Journal of the Korean Data & Information Science Society
2012, **23**(2), 299–307

한국데이터정보과학회지

주변조건부 변수를 이용한 의사결정나무모형 생성에 관한 연구

조광현[1] · 박희창[2]

[1]창원대학교 유아교육학과 · [2]창원대학교 통계학과

접수 2012년 2월 8일, 수정 2012년 3월 12일, 게재확정 2012년 3월 16일

요 약

데이터마이닝은 주어진 데이터베이스에서 항목간의 흥미로운 관계를 찾아내는 기법으로서 의사결정나무는 데이터마이닝의 대표적인 알고리즘이라고 할 수 있다. 의사결정나무는 관심대상이 되는 집단을 몇 개의 소집단으로 분류하거나 예측을 수행하는 방법이다. 일반적으로 연구자가 의사결정나무 모형을 생성 할 때 모형 생성의 기준 및 입력 변수의 수에 따라 복잡한 모형이 생성되기도 한다. 특히 의사결정나무 모형에서 입력 변수의 수가 많을 경우 생성된 모형은 복잡한 형태가 될 수 있고, 모형 분석이 어려울 수도 있다. 만일 입력변수에서 주변조건부 변수 (매개변수, 외적변수)가 존재한다면 이 입력변수는 직접적인 관련성이 없는 것으로 판단한다. 이에 본 논문에서는 주변조건부 변수를 고려하여 의사결정나무모형을 생성하는 방법을 제시하고 그 효율성을 파악하기 위하여 실제 자료에 적용하고자 한다.

주요용어: 데이터마이닝, 매개변수, 외적변수, 의사결정나무, 주변조건부변수.

1. 서론

데이터마이닝이란 대용량의 관측 가능한 데이터를 기반으로 숨겨진 지식, 기대하지 못했던 패턴, 새로운 법칙과 관계를 발견하고 이를 바탕으로 의사결정 등을 위한 정보로 활용하고자 하는 것으로서, 연관성규칙, 군집분석, 의사결정나무 등의 다양한 기법들이 있다. 이 중 의사결정나무는 데이터마이닝의 대표적인 알고리즘으로서 의사결정 규칙을 도표화하여 관심대상이 되는 집단을 몇 개의 소집단으로 분류하거나 예측을 수행하는 분석방법이다. 의사결정나무는 의사결정나무분석을 수행하기 위한 다양한 분리기준, 정지규칙, 가지치기 방법들이 제안되었으며, 이들을 어떻게 결합하느냐에 따라서 서로 다른 의사결정나무가 형성된다. 의사결정나무의 대표적인 알고리즘에는 Hartigan (1975)의 CHAID, Breiman 등 (1984)의 CART, Quinlan (1993)의 C5.0 등의 알고리즘 있다. 현재 Park (2010), Cho와 Park (2011b, 2011c) 등에 의하여 의사결정나무 모형의 모형 구축 시간 단축 및 생성된 모형 정확성 등을 높이기 위한 하이브리드 데이터마이닝의 연구가 진행되고 있다.

의사결정나무의 모형 생성 시, 모형 생성의 기준 및 입력변수의 수에 따라 복잡한 모형이 생성되기도 한다. 특히 입력변수의 분리 기준에 따라 나무 모형이 생성되므로 입력 변수가 많은 경우, 나무 모형이 복잡해 질 수밖에 없으므로 종종 모형 생성 및 해석에 있어 어려움을 겪기도 한다. 이때 생성된 모형에 대한 목표변수와 입력변수와의 관계에서 두 변수의 관계가 우연히 어떤 다른 변수와 연결됨으로써

[1] (641-773) 경상남도 창원시 사림동 9번지, 창원대학교 유아교육학과, 통계학 시간 강사.
[2] 교신저자: (641-773) 경상남도 창원시 사림동 9번지, 창원대학교 통계학과, 교수.
E-mail: hcpark@changwon.ac.kr

관련성이 있는 것으로 나타나는 경우, 실제적으로 두 변수 간에는 관련성이 없으나 관련성이 있는 것으로 해석하는 오류를 범할 수 있다. 일반적으로 목표변수와 입력변수에 대한 상호 관련성을 분석할 때, 변수들 간의 내재적인 관련성은 없고 각 변수가 우연히 어떤 주변조건부변수 (marginally conditional variables)와 연결됨으로써 관련성이 있는 것으로 나타나는 경우가 발생할 수 있다. 이 경우 주변조건부변수에 의하여 목표변수와 입력변수의 관계가 실제적으로 무의미한 관계라고 한다면 모형 생성 시 그 입력변수를 제거하고 모형을 생성하는 것이 효과적일 것이다.

이에 본 논문에서는 의사결정나무 생성 시, 목표변수와 입력변수 사이에 주변조건부변수에 의하여 내재적인 관련성이 없는 입력변수를 파악하여 이를 찾아내는 방법에 대하여 연구하고자 하며, 그 효율성을 파악하기 위하여 실제 자료에 그 방법을 적용하고자 한다. 본 논문의 구성은 다음과 같다. 논문의 2절에서는 논문의 연구 배경에 대하여 기술하고 3절에서는 실제 자료를 통하여 본 논문에서 제안하는 방법의 효과를 살펴본 후, 4절에서 결론을 맺고자 한다.

2. 연구 배경

2.1. 주변조건부변수

일반적으로 독립변수와 종속변수에 대한 상호 관련성을 분석할 때, 각 변수가 우연히 어떤 주변 조건부 변수와 연결됨으로써 관련성이 있는 것으로 나타나는 경우가 발생할 수 있다. 여기서 주변 조건부 변수란 독립변수와 종속변수 사이에 영향을 미치는 변수들로서 매개변수, 외적변수, 구성변수, 선행변수, 억제변수, 왜곡변수 등이 있다. 독립변수와 종속변수 사이에 주변 조건부 변수 중 매개변수와 외적변수가 존재하는 경우 두 변수 간에는 실제적인 관련성이 없으나 이 주변 조건부 변수에 의하여 관련성이 있는 것으로 나타날 수 있다. 이 경우 두 변수간의 관련성을 분석한다면 잘못된 해석을 내릴 수 있다. 주변 조건부 변수 중 간접적 해석에 관여되는 변수는 대표적으로 매개변수와 외적변수가 있다. 매개변수는 독립 변수와 종속 변수 사이에서 독립 변수의 결과인 동시에 종속 변수의 원인이 되는 변수를 의미하고 외적변수는 독립변수와 종속변수 사이에서 우연히 어떤 다른 변수와 연결됨으로써 관계가 있는 것처럼 이 변수를 통제하면 관계가 사라지게 되는 변수를 외적변수라고 한다.

Cho와 Park (2011a)은 연관성 규칙을 이용하여 매개변수를 추출하는 방법에 대하여 연구한 바 있으며, 그 조건은 다음과 같다. 여기서, Y는 후항변수, X_1은 전항변수, 그리고 X_2는 매개변수라고 한다.

[조건 1] Y와 X_1에 대한 연관성규칙의 결과가 지정된 최소 지지도와 최소 신뢰도보다 커야 한다.

[조건 2] X_1과 X_2에 대한 연관성규칙의 결과가 지정된 최소 지지도와 최소 신뢰도보다 커야 한다.

[조건 3] X_1및 X_2와 Y와의 연관성규칙의 결과가 지정된 최소 지지도와 최소 신뢰도보다 커야 한다.

[조건 4] X_1및 X_2와 Y와의 연관성규칙의 신뢰도가 Y와 X_1에 대한 연관성규칙의 신뢰도보다 커야 한다.

Cho와 Park (2011d)은 연관성 규칙을 이용하여 외적변수를 추출하는 방법에 대하여 연구한 바 있으며, 그 조건은 다음과 같다. 여기서 X는 전항변수, Y는 후항변수, Z는 외적변수라고 한다.

[조건 1] 변수 X와 변수 Y에 대한 연관성이 존재해야 한다.

[조건 2] 변수 Z와 변수 X에 대한 연관성이 존재해야 한다.

[조건 3] 변수 Z와 변수 Y에 대한 연관성이 존재해야 한다.

[조건 4] 변수 Z를 통제했을 때, 변수 X와 변수 Y에 대한 연관성이 존재하지 않는다.

2.2. 연구 방법

일반적으로 의사결정나무 모형 생성 시, 모형 생성의 기준 및 입력변수의 수에 따라 의사결정나무 모형이 생성되므로 종종 복잡한 의사결정나무 모형이 생성되기도 한다. 특히 목표 변수에 대한 입력 변수의 분리 기준에 따라서 의사결정나무 모형이 생성되므로 목표 변수에 유의한 입력 변수의 수가 많은 경우 의사결정나무 모형이 복잡해 질 수밖에 없으므로 모형 생성 및 해석에 있어 어려움을 격기도 한다. 그러나 생성된 모형에 대한 목표 변수와 입력 변수와의 관계가 다른 주변의 조건부 변수에 의하여 실제적인 관련성이 없는 것으로 판단된다면 모형 생성 시 실제로 목표변수에 무의미한 입력 변수를 제거하고 모형을 생성하는 것이 효과적일 것이다. 이에 본 논문에서는 의사결정나무 생성 시 목표변수와 입력변수 사이의 주변조건부변수가 존재하는지 파악하여 불필요한 입력변수를 제거하는 방법을 연구를 하고자 한다. 주변조건부변수를 이용한 의사결정나무모형 생성 방안은 그림 2.1과 같다.

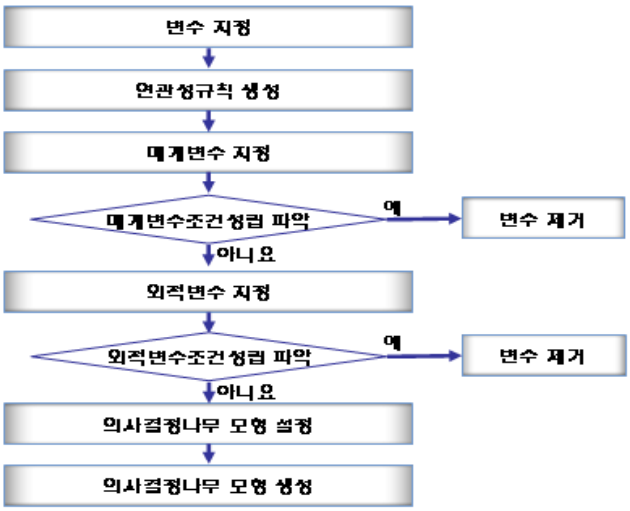

그림 2.1 의사결정나무모형 생성 방안

그림 2.1을 자세하게 설명하면 다음과 같다.

[단계 1] 변수 지정

의사결정나무 모형을 생성하기 위하여 목표 변수와 입력 변수를 결정한다.

[단계 2] 연관성규칙 생성

목표변수와 입력변수에 대한 연관성 규칙의 조건인 최소지지도, 최소신뢰도, 향상도를 지정하여 연관성 규칙을 생성한다.

[단계 3] 매개변수 지정

생성된 규칙에 대한 매개변수가 존재하는 가를 파악하기 위하여 매개변수를 지정한다. 여기서, 매개변수는 생성된 규칙들 중 목표변수가 동일한 입력변수 중에서 선택하도록 한다.

[단계 4] 매개변수 조건 성립 파악

지정된 매개변수에 대하여 목표변수와 입력변수 사이에 실제로 매개변수 조건이 성립하는 가에 대하여 파악한다. 매개변수 조건은 2.1절에서 설명한 네 가지 조건을 만족해야 하며 네 가지 조건을 모두 만

족하는 경우 이 입력변수는 지정된 매개변수에 의하여 의미 없는 입력변수로 판단할 수 있으며, 변수를 삭제할 수 있다. 조건 성립을 위하여 다음의 단계를 거친다.

[단계 4-1] 입력변수와 매개변수와의 연관성 규칙 생성

[단계 4-2] 입력변수 및 매개변수와 목표변수와의 연관성 규칙 생성

[단계 4-3] 단계 2와 단계 4-2의 연관성 규칙 결과 비교

[단계 5] 외적변수 지정

생성된 규칙에 대한 외적변수가 존재하는 가를 파악하기 위하여 외적변수를 지정한다. 외적변수는 규칙에 사용된 목표변수와 입력변수를 제외한 나머지 변수들로 지정한다.

[단계 6] 외적변수 조건 성립 파악

지정된 외적변수에 대하여 목표변수와 입력변수 사이에 실제로 외적변수 조건이 성립하는 가에 대하여 파악한다. 외적변수 조건은 2.1절에서 설명한 네 가지 조건을 만족해야 하며 네 가지 조건을 모두 만족하는 경우 이 입력변수는 지정된 외적변수에 의하여 의미 없는 입력변수로 판단할 수 있으며, 변수를 삭제할 수 있다. 조건 성립을 위하여 다음의 단계를 거친다.

[단계 6-1] 외적변수와 입력변수와의 연관성 규칙 생성

[단계 6-2] 외적변수와 목표변수와의 연관성 규칙 생성

[단계 6-3] 외적변수 통제 후 입력변수와 목표 변수와의 연관성 규칙 생성

[단계 6-4] 단계 2와 단계 6-3의 연관성 규칙 결과 비교

[단계 7] 모형 설정

입력변수와 목표변수 사이에 주변조건부변수의 성립 여부를 파악한 뒤 주변조건부변수가 성립하는 경우의 해당하는 입력 변수를 제거하고 모형을 설정한다. 모형 설정에서는 자료 분할, 모형 알고리즘 선택, 정지 규칙 등을 지정한다.

[단계 8] 모형 생성

지정된 모형에 의하여 모형을 생성한다. 생성된 모형에 대한 예측정확도 및 모형평가 예측정확도를 살펴본 뒤 모형에 대한 해석을 실시한다.

3. 자료 분석

본 절에서는 본 논문에서 제안하는 방법에 대한 효율성을 파악하기 위하여 2010년 C대학교에서 조사한 갱년기 여성의 건강 및 생활환경에 대한 조사 자료를 이용하였다. 설문 문항은 총 59문항으로 구성되어 있고, 조사 대상자는 경상남도에 거주하는 50~ 60대 여성이며, 분석에 사용한 자료 건수는 571명이다. 분석에 사용된 변수는 갱년기 유무, 운동 시간, 수면 시간, 결혼 만족도, 부부 친밀도, 나이, 직업, 학력, 전반적 건강상태, 평상시 식생활의 10개 문항을 추출하였고 갱년기 유무를 목표변수로 지정하였으며, 나머지 9개 문항을 입력변수로 지정하였다. 주변조건부변수를 파악하기 위하여 편의상 비율자료는 평균을 바탕으로 이분형으로 변환한 뒤 분석을 실시하였다. 분석에 사용한 변수는 표 3.1과 같다.

본 논문에서는 기존의 의사결정나무 원 모형과 본 논문에서 제시하는 의사결정나무 모형의 두 가지 모형을 생성 한 뒤, 두 모형을 비교하고자 한다. 첫 번째로 목표변수인 갱년기 유무와 입력변수 9개 문항에 대한 기존의 의사결정나무 모형을 생성한다. 본 논문에서는 입력 변수가 이분형 자료이므로 이지 분리가 가능한 CART 알고리즘으로 의사결정나무 모형을 생성하였다. 모형 생성 시, 훈련 자료 (2/3)와 모형평가 자료 (1/3)로 분할하여 모형을 생성하였으며 생성된 모형은 그림 3.1과 같다.

표 3.1 변수 설명

변수명	구분	설명
갱년기 유무	목표변수	범주 1 : 예, 범주 2 : 아니오
운동 시간	입력변수	범주 1 : 적음, 범주 2 : 많음
수면 시간	입력변수	범주 1 : 적음, 범주 2 : 많음
결혼 만족도	입력변수	범주 1 : 낮음, 범주 2 : 높음
부부 친밀도	입력변수	범주 1 : 낮음, 범주 2 : 높음
나이	입력변수	범주 1 : 적음, 범주 2 : 많음
직업	입력변수	범주 1 : 가정주부, 범주 2 : 기타
학력	입력변수	범주 1 : 낮음, 범주 2 : 높음
전반적 건강상태	입력변수	범주 1 : 나쁨, 범주 2 : 좋음
평상시 식생활	입력변수	범주 1 : 나쁨, 범주 2 : 좋음

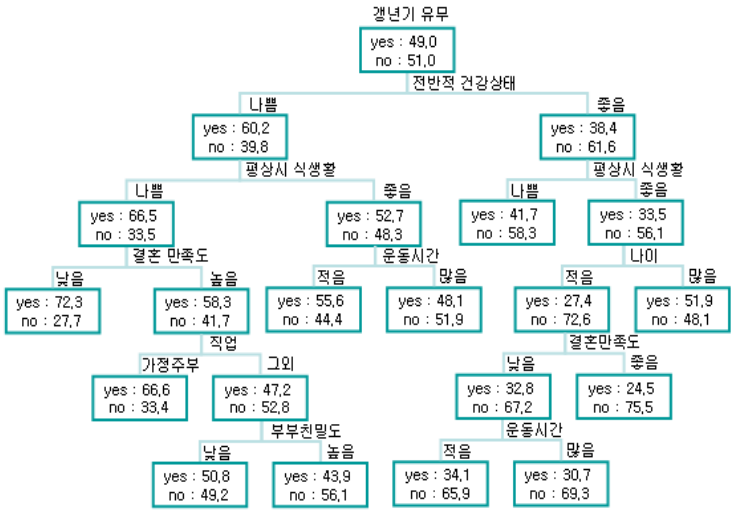

그림 3.1 기존의 의사결정나무 (원 모형)

다음으로 갱년기 유무를 목표변수로 지정하고 운동 시간, 수면 시간, 결혼 만족도, 부부 친밀도, 나이, 직업, 학력, 전반적 건강상태, 평상시 식생활의 9개 문항을 입력변수를 지정하였을 경우, 주변조건부변수의 성립 여부를 파악한 후, 의사결정나무 모형을 생성한다. 입력 변수에 대한 주변조건부변수인 매개변수와 외적변수의 성립여부를 파악할 결과, 주변조건부변수인 매개변수가 성립하는 입력변수가 한개 존재하는 것으로 나타났다 (표 3.2). 또한 외적변수가 성립하는 입력변수가 한개 존재하는 것으로 나타났다 (표 3.3). 여기서 주변조건부변수에 대한 연관성 규칙의 적용 기준은 최소 지지도를 10, 최소 신뢰도를 70, 향상도를 1로 지정하였다 (표에서는 연관성 기준 중 신뢰도만 표시함).

표 3.2 주변조건부변수인 매개변수 성립여부에 대한 결과

조건	입력변수	매개변수	목표변수	신뢰도
1	운동 시간	-	갱년기 유무	74.3
2	운동 시간	전반적 건강상태	-	73.5
3	-	전반적 건강상태	갱년기 유무	76.9
4	운동 시간	전반적 건강상태	갱년기 유무	81.1

표 3.2를 자세하게 살펴보면, 조건 1에서 입력변수와 목표변수와의 관련성이 존재하고, 조건 2에서 입력변수와 매개변수와의 관련성이 존재하며, 조건 3에서 매개변수와 목표변수와의 관련성이 존재한다. 또한 입력변수 및 매개변수와 목표변수와의 관련성이 81.1이고 입력변수와 목표변수와의 관련성이 74.3이므로 조건 4를 만족한다고 할 수 있어, 입력변수인 운동시간과 목표변수인 갱년기 유무 사이에 매개변수인 전반적 건강상태가 존재한다고 할 수 있다. 즉, 입력변수인 운동시간은 매개변수인 전반적 건강상태에 의하여 간접적인 해석만 가능하므로 의미가 없는 변수로 판단되어 9개의 입력변수 중 운동시간은 제거한다. 다음으로 주변조건부변수인 외적변수 성립여부에 대한 결과는 표 3.3과 같다.

표 3.3 주변조건부변수인 외적변수 성립여부에 대한 결과

조건	외적변수	입력변수	목표변수	신뢰도
1	-	부부친밀도	갱년기 유무	71.2
2	결혼만족도	부부친밀도	-	72.4
3	결혼만족도	-	갱년기 유무	72.5
4	결혼만족도	부부친밀도	갱년기 유무	60.1

표 3.3을 자세하게 살펴보면, 조건 1에서 입력변수와 목표변수와의 관련성이 존재하고, 조건 2에서 외적변수와 입력변수와의 관련성이 존재하며, 조건 3에서 외적변수와 목표변수와의 관련성이 존재한다. 또한 조건 4에서 외적변수를 통제했을 때 입력변수와 목표변수와 신뢰도 값이 60.1로 최소 신뢰도의 기준값인 70보다 작으므로 관련성이 존재하지 않아 외적변수의 모든 조건을 만족한다고 볼 수 있다. 즉, 입력변수인 부부친밀도와 목표변수인 갱년기 유무 사이에 외적변수인 결혼만족도가 존재한다고 볼 수 있으므로 입력변수인 부부친밀도는 외적변수인 결혼만족도에 의하여 우연히 나타난 규칙이므로 의미가 없는 변수로 판단되어 9개의 입력변수 중 부부친밀도는 제거한다. 이를 종합하면 위에서 설명한 9개의 입력변수 중 운동시간과 부부친밀도는 주변조건부변수에 의하여 의미 없는 입력변수로 나타났다. 이에 본 논문에서는 9개 문항의 입력 변수 중 운동시간과 부부친밀도를 제외한 7개 문항을 입력 변수로 지정하여 위의 원 모형과 동일한 조건으로 의사결정나무 모형을 생성하였다. 생성된 모형은 그림 3.2와 같다.

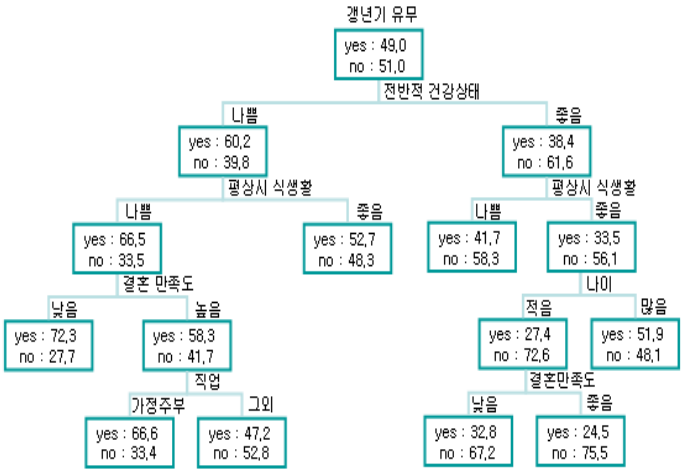

그림 3.2 주변조건부변수를 이용한 의사결정나무 모형

갱년기 유무를 목표변수로 하는 의사결정나무의 원 모형인 그림 3.1과 본 논문에서 제시하는 주변조건부변수를 이용한 의사결정나무 모형인 그림 3.2를 비교하면 표 3.4와 같다.

표 3.4 모형 비교

구분	원 모형	제안 모형
노드의 깊이	6	5
생성된 노드의 수	21	15
끝마디 노드 수	8	7

표 3.4을 살펴보면 노드의 깊이가 6개에서 5개로 줄어들었고 생성된 노드의 수 또한 21에서 15개로 줄어든 것을 알 수 있었으며, 끝마디 노드 수 또한 8개에서 7개로 줄어든 것을 알 수 있다. 이는 불필요한 가지를 생성하지 않으므로 모형의 생성과 생성된 모형의 해석 시 시간과 노력을 단축할 수 있다. 그러나 생성된 모형이 원 모형에 비하여 간결해 졌지만 모형의 정확도가 현저하게 차이가 난다면 이는 좋은 모형이라고 할 수 없다. 이에 본 논문에서는 표 3.5에서와 같이 기존의 나무 모형과 본 논문에서 제시하는 나무 모형의 정확도를 비교하였다. 표 3.5를 살펴보면 주변조건부변수를 이용한 모형의 모형 예측정확도 및 모형평가 예측정확도가 원 모형의 모형 예측정확도 및 모형평가 예측 정확도와 큰 차이를 보이고 있지 않은 것을 알 수 있다. 이에 본 논문에서 제시하는 의사결정나무모형 생성의 방법이 모형의 정확도는 거의 동일하면서 불필요한 가지를 생성하지 않으므로 효율적이라고 할 수 있다.

표 3.5 모형의 정확도 비교

모형	원 모형		제안 모형	
	모형 예측정확도	모형평가 예측정확도	모형 예측정확도	모형평가 예측정확도
퍼센트	68.7%	69.4%	67.2%	68.3%

4. 결론

일반적으로 모형 생성의 기준 및 입력 변수의 수에 따라 의사결정나무 모형이 생성되므로 종종 복잡한 의사결정나무 모형이 생성되기도 한다. 특히 목표 변수에 대한 입력 변수의 분리 기준에 따라서 의사결정나무 모형이 생성되므로 목표 변수에 유의한 입력 변수의 수가 많은 경우 의사결정나무 모형이 복잡해 질 수밖에 없으므로 모형 생성 및 해석에 있어 어려움을 격기도 한다. 이때 생성된 모형에 대한 목표변수와 입력 변수와의 관계에서 두 변수의 관계가 우연히 어떤 주변조건부변수에 의하여 실제적으로 무의미한 관계라고 한다면 모형 생성 시 그 입력 변수를 제거하고 모형을 생성하는 것이 효과적이다. 이에 본 논문에서는 의사결정나무 생성 시 목표변수와 입력변수 사이의 주변조건부변수가 존재하는가를 파악하여 불필요한 입력변수를 제거하는 방법을 연구하였고, 실제 자료에 적용해 보았다. 분석 결과, 목표변수와 입력 변수 사이에 무의미한 입력 변수를 제거함으로서 기존의 모형에 비하여 노드의 깊이나 노드의 수가 줄어든 것을 알 수 있으며, 기존의 모형에 비해서도 모형의 정확도가 큰 차이가 나지 않으므로 본 논문에서 제시하는 방법이 효율적이라고 할 수 있다. 향후 과제로 본 논문에서 제안하는 방법을 국가통계, 기업체 및 연구 자료 등에 다양하게 적용하여 변수들 간의 관계를 명확하게 규명할 필요성이 있으며, 의사결정나무 뿐만 아니라 신경망분석, 로지스틱회귀분석 등에도 적용하고 이를 서로 비교하는 연구도 필요할 것이다.

참고문헌

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

Cho, K. H. and Park, H. C. (2011a). A study on insignificant rules discovery in association rule mining. *Journal of the Korean Data & Information Science Society*, **22**, 81-88.

Cho, K. H. and Park, H. C. (2011b). A study on decision tree creation using intervening variable. *Journal of the Korean Data & Information Science Society*, **22**, 671-678.

Cho, K. H. and Park, H. C. (2011c). A study on removal of unnecessary input variables using multiple external association rule. *Journal of the Korean Data & Information Science Society*, **22**, 877-884.

Cho, K. H. and Park, H. C. (2011d). Discovery of insignificant association rules using external variable. *Journal of the Korean Data Analysis Society*, **13**, 1343-1352.

Hartigan, J. A. (1975). *Clustering algorithms*, John Wiley & Sons, New York.

Park, H. C. (2010). Association rule ranking function by decreased lift influence. *Journal of the Korean Data & Information Science Society*, **21**, 397-405.

Quinlan, J. R. (1993). *C4.5 programs for machine learning,* Morgan Kaufmann Publishers, San Francisco.

Journal of the Korean Data & Information Science Society
2012, **23**(2), 299–307

한국데이터정보과학회지

A study on decision tree creation using marginally conditional variables

Kwang-Hyun Cho[1] · Hee-Chang Park[2]

[1]Department of Early Childhood Education, Changwon National University
[2]Department of Statistics, Changwon National University

Received 8 February 2012, revised 12 March 2012, accepted 16 March 2012

Abstract

Data mining is a method of searching for an interesting relationship among items in a given database. The decision tree is a typical algorithm of data mining. The decision tree is the method that classifies or predicts a group as some subgroups. In general, when researchers create a decision tree model, the generated model can be complicated by the standard of model creation and the number of input variables. In particular, if the decision trees have a large number of input variables in a model, the generated models can be complex and difficult to analyze model. When creating the decision tree model, if there are marginally conditional variables (intervening variables, external variables) in the input variables, it is not directly relevant. In this study, we suggest the method of creating a decision tree using marginally conditional variables and apply to actual data to search for efficiency.

Keywords: Data mining, decision tree, external variable, intervening variable, marginally conditional variables.

[1] A part-time lecturer, Department of Early Childhood Education, Changwon National University, Changwon, Gyeongnam 641-773, Korea.
[2] Corresponding author: Professor, Department of Statistics, Changwon National University, Changwon, Gyeongnam 641-773, Korea. E-mail: hcpark@changwon.ac.kr