The Korean Communications
in Statistics Vol. 7, No. 3, 2000
pp. 937-943

Interpretation of Data Mining Prediction Model Using Decision Tree

Hyuncheol Kang[1], Sang-Tae Han[2], Jong-Hoo Choi[3]

Abstract

Data mining usually deal with undesigned massive data containing many variables for which their characteristics and association rules are unknown, therefore it is actually not easy to interpret the results of analysis. In this paper, it is shown that decision tree can be very useful in interpreting data mining prediction model using two real examples.

Keywords : Data mining, Supervised prediction, Unsupervised prediction, Decision tree

1. 서론

의사결정나무(Decision Tree)는 관심대상이 되는 집단을 몇 개의 소집단으로 분류하거나 예측을 수행하기 위해서 자주 사용되는 분석기법 중의 하나이다. 특히, 의사결정규칙(decision rule)이 나무구조로 표현되기 때문에 분류 또는 예측을 수행하는 다른 방법들에 비해서 연구자가 분석과정을 쉽게 이해하고 설명할 수 있다는 장점을 가지고 있다(최종후 등, 1998). 의사결정나무분석은 예측모형 자체로 사용될 뿐만 아니라 이상치를 검색하거나 분석에 필요한 변수 또는 교호효과를 찾아내는데 많이 이용되고 있다. 의사결정나무분석을 수행하기 위해 CHAID(Kass, 1980), CART(Breiman et. al., 1984), C4.5(Quinlan, 1993), QUEST(Loh & Shih, 1997) 등의 알고리즘들이 제안되어 있으며, 지금도 많은 연구자들에 의해서 다양하게 개선된 알고리즘들이 개발 및 제안되고 있다.

최근에 부각되고 있는 데이터마이닝은 대용량의 자료로부터 기업의 경쟁력 확보를 위한 의사결정을 돕는 유용한 정보를 찾아내는 일련의 분석과정이라고 할 수 있는데 통계학, 인공지능, 기계학습, 계량경제학 등 다양한 분야의 분석방법들이 데이터마이닝을 위해 사용되고 있다. 데이터마이닝에서는 대용량의 데이터를 다룰 뿐만 아니라 변수의 성격이나 그들 간의 관계가 잘 알려져 있지 않은 많은 변수를 사용하여 분석하기 때문에, 분석결과를 해석하는 것이 쉽지 않은 경우가 대부분이다. 데이터마이닝 예측모형을 구축하는 중요한 목적 중의 하나는 정확한 예측을 수행하

1) (136-701) Institute of Statistics, Korea University, Seoul, Korea.
2) (336-795) Department of Mathematics, Hoseo University, Asan, Korea.
3) (339-800) Department of Informational Statistics, Korea University, Chochiwon, Korea.
E-mail : jchoi@tiger.korea.ac.kr

는 것에 있다고 할 수 있지만, 예측모형의 해석적 어려움은 실제적인 문제에 있어서 큰 단점이 될 수 있다. 즉, 각 변수의 중요성과 변수들 간의 상호작용을 이해하는 것은 향후의 데이터 수집과 관리 작업을 향상시킬 수 있다. 또한 신용평가 등과 같은 현실적 문제에 있어서는 왜 그러한 예측이 이루어졌는지에 대한 사유가 법적인 요구사항이 되는 경우도 있다. 더구나, 수리적 지식이 부족한 연구자 또는 고객이 주어진 예측모형을 이해할 수 있다면 그들에게 예측모형의 가치를 납득시키기 훨씬 더 용이할 것이다. 따라서 정확한 예측뿐만 아니라 해석적 용이함도 추구할 수 있다면 예측모형의 실제적 유용성을 훨씬 높일 수 있을 것이다.

본 논문에서는 데이터마이닝의 두 분야라고 할 수 있는 지도예측(supervised prediction)과 자율예측(unsupervised prediction)에 대해 예측모형의 해석적 측면에서 의사결정나무를 활용할 수 있는 방법을 제안하고자 한다. 2절에서는 목표변수(target variable)가 있는 지도예측에 있어서 신경망 모형(Neural Network Model)을 적용한 사례에서 나타나는 해석적 어려움을 의사결정나무 규칙을 적용하여 효율적 해석을 할 수 있음을 보였고, 3절에서는 목표변수가 없는 자율예측에 있어 군집분석을 적용한 사례에 대해 의사결정나무 규칙을 해석적 측면에서 효율적으로 활용하는 경우를 살펴보았다.

2. 지도예측모형의 해석에 대한 사례연구

지도예측은 일반적으로 목표변수(또는 종속변수)가 존재하는 경우의 분석을 의미하는데, 이는 입력변수(input variable)로부터 목표값을 예측하는 모형(규칙)을 개발하기 위해서 사용된다. 대표적인 분석기법으로는 통계학 분야의 판별분석이나 회귀분석 그리고 인공지능 분야에서 개발된 MLP(Multi-Layer Perceptron) 신경망 등이 있다.

자료분석 분야에서 신경망(또는 인공신경망)은 복잡한 구조를 가지고 있는 자료에 대한 예측문제를 해결하기 위해서 사용되는 유연한 비선형모형의 하나로 분류될 수 있다(Smith, 1996). 특히 컴퓨터의 성능향상에 힘입어 최근에 그 사용범위가 급속도로 확산되어 가고 있다. 그러나 최적화 과정에서 발생하는 비수렴성의 문제 등 사용상의 제약이 많으며, 추정된 계수나 예측모형에 대한 해석적 어려움은 가장 큰 단점 중의 하나로 지적되고 있다(Berry & Linoff, 1997). 이 절에서는 사례분석을 통해서 신경망 예측모형의 해석적 어려움을 해결하기 위해 의사결정나무가 효율적으로 사용될 수 있음을 보일 것이다.

이 절의 사례분석에 사용된 자료는 한 은행의 신용평가 부서에서 대출승인에 대한 예측모형을 구축하기 위해 5960명의 고객으로부터 수집된 자료로써, 목표변수인 대출금 상환여부(y)와 대출사유 및 직업 등 17개의 입력변수(x_1, x_2,$\cdots$,x_{17})로 구성되어 있다. 이 자료에서 대출금을 상환하지 않은 고객은 약 20%이다. 하나의 은닉층(hidden layer)에 두 개의 은닉마디(hidden node)를 가지는 매우 간단한 구조의 MLP 신경망을 이 자료에 적용시켜 얻은 사후확률에 대한 예측식은 다음 식 (1)과 같다.

$$f(x) = 1/[1 + \exp(-4.493 + 2.102h_1(x) + 6.699h_2(x))] \quad (1)$$

식 (1)에서 $h_1(x)$와 $h_2(x)$는 각 은닉마디에 대한 활성함수(activation function)를 나타내며, 이를 구체적으로 표현하면 다음과 같다.

$$h_1(x) = 1/[1+\exp(a_{01} + b_{11}x_1 + \cdots + b_{p1}x_p)]$$
$$h_2(x) = 1/[1+\exp(a_{02} + b_{12}x_1 + \cdots + b_{p2}x_p)]$$

위 식에 대해서 자료로부터 추정된 계수값들은 <표 2.1>에 제시되어 있다.

<표 2.1> 은닉마디에 대한 계수값

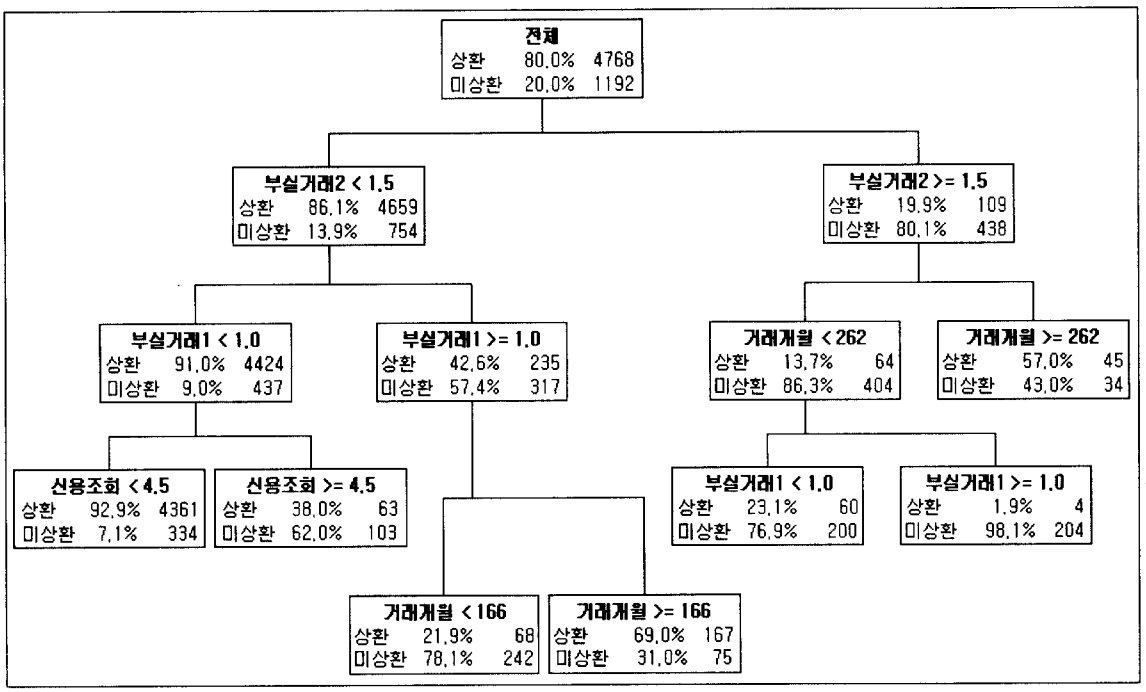
변수	$h_1(x)$	$h_2(x)$	변수	$h_1(x)$	$h_2(x)$
절편항	-0.157	-0.877	자산가치	0.206	0.174
대출사유	0.229	-0.154	부채비율	0.932	0.162
직업1	-0,869	0.015	근무년수	-1.135	0.139
직업2	-0.754	-0.502	부실거래1	2.207	0.162
직업3	0.932	-0.273	부실거래2	1.217	0.388
직업4	-0.753	0.044	금융거래	-1.759	0.174
직업5	1.975	0.344	거래개월	-0.157	-0.389
대출금액	0.000	-0.173	신용조회	1.539	0.051
저당금액	-1.207	-0.005			

한편, 식 (1)에서 $f(x)$는 대출금을 상환하지 않을 확률을 나타내므로, 사전확률(π=0.20)을 고려하여 전체 오분류확률을 최소로 하는 분류규칙은 다음과 같게 된다(Johnson & Wichern, 1992).

$$C = \begin{cases} 1(\text{대출금을상환하지않음}), & f(x) \geq 1-\pi \text{ 이면}, \\ 0(\text{대출금을상환함}), & f(x) < 1-\pi \text{ 이면}, \end{cases} \quad (2)$$

식 (1)과 (2)에서 볼 수 있는 바와 같이 MLP 신경망은 매우 복잡한 비선형 형태를 가지고 있기 때문에 그 결과를 해석하는 것이 쉽지 않다. 이러한 단점을 해결하는 한 가지 방법으로 의사결정나무를 이용하여 근사 분류규칙을 만들고 이를 살펴봄으로써 신경망 분류규칙의 성격을 부분적으로 파악할 수 있다. 다음 <그림 2.1>은 식 (2)와 같은 분류규칙을 목표변수로 하여 생성된 의사결정나무의 일부를 보여주고 있다. 식 (2)의 분류규칙과 이 의사결정나무 분류규칙의 분류일치도(전체 개체들 중 두 분류규칙이 동일한 목표값으로 분류하는 개체들의 비율)는 약 92%이고, 따라서 이 경우 의사결정나무가 신경망 분류규칙을 잘 근사한다고 할 수 있다. 또한 의사결정나무로부터 부실거래의 수가 많고, 거래개월수가 적으며, 신용조회수가 많을수록 신경망 예측모형에 의한 미상환 확률 $f(x)$가 크다는 것을 쉽게 추측할 수 있다.

<그림 2.1> 신경망 예측모형에 대한 근사 의사결정나무



3. 자율예측모형의 해석에 대한 사례연구

자율예측은 넓은 의미에서는 흔히 목표변수가 존재하지 않는 경우의 분석을 일컫는데, 대표적인 분석방법들로는 연결분석(link analysis), 연관성 규칙발견(association rule discovery), 군집분석(cluster analysis) 등이 있다. 좁은 의미에서의 데이터마이닝은 소위 세분화(segmentation)라고 불리기도 하는 군집분석과 거의 유사한 의미로 사용된다.

데이터마이닝에서 사용되는 대표적인 군집분석 방법으로는 k-평균 군집분석, SOM(self organizing maps or Kohonen network; Kohonen, 1997), k-최단근접 등과 같은 방법이 사용된다. 데이터마이닝에서는 일반적으로 군집분석에 사용되는 변수의 수가 매우 많게 되는데, 따라서 그 결과를 해석하는 것이 수월하지 않다(Berry & Linoff, 1997). 이 절에서는 앞에서와 마찬가지로 사례분석을 통해서 군집분석의 해석적 어려움을 해결하기 위해 의사결정나무가 효율적으로 사용될 수 있음을 보일 것이다.

이 절의 사례분석에 사용된 자료는 한 백화점에서 10,000명의 고객에게 광고인쇄물(DM: Direct Mail)을 발송하고 그 반응을 조사하여 얻은 자료이다. 이 자료에는 인구·사회적 속성(성, 연령, 결혼여부, 년수입, 거주지, 집의 소유여부), 거래속성(할인고객여부, 7개 제품유형별 구입금액, 총 구입금액) 등이 포함되어 있다.

여기서는 코호넨 신경망을 이용하여 군집분석을 수행하였으며, 편의상 군집의 개수는 5개로 하였다. 군집분석을 수행한 결과 각 군집의 크기는 75, 3143, 6233, 334, 206이고, 5개 군집의 프로파일은 <표 3.1>에 요약되어 있다.

<표 3.1> 각 군집의 변수별 빈도 및 평균

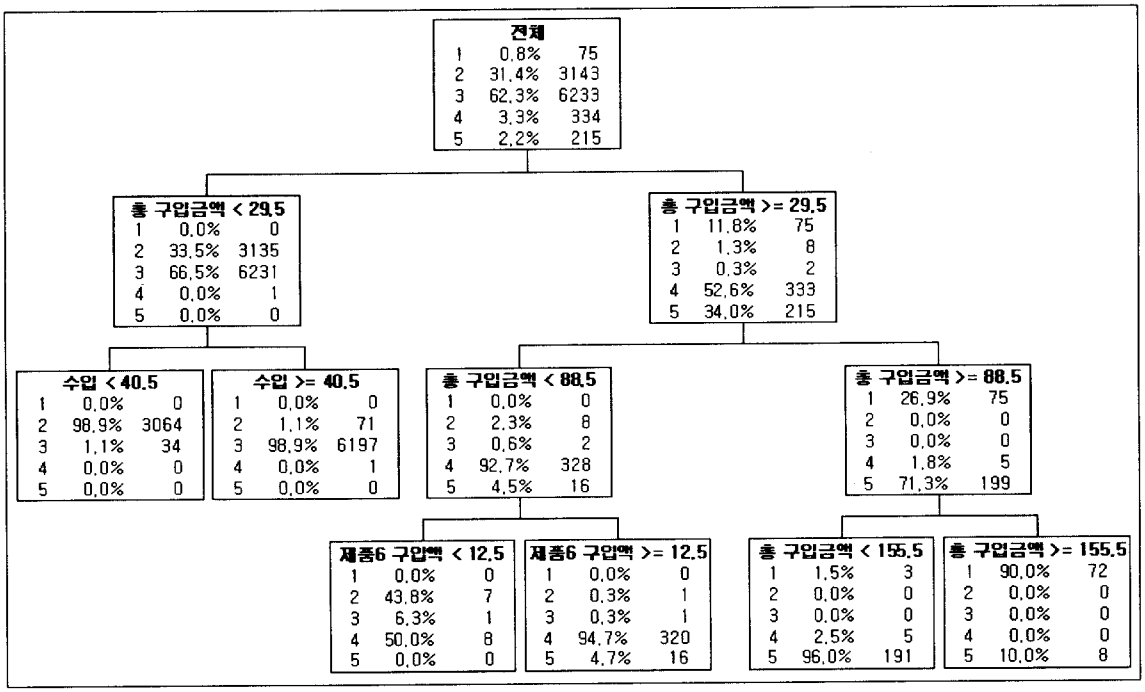
군집	성		결혼여부		집의 소유여부		할인고객여부	
	남자	여자	미혼	결혼	미소유	소유	일반	할인
1	33(48%)	36(52%)	23(33%)	46(67%)	51(74%)	18(26%)	55(73%)	20(26%)
2	1870(60%)	1224(40%)	1263(41%)	1831(59%)	2258(73%)	836(27%)	2287(73%)	856(27%)
3	2339(39%)	3741(61%)	2588(43%)	3492(57%)	3804(63%)	2276(37%)	4540(73%)	1693(27%)
4	159(50%)	158(50%)	121(38%)	196(62%)	229(72%)	88(28%)	253(76%)	81(24%)
5	88(43%)	118(57%)	66(32%)	140(68%)	161(78%)	45(22%)	265(77%)	50(23%)

군집	거주지							
	A	B	C	D	E	F	G	H
1	4(5%)	11(15%)	6(8%)	2(3%)	18(24%)	8(11%)	13(17%)	13(17%)
2	177(6%)	561(18%)	215(7%)	211(7%)	652(21%)	604(19%)	342(11%)	381(12%)
3	370(6%)	1161(19%)	288(5%)	294(5%)	1478(24%)	1454(23%)	548(9%)	640(10%)
4	23(7%)	60(18%)	30(9%)	32(10%)	67(20%)	65(20%)	20(6%)	37(11%)
5	11(5%)	35(16%)	22(10%)	15(7%)	46(21%)	37(17%)	27(13%)	22(10%)

군집	나이 (평균)	수입 (평균)	구입금액(평균)							
			제품1	제품2	제품3	제품4	제품5	제품6	제품7	전체
1	40.86	45.88	5.65	30.30	22.05	17.73	16.17	110.32	8.28	210.52
2	45.13	29.23	0.00	0.01	0.04	0.07	0.00	0.21	0.00	0.36
3	44.51	57.49	0.00	0.03	0.00	0.01	0.00	0.15	0.00	0.22
4	41.45	50.19	1.53	11.41	4.23	4.18	2.00	34.48	0.86	58.71
5	43.34	44.59	2.74	17.50	11.06	10.59	7.45	62.81	3.70	115.89

다음 <그림 3.1>은 군집을 나타내는 변수를 목표변수로 하여 생성된 의사결정나무의 일부를 보여주고 있으며, 군집결과와 이 의사결정나무의 분류일치도는 약 99%이다. <그림 3.1>로부터 총 구입금액, 수입, 제품6의 구입금액 등이 군집결과에 큰 영향을 주고 있음을 알 수 있다. 예를 들어, 총 구입금액이 295,000원 이하이면서 수입이 400백만원 이하인 고객들은 대부분 군집2에 속한다는 것을 쉽게 파악할 수 있다. 연구자는 이와 같은 결과를 기초로 하여 <표 3.1>과 같은 결과를 살펴봄으로써 보다 쉽게 군집분석의 결과를 파악할 수 있을 것이다.

<그림 3.1> 군집분석 결과에 대한 근사 의사결정나무



4. 결론

본 논문에서는 지도예측 또는 자율예측의 결과를 해석함에 있어서 의사결정나무가 유용하게 활용될 수 있음을 살펴보았다. 즉, 예측모형에 대한 근사 의사결정나무를 구성함으로써 어느 변수들이 예측모형에 큰 영향을 주고 있으며 변수들 간의 상호작용은 어떻게 이루어지고 있는 지 등을 보다 쉽게 파악할 수 있다. 실제 문제에 있어서 의사결정나무가 데이터마이닝 예측모형을 항상 잘 근사하는 것은 아니다. 그러나 예측모형에 대한 해석을 왜곡시킬 정도로 의사결정나무에 의한 근사정도(분류일치도)가 작은 경우가 아니라면, 이와 같은 시도가 현실의 문제를 해결함에 있어서 연구자에게 많을 도움을 줄 것으로 기대된다.

참 고 문 헌

[1] 최종후 · 한상태 · 강현철 · 김은석(1998). 「AnswerTree를 이용한 데이터마이닝 의사결정나무 분석」, SPSS 아카데미, 서울.

[2] Berry, M. J. A. and Linoff, G.(1997). *Data Mining Techniques for Marketing, Sales, and Customer Support*, John Wiley & Sons, New York.

[3] Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone.(1984). *Classification and regression trees*. Wadsworth.

[4] Johnson, R. A. and Wichern, D. W.(1992). *Applied Multivariate Statistical Analysis*, Prentice-Hall.

[5] Kass, G.(1980). An exploratory technique for investigating large quantities of categorical data, *Applied Statistics*, Vol. 29, 119-129.

[6] Kohonen, T.(1997), *Self-Organizing Maps*, Springer-Verlag, Berlin.

[7] Loh, W., Shih, Y.(1997). Split selection methods for classification trees, *Statistica Sinica*, Vol. 7, 815~840.

[8] Quinlan, J. R.(1993). *C4.5 Programs for machine learning*. Morgan Kaufmann, San Mateo.

[9] Smith, M.(1996). *Neural Networks for Statistical Modeling*, International ThomsonComputer Press.