증가하고 있는 디지털 데이터로부터 유용한 정보(지식)를 끄집어낼 수 있도록 사람들을 보조하기 위한 도구와 컴퓨터 이론의 새로운 세대에 대한 요구가 있다. 이러한 요구가 데이터 마이닝의 주요 목적이다 [1].

본 논문은 A카드社의 텔레마케팅 시스템에 데이터마이닝의 적용을 통해 실질적으로 데이터베이스 마케팅을 위하여 데이터 마이닝이 어떻게 활용될 수 있는가를 보여주고 동시에 데이터 마이닝을 통해 발견된 결과와 효과를 분석하였다.

2. 데이터 마이닝의 개념

데이터 마이닝에 대한 고찰을 위하여는 우선 데이터 마이닝의 개념적 정의를 명확히 하는 것이 선행되어야 할 것이다. 데이터 마이닝은 방대한 양의 데이터로부터 의미 있는 패턴, 규칙들을 발견하기 위하여 자동적인 혹은 반자동적인 방법으로 데이터를 분석하고 탐색하는 것을 말한다.[28] 또한, 1996년 Gartner Group의 E.Brethenoux, H. Drenser, K.Strange 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 것으로써 가치사슬을 "데이터에서 정보로 정보에서 지식으로 지식에서 의사결

정"에 이르는 경로를 설명함과 동시에 데이터에서 의사결정까지 이르는 프로세스의 주요한 세 가지 정보기술로서 데이터베이스, 인프라구조 기술, 데이터 마이닝을 강조하였다. 이 리포트에서 "데이터 마이닝은 단순한 도구가 아닌 여러 단계의 프로세스이다"라고 정의하고 있다.[23]

3. 데이터 마이닝 종류 및 사용기법

특정 문제에 적용하는 데이터 마이닝 기법이 정해져 있는 것은 아니다. 얻고자 하는 결과나 데이터의 상태등에 따라 적용할 수 있는 기법은 다를 수 있다 [42]. 따라서 최상의 모델을 찾아내기 위해서는 다양한 데이터 마이닝 기법과 알고리즘을 갖추어야 한다.

데이터 마이닝의 기법에는 일반적으로 통계학에서 사용되는 여러 분석 기법들을 포함하여 연관규칙(Associations), 클러스터링(Clustering), 의사결정 트리(Decision Tree) 그리고 신경망(Neural networks)과 같은 기법들이 있다. 본 논문은 데이터 마이닝의 기법인 의사결정 트리에 기초한 CHAID 알고리즘에 초점을 두고 있다. 1975년 J.A. Hatigan에 의해 처음 발표된 CHAID (CHI-SQUARE Automatic Interaction Detection) 알고리즘은 1963년 J.A.Morgan과 J.N. Sonquist에 의해 발표된 AID(Automatic Interaction Detection) 시스템에서 유래된다. "AID"에서 암시하고 있는 것과 같이 CHAID에 대한 원래 동기는 변수들 간의 통계적 관계를 찾는 것이 목적이었다. 이것은 다시 의사결정 트리를 통해 변수들간의 통계적인 관계를 찾을 수 있으므로 이 방법은 분류기법 (classification technique)으로서 사용하게 되었다 [28]. CHAID는 시장 분할(market segmentation)이 요구될 때 종종 사용되는 통계적 절차이다. CHAID는 변수가 범주형 데이터이고 예측 변수 (predictor variable)와 범주형 결과 변수간의 관계를 찾아야 할 때 가장 유용하다 [45]. 다른 의사결정 트리와 같이 CHAID알고리즘은 두개 이상의 자식 노드로 트레이닝 데이터를 쪼개기 위한 입력변수(input variables)를 사용하는 방법을 찾는다. 즉, CHAID는 분리기준(split)를 찾는 것을 시발점으로하며 자식노드는 특정 변수가 갖고 있는 결과변수의 확률이 각 노드마다 다르게 하는 방식으로 선택된다. CHAID는 데이터의 집합을 검

색하여 예측변수의 예측치로서 가장 유의성이 높은 변수를 결정한다. 데이터베이스에서 어떤 고객이 DM(direct mail)에 가장 응답하기 쉬운가를 예측하려 한다면, CHAID 알고리즘은 최상의 예측 변수로서 결정된 변수를 이용하여 응답률에서 가장 큰 차이를 갖는 두개 이상의 구분된 집단으로 나누고 그 결과를 트리로 나타낸다 [44]. 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 $r \times c$ 분할표(contingency table)로부터 계산된다. 이때, Pearson의 카이제곱 통계량은

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

f_o : 관찰치, f_e : 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 노드(NODE)는 p-값이 가장 작은 변수를 기준으로 가지가 형성되는 것이다 [32,46].

4. 데이터마이닝 적용

4.1. A카드사의 텔레마케팅 시스템

현재 14명으로 구성된 A카드사의 텔레마케터의 주 업무는 카드 발급 후 카드를 사용하지 않았거나 연회비를 납부하지 않은 고객을 대상으로 회원 유지를 목표로 텔레마케팅을 실시하고 있다.

본 논문은 텔레마케팅 대상 고객 가운데 텔레마케팅 실행에 대해 응답율이 높은 고객 패턴을 찾아 내고 발견된 패턴을 이용하여 텔레마케팅 실행 대상이 되는 고객리스트를 제공함으로써 텔레마케팅의 성공율을 높이는 효과적인 데이터베이스 마케팅에 관한 연구이다. 즉, 데이터베이스로부터 텔레마케팅 대상 고객을 무작위 추출 방법이 아닌 데이터 마이닝을 통해 발견된 패턴에 따라 추출된 대상 고객을 텔레마케터에게 할당함으로써 텔레마케터의 업무성과를 향상시키는 것을 목표로하고 있다.

4.2. 데이터 생성 및 데이터 사전처리

데이터 추출은 과거 1년 동안 텔레마케팅을 실시한 고객 중 카드발급 후 14개월 미사용 고객을 추출하였으며, 입력변수는 데이터베이스에 저장되어 있는 고객의 인구통계적 변수, DM발송여부, 쿠폰발송여부, 자동이체 신청고객여부와 카드유치구분 등 15개 필드를 포함하였다. 그러나, 필드에 포함되어 있는 데이터 중 20%미만이 Null 데이터 값을 가지고 있는 레코드는 대상 변수 항목에서 제외[29] 시켰으며, 필터링,

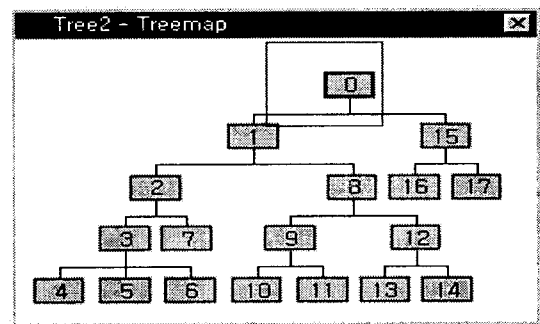
중복 데이터 처리, 비대칭 분포 처리, 그리고 결측치 처리과정을 통해 데이터 마이닝 도구에 맞게 데이터 사전처리를 하였다. 이렇게 추출된 데이터는 총 33,675건이며 텔레마케팅 실시 후 응답한 고객은 4.7%이다.

4.3. 데이터 마이닝 실행

의사결정 트리 알고리즘을 사용하여 만들어진 모델은 검증절차를 거쳐야하는데 이러한 검증방법 가운데 하나는 데이터를 두개의 데이터 셋으로 나누는 것이다. 즉, 모델을 만들기 위한 트레이닝 데이터 셋(Training Data Set)과 트레이닝 데이터 셋으로 만들어진 모델을 테스트하기 위한 테스트 데이터 셋(Test Data set)으로 전체 데이터 셋을 나눈다. 트레이닝 데이터 셋에서 만들어진 모델이 테스트 데이터 셋에서 적용하였을 경우 동일한 결과를 나타낸다면, 이것은 현재 데이터와 유사한 더 큰 데이터 셋에서도 유사한 결과를 생성된다는 것을 의미한다.[42] 따라서, 본 실험에서도 이와 같은 방법으로 데이터 셋을 트레이닝 셋과 테스트 셋으로 나누어 검증하였다.

4.4. 14개월 미사용 데이터 셋 실험

14개월 미사용 고객 데이터의 트레이닝 셋은 전체 데이터 셋의 70%로 총 23,572건이고 테스트 셋은 전체 데이터 셋의 30%를 차지하는 10,103건에 대하여 데이터 마이닝을 실시하였다.



[그림 4-1] 14개월 미사용 트리 구조도

그림 4-1 트리 구조도(Tree Map)와 같이 14개월 미사용 트레이닝 셋은 총 17개의 노드를 갖는 트리가 생성되었고 각 리프 노드(leaf node)는 정보 이익 요약표(Information Gain Summary)에 의해 트리를 분석할 수 있다. 본 트리 구조의 숫자는 단지 각각의 노드에 번호를 부여하여 트리를 해석하는데 용이성을 제공하는 목적으로 번호가 부여되었다.

일반적인 트리 생성 알고리즘에서는 언제 노드의 확장을 멈춰야 하는가를 결정하는 규칙이 필요하다. 이러한 규칙을 정지 법칙(Stopping Rule)이라고 한다 [42,43]. 14개월 미사용 트레이닝 셋의 경우 정지 규칙을 적용하였을 때, 최대 노드 깊이(Node Depth)를 4로 설정하였으며 부모노드(Parent Node)와 자식노드(Child Node) 각각에 노드를 형성하기 위한 필요조건으로 최소 데이터 수를 부모 노드는 100건, 자식 노드

Misclassification Matrix (14개월 데이터 셋)							
		Actual Category(14개월 트레이닝 셋)			Actual Category(14개월 테스트 셋)		
		Non-Response	Response	Total	Non-Response	Response	Total
Predicted Category	Non-Response	22,396	722	23,118	9,612	332	9,944
	Response	43	411	454	14	145	159
	Total	22,439	1,133	23,572	9,626	477	10,103
Learning Sample				Testing Sample			
Risk Estimate				0.0324538			
SE of Risk Estimate				0.00115417			
				0.0342473			
				0.00180934			

[표 4-2] 14개월 미사용 데이터 셋 오차분석표(Misclassification Matrix)

는 50건의 최소 데이터 수를 갖도록 설정하였다.

표 4-1은 14개월 미사용 트레이닝 셋을 CHAID 알고리즘에 적용한 결과 생성된 의사결정 트리의 정보 이익요약표이다. 트리의 모든 노드는 텔레마케팅 실시 후 카드를 사용하거나 연회비를 입금한 응답고객과 그렇지 않은 무응답고객으로 구분하여 각각의 분포를 보여주고 있다. 총 응답과 무응답 건수를 나타내 주는 뿌리 노드의 응답(Response)은 총 23,572건 중 4.81%인 1,133명이 텔레마케팅 실시 후 응답을 보이고 있다. 따라서 본 실험의 결과에서 얻은 4.81%의 응답률 이상을 얻은 노드를 데이터 마이닝의 실행을 통해 얻은 의미 있는 정보로 간주하여 결과를 해석할 수 있다. 각 의사결정 트리 노드의 입력변수(데이터 속성)와 결과변수(응답여부)간의 상관관계가 높은 변수를 뿌리노드에 가장 가까운 입력변수로 선정하였고 의사결정 트리 형성시 노드의 쪼개짐(split)과 결합(merge)을 결정하기 위한 유의수준을 0.05로 하였다.

표 4-1에서 4, 6, 7, 11, 13, 14, 16, 17번 리프 노드가 전체 평균 응답률 4.81% 이상의 응답률을 나타내고 있다. 정보 이익율이 4.81% 이상인 노드의 전체 데이터는 5,931건(노드 4, 6, 7, 11, 13, 14, 16, 17 데이터 합)이고 텔레마케팅 실행에 응답한 건수는 796건으로 응답률 13.42%을 보이고 있다. 선택된 노드들은 트레이닝 셋에서 만들어진 모델이 되며 테스트 셋을 본 모델에 적용하여 모델의 적합성을 평가할 수 있다.

Gain(%) : 현재 노드에 분류된 총 데이터 건수 (Non-Response건수 + Response건수)에서 Response의 건수가 현 노드에서 차지하고 있는 비율

Index(%) : 현재 노드에 포함된 Response 건수와 전체 데이터 셋 건수 간의 비교 측정치 제공

정보이익표의 열 Gain(%)은 이 노드에 속한 고객에게만 텔레마케팅을 실시하였을 경우 예상되는 응답률로서 Gain(%)와 Index(%)를 수식으로 나타내면 다음과 같다.

Gain(%) = COUNT(Response in this node) / COUNT(Response + Non-Response in this node)

Index(%) = Gain(%) / 데이터 셋 전체 Resp:%

일단 의사결정 트리 모델이 만들어지면, 모델을 평가하고 트리의 타당성을 검증해야만 한다. 이러한 모델의 적합도를 판단하기 위해 오

차 분석표(Misclassification Matrix)가 사용된다. 오차 분석표는 실제 데이터 수와 예측된 데이터 수의 관계를 보여주고 있다. 이것은 얼마나 모델이 잘 만들어졌는지 뿐만 아니라 분류를 잘못된 결과에 대하여도 보여주고 있다.

표4-2 오차 분석표에서 보는 것과 같이 한 노드에서 하나의 쪼개짐을 갖는 오류 측정치(Risk estimate)는 실제로 무응답한 건수를 응답으로 잘못 예측한 43건의 데이터와 응답한 실제 데이터를 무응답으로 잘못 분류한 722건의 합계를 총 데이터 건수인 23,572로 나눈 0.0324538이고 표준 오차는 0.00115417이다. 이것은 현재 모델을 적용할 때 의사결정 트리 모델에서 생성된 의사결정 규칙(decision rule)을 적용할 때, 총 데이터 셋 23,572건 중 올바르게 분류된 데이터는 22,807건으로 전체 정확성은 22,807 / 23,572로서 96.76%를 나타내고 있다. 즉, 14개월 미사용 트레이닝 셋의 오차 분석표는 정확하게 데이터의 96.76%를 분류하고 있는 것을 보여주고 있다. 이 트리에서 텔레마케팅에 응답한 고객(Response)이 응답하지 않은 고객(Non-Response)으로 분류될 가능성은 거의 없다.[41] 이렇게 만들어진 모델은 다음과 같은 가설을 만들 수 있다.

“데이터 마이닝 기법을 사용하여 발견된 모델을 이용하여 텔레마케팅 데이터 베이스에서 텔레마케팅 대상 고객을 추출할 경우 응답률을 13.42%까지 향상시킬 수 있다”

트레이닝 셋의 실험 결과 즉, 가설은 테스트 셋의 실험에서 검증하게 된다. 테스트 셋은 트레이닝 셋의 실험에서 만들어진 모델에 대한 검증을 목적으로 한다. 테스트 셋은 위에서 언급한 것과 같이 14개월 미사용 전체 데이터 셋에서 무작위로 추출한 트레이닝 셋을 제외한 나머지 데이터로 총 데이터 셋의 30%인 10,103건을 대상이 된다. 표 4-1에서 트레이닝 셋의 데이터 마이닝 실험 결과와 같이 4, 6, 7, 11, 13, 14, 16, 17번 리프 노드는 뿌리 노드의 텔레마케팅 실시 후 응답한 고객의 응답률 4.72% 이상의 응답률을 나타내고 있다. 응답률이 4.72% 이상인 노드의 전체 데이터는 2,564건이고 텔레마케팅 실행에 응답한 건수는 321건으로 응답률 12.52%을 보이고 있다.

표 4-2 테스트 오차 분석표에서와 같이 오차 측정치(Risk Estimate)는 0.0342473으로 테스트 셋의 트리는 올바르게 96.58% 데이터가 분류되었음을 보여주고 있다. 트레이닝 셋의 실험에서 만들어진 모델은 테스트 셋의 적용 결과

동일한 노드에서 뿌리노드의 정보 이익율보다 높은 정보 이익율의 결과를 나타냈다. 그러므로, 트레이닝 셋의 실험에서 생성된 가설은 검증되었다고 볼 수 있으며 이것은 다시 현실에 적용하기에 적합한 모델이다라는 결론을 내릴 수 있다.

위 데이터 마이닝 적용 결과는 SQL문이나 IF문과 같은 상위단계 언어로 표현할 수 있으며 이를 토대로 실제 현실에 적용이 용이하다.

5. 결론

대다수의 기업은 많은 자료가 여러 시스템에 산재되어 있는 실정이다. 어떤 측면에서 데이터가 정보로 전환되지 못한다면 많은 양의 데이터를 보유하는 것 자체가 비용낭비라고 볼 수 있다. 본 연구는 이러한 문제점을 갖고 있는 기업에 데이터 마이닝을 이용하여 데이터베이스 마케팅에서 잠재고객을 발굴하거나 고객의 문제를 발견/해결하는데 중점을 둘 수 있는 해결방안을 제시하였다. 데이터 마이닝 분야의 빠른 발전은 데이터베이스, 데이터 웨어하우스, 지식 기반 시스템, 통계학, 기계학습, MIS, 그리고 다른 관련 분야 등에서 연구자와 개발자의 접목된 노력을 해야 할 것이다. 또한, 데이터의 종류가 다양하듯 데이터 마이닝 기술을 통해 얻을 수 있는 지식의 종류 또한 그 지식의 형태는 업무 현장의 요구에 따라 결정되므로 새로운 요구가 제기되면 그에 따른 새로운 데이터 마이닝 기법이 개발되어야 한다.

참고 문헌

1. Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From Data Mining to Knowledge Discovery in Databases". American Association for Artificial Intelligence. AI magazine. Fall. 1996
2. Ronald J. Brachman, Tej Anand, "The Process of Knowledge Discovery in Databases: A First Sketch", AAAI-94 Workshop on Knowledge Discovery in Databases, KDD-94
3. Tomasz Imielinski and Heikki Mannila. "A Database Perspective on Knowledge Discovery". November 1996/Vol.39, No.11 Communication of ACM
4. 나민영/육군사관학교 전산학과, 1998, "'데이터 마이닝' 대규모 지식데이터베이스에서 유용한 지식 추출하는 기법"
5. Usama Fayyad. "Diving into Databases". Database Programming & Design March, 1998
6. Robert B. Evans. "Driving Down Coasts : A Case Study in Data Mining". Database Programming & Design April 1997.
7. 김신곤. 광운대학교 경영정보학과. "데이터 마이닝과 지식발견"
8. Dorian Pyle. "Putting Data Mining In Its Place". Database Programming & Design. March 1998.

9. Kurt Thearling. "From Data Mining to Database Marketing". DIG White Paper. October 1995. www.santafe.edu/~kurt/text/wp9502/wp9502.shtm
10. Micheline Kamber, 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. Heikki Mannila. Department of Computer Science University of Helsinki. "Methods and Problems in data mining" Heikki.Mannila@cs.helsinki.fi
12. Kamran Parsaye. "OLAP & Data Mining : Bridging the Gap". Database Programming & Design. February 1997.
13. Kamran Parsaye. "New Realms of Analysis", Database Programming & Design, April 1996.
14. Pieter Adriaans, Dolf Zantinge "Data Mining", Addison Wesley, 1996
15. Rakesh Agrawal, Ashish Gupta, Sunita Sarawagi, "Research Report: Modeling Multidimensional Databases", IBM Almaden Research Center
16. Jeffrey D. Ullman, "Efficient Implementation of Data Cubes Via Materialized Views", Department of Computer Science, Stanford University. <http://db.stanford.edu/~ullman>
17. Information Discovery, Inc white paper, "Rules are Much More than Decision Trees", 1999, <http://www.datamining.com/datamine/trees.html>
18. Agrawal, R., and Psaila, G. "Active Data Mining", In Proceedings of the first International Conference On Knowledge Discovery and Data Mining (KDD-95), 3-8.
19. Clark Glymour, David Madigan, Daryl Pregibon, and Padhraic Smyth, "Statistical Inference and Data Mining", COMMUNICATIONS OF THE ACM, November 1996, Vol.39, No.11
20. Tendem Computer Incorporated, "Knowledge Discovery through Data Mining", White Paper
21. Tendem Computer Incorporated, "Objects Relational Data Mining Solutions for Card Issuers and Acquirers", Decision Support Solution White paper.
22. E.Brethenoux, 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.
23. Ronald J. Brachman, Tom Khabaza, Willi Kloesgen, Gregory Piatetsky-Shapiro, and

- Evangelos Simoudis, "Mining Business Databases", COMMUNICATIONS OF THE ACM, November 1996, Vol. 39, No.11.
24. John Shager, Rakesh Agrawal, Manish Mehta, "SPRINT: A Scalable Parallel Classifier for Data Mining", IBM Almaden Research Center, Proceedings of the 22nd VLDB Conference Mumbai(Bombay), India, 1996
25. Boris Gendele, "Closing the OLAP Gap", Database Programming & Design, April 1998.
26. Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, "The KDD Process for Extracting Useful Knowledge from Volumes of Data", COMMUNICATIONS OF THE ACM, November 1996, Vol.39, No.11.
27. Stuart Russel, Peter Norvig, "Artificial Intelligence A Modern Approach", Prentice Hall International Editions, 1995
28. Michael J. A. Berry, Gordon Linoff, "Data Mining Techniques for Marketing, Sales, and Customer Support", Wiley Computer Publishing-John Wiley & Sons, Inc, 1997
29. A. Famili, Wei-Min Shen, Richard Weber, Evangelos Simoudis, "Data Preprocessing and Intelligent Data Analysis", Intelligent Data Analysis:Elsevier Science Inc., 1996. <http://www-east.elsevier.com/ida/browse/96-1/ida96-1.htm>
30. Stephen Deng, "Better segmentation using SPSS CHAID", SPSS Inc., <http://www.spss.com/cool/papers/chaid1.htm>
31. B. de la Iglesia, J.C.W. Debusse and V.J. Rayward-Smith "Discovering Knowledge in Commercial Databases Using Modern Heuristic Techniques" KDD-96
32. SAS Corp. "데이터 마이닝 솔루션" 백서. 1 9 8 . <http://www.sas.com/offices/asiapacific/korea/solution/mining/wp/mining-wp.html>
33. Andreas Ittner, Michael Schlosser "Discovery of Relevant New Features By Generating Non-Linear Decision Trees" Dept. of Computer Science, Chemnitz University of Technology, Germany, Dept. of Electrical Engineering, Fachhochschule Koblenz, Germany, andreas.ittner@informatik.tu-chemnitz.de, schlosser@koblenz.fh-rpl.de
34. Pilot Software Corp., "An Introduction to Data Mining : Discovering hidden value in your data warehouse", Pilot Software white paper.
35. Kolluru Venkata Sreerama Murthy, " On Growing Better Decision Trees from Data", The JohnsHopkins University, disertation for the degree of Doctor of Philosophy, 1995.
36. David Jensen, 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
37. Manish Mehta, Jorma Rissanen, Rakesh Agrawal, "MDL-based Decision Tree Pruning", IBM, Almaden Research Center, {mmehta, rissanen, agrawal}@almaden.ibm.com }
38. Estelle Brand, Rob Gerritsen , " Decision Trees ", DBMS, Data Mining Solutions Supplement, July, 1998
39. David Jensen and Matt Schmill, "Adjusting for multiple comparisons in decision tree pruning", Department of Computer Science, University of Massachusetts, { jensens|schmill }@cs.umass.edu
40. 박종수, "연관 규칙 탐사와 그 응용", 데이터 베이스 연구회 1998년도 춘계 튜토리얼, 한국정보과학회 데이터베이스연구회, 1998
41. Herb Edelstein, "Data Mining -- Let's Get Practical", DB2 magazine, Summer 1998 Data Mining
42. SPSS Corp, "AnswerTree 1.0 User's Guide", 1998
43. SPSS Inc, "AnswerTree Algorithm Summary", SPSS White paper, 1998 http://www.spss.com/cool/papers/algo_sum.htm
44. Richard Deere, "Using CHAID Analysis in Direct Marketing", Acxiom Case-in-Point Vol2 No 2 report, <http://www.acxiom.com/cip-rpt-c.htm>
45. Market Strategies Corp., "Market Segmentation CHAID", Market Strategies white paper, <http://www.marketstrategies.com/itmethod/chaidh.htm>
46. 이성근, 김범중,채서일, "SPSS/PC+를 이용한 통계분석", 학현사
47. Lelia Morrill, "Enterprise Mining: More than a Tool", Decision Support, Database programming & design, February, 1998
48. 김상훈, "경영정보시스템(MIS) 연구현황 및 발전방향", 한국경영과학회, 1987