Journal of the Korean Data & Information Science Society
2011, **22**(5), 877–884

한국데이터정보과학회지

다중외적연관성규칙을 이용한 불필요한 입력변수 제거에 관한 연구

조광현[1] · 박희창[2]

[1]창원대학교 유아교육학과 · [2]창원대학교 통계학과

접수 2011년 7월 18일, 수정 2011년 8월 11일, 게재확정 2011년 8월 21일

요 약

의사결정나무는 데이터마이닝의 대표적인 알고리즘으로서, 의사결정 규칙을 도표화하여 관심대상이 되는 집단을 몇 개의 소집단으로 분류하거나 예측을 수행하는 방법이다. 일반적으로 의사결정나무의 모형 생성 시, 입력 변수의 수가 많을 경우 생성된 의사결정모형은 복잡한 형태가 될 수 있고, 모형 탐색 및 분석에 있어 어려움을 겪기도 한다. 이때 입력변수들 간의 내재적인 관련성은 없으나, 외적 변수에 의하여 각 변수가 우연히 어떤 다른 변수와 연결됨으로써 관련성이 있는 것으로 나타나는 것을 종종 볼 수 있다. 이에 본 논문에서는 의사결정나무 생성 시, 입력 변수에 대한 외적 관계를 파악할 수 있는 다중외적연관성규칙을 이용하여 의사결정나무 생성에 불필요한 입력변수를 제거하는 방법을 제시하고 그 효율성을 파악하기 위하여 실제 자료에 적용하고자 한다.

주요용어: 데이터마이닝, 신뢰도, 외적변수, 연관성규칙, 의사결정나무.

1. 서론

의사결정나무 (decision tree)는 의사결정 규칙을 도표화하여 관심대상이 되는 집단을 몇 개의 소집단으로 분류하거나 예측을 수행하는 분석방법이다. 의사결정나무는 다른 분석방법에 비해 연구자가 분석과정을 쉽게 이해하고 설명할 수 있다는 장점이 있어 분류 (classification)와 예측 (prediction)을 수행하는 분석방법 중 강력하고 많이 사용하는 방법이라고 할 수 있다. 의사결정나무의 대표적인 알고리즘에는 Hartigan (1975)의 CHAID (Chi-squared Automatic Interaction Detection), Breiman 등 (1984)의 CART (Classification and Regression Trees), Quinlan (1993)의 C5.0 등의 알고리즘 있다.

일반적으로 의사결정나무의 모형 생성 시, 입력 변수의 수가 많을 경우 종종 모형 생성 및 해석에 있어 어려움을 격기도 하며, 모형 생성의 기준 및 입력 변수의 수에 따라 복잡한 모형이 생성되기도 한다. 이때 생성된 모형에 대한 목표 변수와 입력 변수와의 관계에서 두 변수의 관계가 우연히 어떤 다른 변수와 연결됨으로써 관련성이 있는 것으로 나타나는 경우가 발생하여 실제적으로 두 변수 간에는 관련성이 없으나 관련성이 있는 것으로 해석하는 오류를 범할 수 있다. 이 경우 외적 변수 (external variable)에 의하여 목표변수와 입력변수의 관계가 실제적으로 무의미한 관계라고 한다면 모형 생성 시 그 입력 변수를 제거하고 모형을 생성하는 것이 효과적일 것이다. 이에 본 논문에서는 의사결정나무 생성 시, 목표변수와 입력 변수 사이에 외적 변수를 명확하게 파악할 수 있는 다중외적연관성규칙 (multiple external

[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

association rule)을 이용하여 의사결정나무 생성에 불필요한 입력 변수를 제거할 수 있는 방법을 연구하고자 한다.

일반적으로 외적 변수는 목표 변수와 입력 변수 간에 내제적인 연결은 없고 어떤 외부 변수에 의하여 관계가 있는 것처럼 보일 경우, 이 외부 변수를 통제하면 두 변수의 관계가 사라지게 되며, 이때 통제되는 외부 변수를 외적 변수라고 한다. 외적 변수를 규명하는 방법에는 다양한 방법이 있으나, 본 논문에서는 연관성 규칙의 평가 기준을 바탕으로 외적 변수를 규명하므로 외적연관성규칙이라고 명하였고 목표 변수와 입력 변수와의 관계에서 외적 변수가 여러 개 존재 할 수 있으므로 이를 모두 밝혀낸다는 의미에서 다중외적연관성규칙이라고 명하였다. 다중외적연관성규칙에서는 목표 변수, 입력 변수, 외적 변수 사이에 연관성규칙을 살펴보고 외적 변수를 통제한 경우의 연관성 규칙을 살펴본 후 외적 변수의 조건이 만족하면 그 외적 변수를 의사결정나무 모형 생성에서 제거하게 된다.

본 논문에서 제시하는 방법은 하이브리드 (hybrid) 데이터마이닝이 방법이라고 할 수 있으며, 현재 모형 구축 시간 단축 및 생성된 모형 정확성 등의 데이터마이닝 효율성을 높이기 위하여 각각의 알고리즘을 혼합하여 사용하는 하이브리드 (hybrid) 데이터마이닝이 등장하였다. Cho와 Park (2006)은 연관성 규칙의 평가 기준을 바탕으로 매개 효과를 찾아내는 연구를 하였고, Lee와 Park (2007)은 군집분석을 연관성 규칙에 적용하였으며, Kim과 Park (2007)은 왜곡변수를 이용하여 연관성 규칙을 찾아내는 방법에 대하여 연구하였다. 또한 Kim과 Park (2008b)은 연관성 규칙에서 외적 변수를 찾는 연구를 실시한 바 있으며, 그 밖의 여러 연구자들에 의하여 하이브리드 데이터마이닝의 연구가 활발하게 진행되고 있다 (Park와 Cho, 2006a; Park와 Cho, 2006b; Kim과 Park, 2008a; Lee와 Park, 2008). 본 논문의 구성은 다음과 같다. 논문의 2절에서는 다중외적연관성규칙에 대하여 기술하고 3절에서는 실제 자료를 통하여 본 논문의 효과를 살펴본 후, 4절에서 결론을 맺고자 한다.

2. 다중외적연관성규칙

일반적으로 변수들 간의 내재적인 관련성은 없고 각 변수가 우연히 어떤 다른 변수와 연결됨으로써 관련성이 있는 것으로 나타나는 경우, 실제적으로 두 변수 간에는 관련성이 없으나 다른 외적인 변수에 의하여 관련성이 있는 것으로 나타나 두 변수간의 관련성을 분석한다면 잘못된 해석을 내릴 수 있다. 즉, 목표변수와 입력변수 간에 내제적인 연결은 없고 각 변수가 우연히 어떤 다른 변수와 연결됨으로써 관계가 있는 것처럼 보일 경우, 검정요인을 통제하면 두 변수의 관계가 사라지게 된다. 이때 통제되는 검정요인을 외적변수라고 한다. 예를 들어 화재장소에 소방차가 많이 나타날수록 화재피해액이 크다는 관찰의 결과가 있다고 하자. 이 경우 소방차의 수를 전항변수로 설정하고 화재피해액을 후항변수로 설정한다면 소방차의 수가 화재피해액의 원인으로 볼일 수가 있다. 그러나 실제로는 그 화재의 초기 발생시 규모가 소방차의 수와 화재피해액 모두의 원인이 되는 것이며, 소방차의 수와 화재피해액의 관계는 화재의 규모라고 하는 변수에 의해 우연히 연결된 관계에 불과하다. 그러므로 만약 화재의 규모를 통제한다면 소방차의 수와 화재피해액의 규모 사이에는 별다른 관계가 나타나지 않게 될 것이다. 이때 화재규모가 외적 변수가 되고 소방차의 수와 화재피해액의 관계의 해석은 잘못된 것이다. Kim과 Park (2008b)은 SAS 매크로를 이용하여 연관성 규칙에서 외적 변수를 찾는 연구를 실시한 바 있으며, 연관성 규칙에서 변수 X (전항 변수), Y (후항 변수), Z (외적 변수)가 있다고 가정했을 때, 외적 변수가 존재하기 위한 조건은 다음과 같다.

<조건 1> 변수 X 와 변수 Y 에 대한 연관성이 존재해야 한다.

<조건 2> 변수 Z 와 변수 X 에 대한 연관성이 존재해야 한다.

<조건 3> 변수 Z 와 변수 Y 에 대한 연관성이 존재해야 한다.

<조건 4> 변수 Z 를 통제했을 때, 변수 X 와 변수 Y 에 대한 연관성이 존재하지 않는다.

위의 4가지 조건이 만족하면 통제 변수가 외적변수가 되며, 전항 변수 와 후항 변수 에 대한 관련성은 외적 변수 에 의하여 우연히 연결된 관계에 불과하여 두 변수간의 관련성을 해석해서는 안 된다. Kim과 Park (2008b)은 SAS 매크로를 이용하여 연관성 규칙에서 외적 변수를 찾는 연구를 실시한 바 있으나, 이 방법은 전항변수와 후항변수 및 외적변수를 일일이 지정하여 결과를 해석하므로 연관성 규칙의 결과에서 외적 변수에 의한 의미 없는 규칙을 모두 찾는 다기 보다는 지정된 외적변수 하나만으로 전항변수와 후항변수와의 관계를 살펴봐야 한다는 제한점이 있었다. 이에 생성된 연관성규칙의 결과에 대하여 모든 통제변수를 이용하여 외적변수를 찾을 수 있을 뿐만 아니라 전항변수와 후항변수와의 관계에서 외적변수가 여러 개 존재 할 수 있으므로 이를 모두 밝혀낼 수 있는 방법 다중외적연관성규칙이라고 하며, 이를 그림으로 표현하면 그림 2.1과 같다.

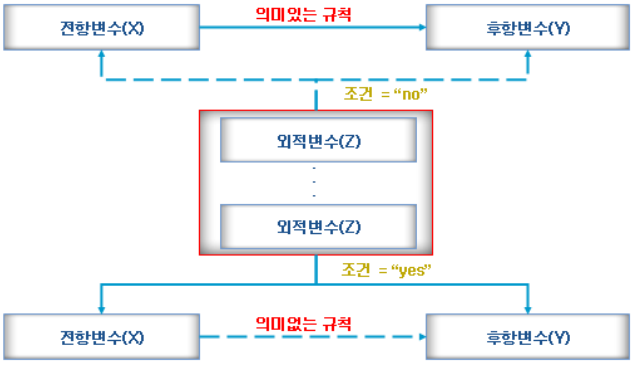

그림 2.1 다중외적연관성규칙

일반적으로 모형 생성의 기준 및 입력 변수의 수에 따라 의사결정나무 모형이 생성되므로 종종 복잡한 의사결정나무 모형이 생성되기도 한다. 특히 하나의 목표 변수에 여러 개의 입력 변수가 존재하는 경우 종종 모형 생성 및 해석에 있어 어려움을 격기도 한다. 다시 말해서, 목표 변수에 대한 입력 변수의 분리 기준에 따라서 의사결정나무 모형이 생성되므로 목표 변수에 유의한 입력 변수의 수가 많은 경우 의사결정나무 모형이 복잡해 질 수밖에 없다. 그러나 생성된 모형에 대한 목표 변수와 입력 변수와의 관계가 다른 외적 변수에 의하여 실제적으로 무의미한 관계라고 판단된다면 모형 생성 시 실제로 목표변수에 무의미한 입력 변수를 제거하고 모형을 생성하는 것이 효과적일 것이다. 이에 본 논문에서는 의사결정나무 생성 시 목표 변수와 입력 변수 사이의 외적 변수가 존재하는지 파악하기 위하여 다중외적연관성규칙을 적용하는 연구를 하고자 한다. 다중외적연관성규칙을 이용한 의사결정나무 생성의 적용 방안은 그림 2.2와 같다.

그림 2.2를 자세하게 설명하면 다음과 같다.

[단계 1] 변수 결정

의사결정나무 모형을 생성하기 위하여 목표 변수와 입력 변수를 결정한다.

[단계 2] 다중외적연관성규칙 생성

결정된 입력변수에 대하여 외적 변수가 존재하는 가를 파악하기 위하여 최소 지지도, 최소 신뢰도, 향상도를 결정하여 다음의 4가지 경우에 대한 연관성규칙을 생성한다.

[단계 2-1] 목표 변수와 입력 변수와의 연관성 규칙 생성

[단계 2-2] 외적 변수와 입력 변수와의 연관성 규칙 생성

그림 2.2 적용 단계

[단계 2-3] 외적 변수와 목표 변수와의 연관성 규칙 생성
[단계 2-4] 외적 변수를 통제한 후 목표 변수와 입력 변수와의 연관성 규칙 생성

[단계 3] 다중외적연관성규칙 성립 파악
목표 변수와 입력 변수들 간의 외적 변수를 찾아내기 위하여 [단계 2]에서 생성된 4가지 연관성규칙에 대한 다중외적연관성규칙의 성립 여부를 파악한다.

[단계 4] 모형 설정
다중외적연관성규칙 성립 여부를 파악한 뒤 외적관계가 성립하는 경우의 입력 변수를 제거하고 모형을 설정한다. 모형 설정에서는 자료 분할, 모형 알고리즘 선택, 정지 규칙 등을 지정한다.

[단계 5] 모형 생성
지정된 모형에 의하여 모형을 생성한다. 생성된 모형에 대한 예측정확도 및 모형평가 예측정확도를 살펴본 뒤 모형에 대한 해석을 실시한다.

3. 자료 분석을 통한 효율성 파악

본 절에서는 다중외적연관성규칙을 이용한 의사결정나무 모형의 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, 건강식품 섭취 유무, 운동 유무, 음주 유무, 흡연 유무, 결혼 만족도, 부부 친밀도, 나이, 직업, 학력의 10개 문항을 추출하였으며, 연관성 규칙에 적용하기 위하여 편의상 비율자료는 평균을 바탕으로 이분형으로 변환한 뒤 분석을 실시하였다. 분석에 사용한 변수는 표 3.1과 같으며, 목표변수인 갱년기 유무에 대해서는 “예”가 280명 (49%), “아니오”가 291명 (51%)으로 응답을 하였다.

본 논문에서는 기존의 의사결정나무 원 모형과 다중외적연관성규칙을 이용한 의사결정나무 모형의 두 가지 모형을 생성 한 뒤, 두 모형을 비교하고자 한다. 첫 번째로 갱년기 유무를 목표변수로 지정하고 건강식품 섭취 유무, 운동 유무, 음주 유무, 흡연 유무, 결혼 만족도, 부부 친밀도, 나이, 직업, 학력의 9개

표 3.1 변수 설명

변수명	구분	형태	설명
갱년기 유무	목표변수	이분형	범주 1 : 예, 범주 2 : 아니오
건강식품 섭취 유무	입력변수	이분형	범주 1 : 예, 범주 2 : 아니오
운동 유무	입력변수	이분형	범주 1 : 예, 범주 2 : 아니오
음주 유무	입력변수	이분형	범주 1 : 예, 범주 2 : 아니오
흡연 유무	입력변수	이분형	범주 1 : 예, 범주 2 : 아니오
결혼 만족도	입력변수	이분형	범주 1 : 낮음, 범주 2 : 높음
부부 친밀도	입력변수	이분형	범주 1 : 낮음, 범주 2 : 높음
나이	입력변수	이분형	범주 1 : 낮음, 범주 2 : 높음
직업	입력변수	이분형	범주 1 : 가정주부, 범주 2 : 기타
학력	입력변수	이분형	범주 1 : 낮음, 범주 2 : 높음

문항을 입력 변수를 지정하여 기존의 의사결정나무 모형을 생성한다. 본 논문에서는 입력 변수가 이분형 자료이므로 이지 분리가 가능한 CART 알고리즘으로 의사결정나무 모형을 생성하였으며, 자료의 형태가 연속형일 경우에는 C5.0 등의 알고리즘을 이용하여 의사결정나무 모형을 생성할 수도 있다. 모형 생성 시, 훈련 자료 (2/3)와 모형평가 자료 (1/3)로 분할하여 모형을 생성하였으며 생성된 모형은 그림 3.1과 같다.

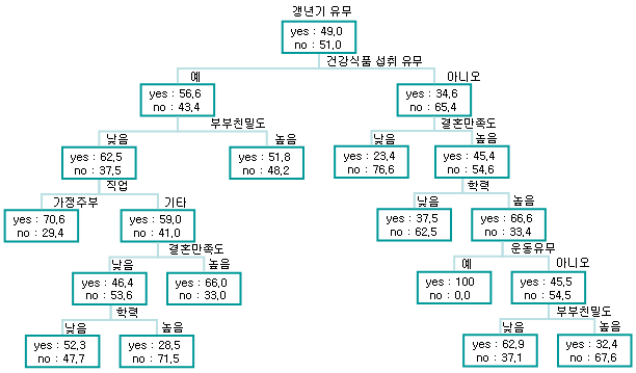

그림 3.1 의사결정나무 원 모형

다음으로 갱년기 유무를 목표변수로 지정하고 건강식품 섭취 유무, 운동 유무, 음주 유무, 흡연 유무, 결혼 만족도, 부부 친밀도, 나이, 직업, 학력의 9개 문항을 입력변수를 지정하였을 경우, 다중외적연관성규칙의 성립 여부를 파악한 후, 의사결정나무 모형을 생성한다. 입력 변수에 대한 다중외적연관성규칙을 적용한 결과, 다중외적연관성규칙이 성립하는 변수가 1개 존재하였고 결혼만족도가 부부친밀도와 갱년기 유무 사이에 외적 변수로 도출되었다. 다중외적연관성규칙의 결과는 표3.2와 같으며, 다중외적연관성규칙의 기준은 최소 지지도를 10, 최소 신뢰도를 70, 향상도를 1로 지정하였다 (표에서는 연관성 기준 중 신뢰도만 표시함).

표 3.2를 자세하게 살펴보면, 조건 1에서 전항변수와 후항변수와의 관련성이 존재하고, 조건 2에서 외적변수와 전항변수와의 관련성이 존재하며, 조건 3에서 외적변수와 전항변수와의 관련성이 존재한다. 또한 조건 4에서 외적변수를 통제했을 때 전항변수와 후항변수와 신뢰도값이 60.1로 최소 신뢰도의 기준값인 70보다 작으므로 관련성이 존재하지 않아 다중외적연관성규칙이 성립한다. 즉, “부부친밀도가 낮으면 갱년기 여성이 많다”라는 규칙은 결혼만족도란 외적변수에 의하여 우연히 나타난 규칙이므로 해

표 3.2 다중외적연관성규칙 결과

조건	전항 변수	후항 변수	외적 변수	신뢰도
1	부부친밀도	갱년기 유무	-	71.2
2	부부친밀도	-	결혼만족도	72.4
3	-	갱년기 유무	결혼만족도	72.5
4	부부친밀도	갱년기 유무	결혼만족도	60.1

석하는 것을 바람직하지 않다. 이에 본 논문에서는 부부친밀도가 결혼만족도 (외적 변수)에 의하여 의미가 없는 변수로 판단되었으므로 9개 문항의 입력 변수 중 부부친밀도를 제외한 8개 문항을 입력 변수로 지정하여 위의 원 모형과 동일한 조건으로 의사결정나무 모형을 생성하였다. 생성된 모형은 그림 3.2와 같다.

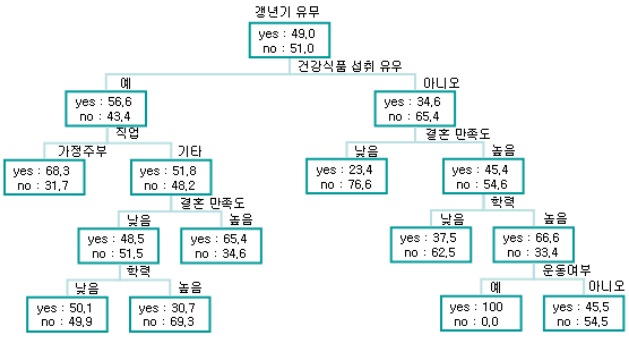

그림 3.2 다중외적연관성규칙을 이용한 의사결정나무 모형

갱년기 유무를 목표변수로 하는 의사결정나무의 원 모형인 그림 3.1과 본 논문에서 제시하는 다중외적연관성규칙을 이용한 의사결정나무 모형인 그림 3.2를 비교하면 표 3.3과 같다.

표 3.3 모형 비교

구분	원 모형 (그림 3.1)	제안 모형 (그림 3.2)
최대 노드의 깊이	5	4
노드의 수	19	15

표 3.3을 살펴보면 최대 노드의 깊이가 5개에서 4개로 줄어들었고 노드의 수 또한 19에서 15개로 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 3.4에서와 같이 기존의 나무모형과 본 논문에서 제시하는 나무 모형의 정확도를 비교하였다.

표 3.4 모형의 정확도 비교

모형	원 모형 (그림 3.1)		제안 모형 (그림 3.2)	
	모형 예측정확도	모형평가 예측정확도	모형 예측정확도	모형평가 예측정확도
퍼센트	71.9%	72.3%	69.5%	70.1%

표 3.4를 살펴보면 다중외적연관성규칙을 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 외적변수를 이용한 의사결정나무모형 생성의 방법이 모형의 정확도는 거의 동일하면서 불필요한 가지를 생성하지 않으므로 효율적이라고 할 수 있다.

4. 결론

데이터마이닝은 기업 등에서 보유하고 있는 사용 가능한 데이터를 기반으로 숨겨진 지식, 기대하지 못했던 패턴, 새로운 법칙과 관계를 발견하고 이를 의사결정 등을 위한 정보로 활용하고자 하는 기법으로서, 의사결정나무, 연관 규칙, 군집분석, 신경망 분석 등의 알고리즘이 있으며, 이중 의사결정나무 알고리즘은 의사결정 규칙을 도표화하여 관심대상이 되는 집단을 몇 개의 소집단으로 분류하거나 예측을 수행하는 방법이다. 일반적으로 의사결정나무의 모형 생성 시, 입력 변수의 수가 많을 경우 생성된 의사결정모형은 복잡한 형태가 될 수 있고, 모형 탐색 및 분석에 있어 어려움을 격기도 한다. 이때 생성된 모형에 대한 목표 변수와 입력 변수와의 관계에서 두 변수의 관계가 우연히 어떤 외적 변수에 의하여 실제적으로 무의미한 관계라고 한다면 모형 생성 시 그 입력 변수를 제거하고 모형을 생성하는 것이 효과적이다. 이에 본 논문에서는 의사결정나무 생성 시, 목표 변수와 입력 변수에 대한 관계를 명확하게 파악할 수 있는 다중외적연관성규칙을 적용하여 불필요한 입력 변수를 제거할 수 있는 방법을 제안하였고, 실제 자료에 적용해 보았다.

분석 결과, 목표 변수와 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 국가 통계, 기업체 및 연구 자료 등에 다양하게 적용하여 변수들 간의 관계를 명확하게 규명할 필요성이 있으며, 다중외적연관성규칙을 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 Brooks, California.

Cho, K. H. and Park, H. C. (2006). A study for intervening effect verification using association rules. *Journal of the Korean Data Analysis Society*, **8**, 1905-1914.

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

Kim, M. H. and Park, H. C. (2007). A study for discovery of distorter variable using association rules. *Journal of the Korean Data Analysis Society*, **9**, 711-719.

Kim, M. H. and Park, H. C. (2008a). Development of component a ssociation rule s and macro algorithm. *Journal of the Korean Data & Information Science Society*, **19**, 197-207.

Kim, M. H. and Park, H. C. (2008b). Development of extraneous association rules and SAS macro algorithm. *Journal of the Korean Data Analysis Society*, **10**, 1141-1152.

Lee, K. W. and Park, H. C. (2007). A study of k-means clustering for association rule. *Journal of the Korean Data Analysis Society*, **9**, 2919-2930.

Lee, K. W. and Park, H. C. (2008). A study for statistical criterion in negative association rules using boolean analyzer. *Journal of the Korean Data & Information Science Society*, **19**, 569-576.

Park, H. C. and Cho, K. H. (2006a). Discovery of association rules using latent variables. *Journal of the Korean Data & Information Science Society*, **17**, 149-160.

Park, H. C. and Cho, K. H. (2006b). A study for antecedent association rules. *Journal of the Korean Data & Information Science Society*, **17**, 1077-1083.

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

Journal of the Korean Data & Information Science Society
2011, **22**(5), 877–884

한국데이터정보과학회지

A study on removal of unnecessary input variables using multiple external association rule

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 18 July 2011, revised 11 August 2011, accepted 21 August 2011

Abstract

The decision tree is a representative algorithm of data mining and used in many domains such as retail target marketing, fraud detection, data reduction, variable screening, category merging, etc. This method is most useful in classification problems, and to make predictions for a target group after dividing it into several small groups. When we create a model of decision tree with a large number of input variables, we suffer difficulties in exploration and analysis of the model because of complex trees. And we can often find some association exist between input variables by external variables despite of no intrinsic association. In this paper, we study on the removal method of unnecessary input variables using multiple external association rules. And then we apply the removal method to actual data for its efficiencies.

Keywords: Association rule, confidence, data mining, decision tree, external variable.

[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