Journal of the Korean Data & Information Science Society
2011, **22**(4), 671–678

한국데이터정보과학회지

매개 변수를 이용한 의사결정나무 생성에 관한 연구

조광현[1] · 박희창[2]

[1]창원대학교 유아교육학과 · [2]창원대학교 통계학과

접수 2011년 5월 19일, 수정 2011년 6월 17일, 게재확정 2011년 6월 22일

요 약

데이터마이닝은 방대한 양의 데이터 속에서 쉽게 드러나지 않는 유용한 정보를 찾아내는 기법으로서 의사결정나무, 연관 규칙, 군집분석, 신경망 분석 등의 기법이 있으며, 이중 의사결정나무 알고리즘은 의사결정 규칙을 도표화하여 관심대상이 되는 집단을 몇 개의 소집단으로 분류하거나 예측을 수행하는 방법으로서 고객세분화, 고객 분류, 문제 예측 등의 여러 분야에서 유용하게 활용되고 있다.일반적으로 의사결정나무의 모형 생성 시, 모형 생성의 기준 및 입력 변수의 수에 따라 복잡한 모형이 생성되기도 하며 특히 입력 변수의 수가 많을 경우 종종 모형 생성 및 해석에 있어 어려움을 격기도 한다. 이에 본 논문에서는 의사결정나무 생성 시, 입력 변수에 대한 매개 관계를 파악하여 나무 생성에 불필요한 입력 변수를 제거하는 방법을 제시하고 그 효율성을 파악하기 위하여 실제 자료에 적용하고자 한다.

주요용어: 다중매개연관성규칙, 데이터마이닝, 매개 변수, 연관성규칙, 의사결정나무.

1. 서론

데이터마이닝이란 방대한 양의 데이터 속에서 쉽게 드러나지 않는 유용한 정보를 찾아내는 과정으로, 대용량의 관측 가능한 데이터를 기반으로 숨겨진 지식, 기대하지 못했던 패턴, 새로운 법칙과 관계를 발견하고 이를 바탕으로 의사결정 등을 위한 정보로 활용하고자 하는 것이다.

데이터마이닝 기법으로는 군집분석, 연관성규칙, 의사결정나무, 신경망모형 등의 분석 기법이 있으며, 현재 모형 구축 시간 단축 및 생성된 모형 정확성 등의 데이터마이닝 효율성을 높이기 위하여 각각의 알고리즘을 혼합하여 사용하는 하이브리드 (hybrid) 데이터마이닝의 연구가 활발하게 진행되고 있다 (Lee 등, 2010; Choi와 Kang, 2011). 본 논문에서 적용하고자 하는 의사결정나무는 의사결정 규칙을 도표화하여 관심대상이 되는 집단을 몇 개의 소집단으로 분류하거나 예측을 수행하는 분석방법으로 다른 분석방법에 비해 연구자가 분석과정을 쉽게 이해하고 설명할 수 있다는 장점이 있다. 그 동안의 연구를 살펴보면 의사결정나무분석을 수행하기 위한 다양한 분리기준, 정지규칙, 가지치기 방법들이 제안되었으며, 이들을 어떻게 결합하느냐에 따라서 서로 다른 의사결정나무가 형성된다. 또한 정확하고 빠르게 의사결정나무를 형성하기 위해 다양한 알고리즘이 제안되고 있다. 대표적인 의사결정나무 알고리즘에는 Hartigan (1975)에 의하여 제시된 CHAID (Chi-squared Automatic Interaction detecion), Breiman 등 (1984)에 의하여 제시된 CART (Classification and Regression Trees), Quinlan (1993)의 ID3을 기반으로한 C5.0 등의 알고리즘 있으며, CHAID는 의사결정나무의 가장 오래된 알고리즘으로 분리기준

[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

으로 카이제곱통계량을 사용하고 CART는 분리기준으로 지니 (Gini) 지수를 사용하여 이지 분리를 수행하는 알고리즘이며 C5.0은 분리기준으로 엔트로피 (entropy)를 사용하여 다지 분리를 수행하는 알고리즘이다. 의사결정나무의 모형 생성 시, 모형 생성의 기준 및 입력 변수의 수에 따라 복잡한 모형이 생성되기도 하며 특히 입력 변수의 수가 많을 경우 종종 모형 생성 및 해석에 있어 어려움을 겪기도 한다. 즉, 목표 변수와 입력 변수의 분리 기준에 따라 나무 모형이 생성되므로 입력 변수가 많은 경우 모형이 복잡해 질 수밖에 없다. 이때 생성된 모형에 대한 목표 변수와 입력 변수와의 관계가 다른 외부 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 매개 관계를 명확하게 파악할 수 있는 다중매개연관성규칙 (multi intervening association rule)을 적용하여 불필요한 입력 변수를 제거할 수 있는 방법을 연구하고자 한다.

2. 이론적 배경

연관성규칙은 Agrawal 등 (1993)에 의해 처음 소개된 이후 연관성규칙의 효율성을 개선하기 위하여 여러 가지 제약 기반 연관성규칙의 연구가 활발하게 진행되고 있고, Park과 Cho (2006a, 2006b), Kim과 Park (2008), Lee와 Park (2008) 등에 의하여 제약 기반 연관성규칙의 연구가 진행된 바 있으며, Cho와 Park (2011)은 다중매개연관성규칙을 제안하였다. 다중매개연관성규칙의 적용 단계는 그림 2.1과 같다.

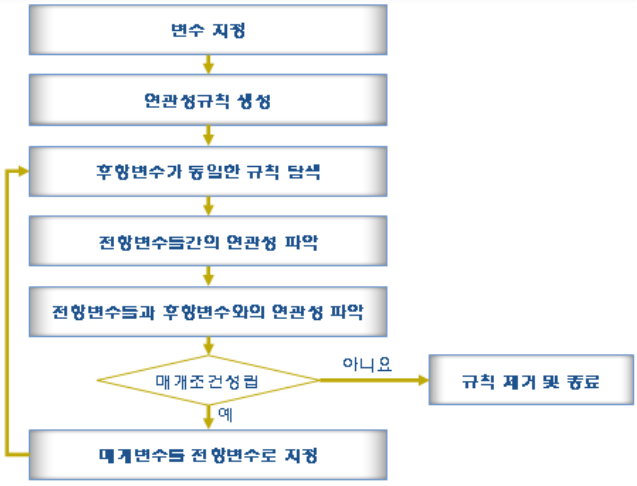

그림 2.1 다중매개연관성규칙의 적용 단계

그림 2.1의 다중매개연관성규칙에서의 매개 변수가 성립하기 위한 조건은 그림 2.2와 같다.

조건 1에서는 후항변수와 전항변수에 대한 관련성이 있어야하고, 조건 2에서는 전항변수와 매개 변수에 대한 관련성이 있어야 하고, 조건 3에서는 전항변수 및 매개 변수와 후항변수에 대한 관련성이 있어야 하며, 조건 4에서는 조건 1에서의 관련성 보다 조건 3에서의 관련성이 더 클 경우 전항변수와 후항변수 사이에 매개 변수가 존재한다고 할 수 있으며, 이 매개 변수에 의하여 전항변수와 후항변수와의 관련성은 의미가 없다고 판단한다. 본 논문에서 제안하는 다중매개연관성규칙을 이용한 의사결정나무 적용

그림 2.2 매개 변수 성립 조건

방안은 그림 2.3과 같다. 여기서, 앞에서 설명한 전항변수는 입력 변수를 의미하며 후항변수는 목표 변수를 의미한다.

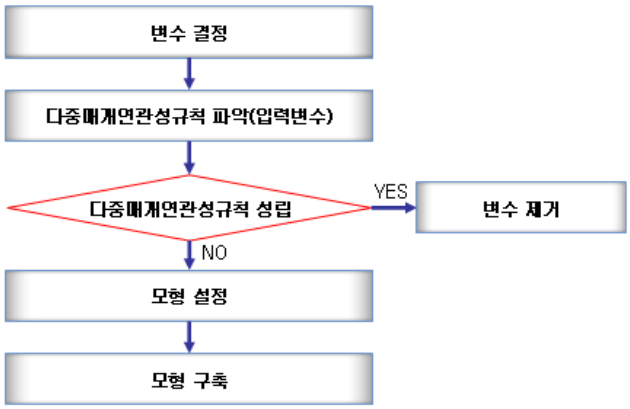

그림 2.3 의사결정나무 적용 방안 (다중매개연관성규칙)

[단계 1] 변수 결정 : 의사결정나무 모형을 생성하기 위하여 목표 변수와 입력 변수를 결정한다.

[단계 2] 다중매개연관성규칙 파악 : 결정된 입력 변수에 대하여 매개 변수가 존재하는 가를 파악하기 위하여 최소 지지도, 최소 신뢰도, 향상도를 결정하여 다중매개연관성규칙을 적용한다.

[단계 3] 다중매개연관성규칙 성립 : 목표 변수와 입력 변수들 간의 매개 관계를 파악하기위하여 그림 2.2의 매개 변수 성립 조건 4가지를 파악한다.

[단계 4] 모형 설정 : 다중매개연관성규칙 성립 여부를 파악한 뒤 매개관계가 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 분석

본 장에서는 다중매개연관성규칙을 이용한 의사결정나무 모형의 효용성을 파악하기 위하여 통계청의 통계정보시스템인 KOSIS (www.kosis.kr) 자료를 이용하였다. 자료는 2009년 조사된 시, 군, 구의 인구수, 가구수, 인구증가율, 취업률 등의 총 7개 문항을 추출하였고, 원 모형과 본 논문에서 제시하는 모형의 효율성을 파악하기 위함이므로 간단하게 모든 자료를 평균을 바탕으로 이분형으로 변환한 뒤 분석을 실시하였으며, 표 3.1과 같다.

표 3.1 변수 설명

변수	구분	형태	설명
인구수	입력 변수	이분형	범주 1 : 많음, 범주 2 : 적음
세대수	입력 변수	이분형	범주 1 : 많음, 범주 2 : 적음
인구 증가율	입력 변수	이분형	범주 1 : 높음, 범주 2 : 낮음
노령 인구율	입력 변수	이분형	범주 1 : 높음, 범주 2 : 낮음
자동차 등록률	입력 변수	이분형	범주 1 : 높음, 범주 2 : 낮음
취업률	입력 변수	이분형	범주 1 : 높음, 범주 2 : 낮음
교통사고율	목표 변수	이분형	범주 1 : 높음, 범주 2 : 낮음
제조업체수	목표 변수	이분형	범주 1 : 많음, 범주 2 : 적음

본 논문에서는 기존의 의사결정나무 원 모형과 다중매개연관성규칙을 이용한 의사결정나무 모형의 두 가지 모형을 생성 한 뒤, 두 모형을 비교하고자 한다.

첫 번째로 교통사고율을 목표 변수로 지정하고 인구수, 세대수, 인구증가율, 노령인구율, 자동차 등록률, 취업률의 6개 문항을 입력 변수를 지정하여 기존의 의사결정나무 모형을 생성한다. 모형 생성에서는 비교적 모형이 간단하게 생성되는 CART 모형을 선택하였으며, 훈련 자료와 모형 평가 자료로 분할하여 모형을 생성하였다. 생성된 모형은 그림 3.1과 같다.

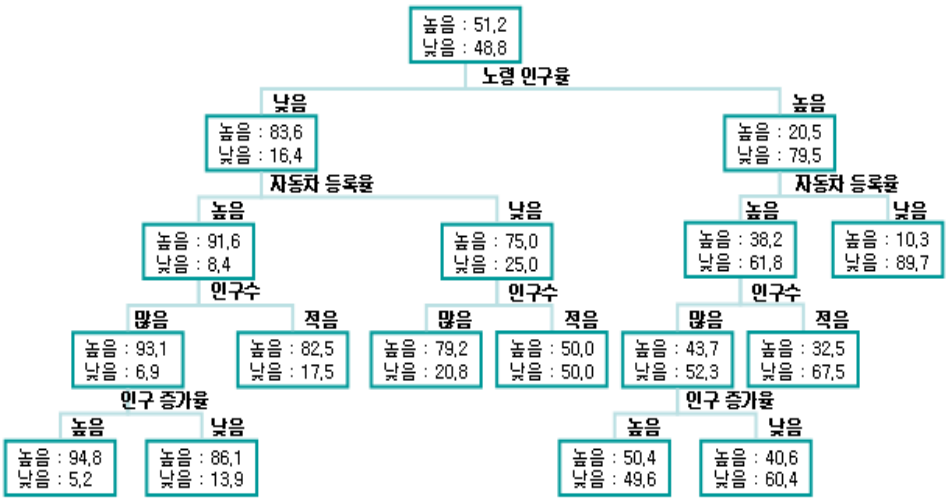

그림 3.1 교통사고율 나무모형 (원 모형)

다음으로 교통사고율을 목표 변수로 지정하고 인구수, 세대수, 인구증가율, 노령인구율, 자동차 등록

률, 취업률의 6개 문항을 입력 변수를 지정하였을 경우, 다중매개연관성규칙의 성립 여부를 파악한 후, 의사결정나무 모형을 생성한다. 입력 변수에 대한 다중매개연관성규칙의 결과는 표 3.2와 같다 (최소 지지도 : 10, 최소 신뢰도 : 70, 향상도 : 1, 연관성 기준 중 신뢰도만 표시함).

표 3.2 다중매개연관성규칙 결과

조건	목표 변수	입력 변수	매개 변수	신뢰도
1	교통사고율	인구수	-	70.4
2	교통사고율	-	자동차 등록률	71.8
3	교통사고율	인구수	자동차 등록률	74.6

표 3.2의 다중매개연관성규칙의 결과를 살펴보면, 목표 변수인 교통사고율과 입력 변수인 인구수 사이에 매개 변수를 자동차 등록률로 지정하였을 경우, 매개 변수의 조건 4가지를 모두 만족하고 있으므로 입력 변수 중 인구수가 자동차 등록률 (매개 변수)에 의하여 의미가 없는 변수로 판단되었으므로 6문항의 입력 변수 중 인구수를 제외한 5문항을 입력 변수로 지정하여 위의 원 모형과 동일한 조건으로 의사결정나무 모형을 생성하였다. 생성된 모형은 그림 3.2와 같다.

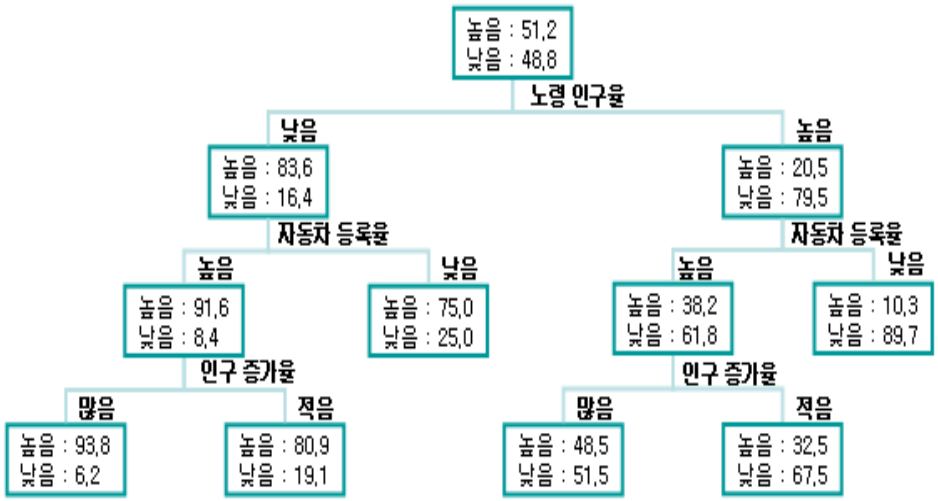

그림 3.2 교통사고율 나무모형 (다중매개연관성규칙을 이용한 모형)

교통사고율의 원래 나무모형과 다중매개연관성규칙을 이용한 나무모형의 차이를 나타내면 표 3.3과 같다.

표 3.3 모형 비교

구분	원 모형	다중매개연관성규칙을 이용한 모형
최대 노드의 깊이	4	3
노드의 수	9	6

표 3.3을 살펴보면 최대 노드의 깊이가 4개에서 3개로 줄어들었고 최종 노드의 수 또한 9에서 6개로 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 그림 3.1의 원 모형과 그림 3.2의 다중매개연관성규칙을 이용한 모형의 정확도를 비교하였다.

표 3.4 모형의 정확도 비교

	모형			
	원 모형		다중매개연관성규칙을 이용한 모형	
목표 변수	모형 예측정확도	모형평가 예측정확도	모형 예측정확도	모형평가 예측정확도
교통사고율	73.8%	73.1%	71.4%	71.2%

표 3.4를 살펴보면, 다중매개연관성규칙을 이용한 모형의 모형 예측정확도 및 모형평가 예측정확도가 원 모형의 모형 예측정확도 및 모형평가 예측정확도와 큰 차이를 보이고 있지 않은 것을 알 수 있다. 이에 본 논문에서 제시하는 다중매개연관성규칙을 이용한 의사결정나무모형 생성의 방법이 모형의 정확도는 거의 동일하면서 불필요한 가지를 생성하지 않으므로 효율적이라고 할 수 있다.

4. 결론

데이터마이닝은 방대한 양의 데이터 속에서 쉽게 드러나지 않는 유용한 정보를 찾아내는 기법으로서 의사결정나무, 연관 규칙, 군집분석, 신경망 분석 등의 기법이 있으며, 이중 의사결정나무 알고리즘은 의사결정 규칙을 도표화하여 관심대상이 되는 집단을 몇 개의 소집단으로 분류하거나 예측을 수행하는 방법으로서 고객세분화, 고객 분류, 문제 예측 등의 여러 분야에서 유용하게 활용되고 있다.

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 매개 변수에 의하여 실제적으로 무의미한 관계라고 판단된다면 모형 생성 시 실제로 목표 변수에 무의미한 입력 변수를 제거하고 모형을 생성하는 것이 효과적일 것이다. 여기서 변수들 간의 직접적인 관련성은 없고 각 변수가 매개 변수에 의하여 간접적인 관련성이 있는 것으로 나타나는 경우, 두 변수간의 관련성을 분석한다면 의미 없는 해석을 내릴 수 있다.

이에 본 논문에서는 의사결정나무 생성 시, 목표 변수와 입력 변수에 대한 관계를 명확하게 파악할 수 있는 다중매개연관성규칙을 적용하여 불필요한 입력 변수를 제거할 수 있는 방법을 제안하였고, 실제 자료에 적용해 보았다. 분석 결과, 본 논문에서 제시하는 모형의 모형 예측정확도 및 모형평가 예측정확도가 원 모형의 모형 예측정확도 및 모형평가 예측 정확도와 큰 차이를 보이고 있지 않으면서 목표 변수와 입력 변수 사이에 무의미한 입력 변수를 제거함으로서 의사결정나무 모형의 생성 및 해석의 시간과 노력을 단축할 수 있으므로 본 논문에서 제시하는 방법이 효율적이라고 할 수 있다.

향후 과제로 본 논문에서 제안하는 방법을 국가 통계, 기업체 및 연구 자료 등의 조금 더 실제적인 자료에 적용하여 생성된 모형을 분석 할 필요성이 있다.

참고문헌

Agrawal, R., Imielinski, R. and Swami, A. (1993). Mining association rules between sets of items in large databases. *Proceedings of the ACM SIGMOD Conference on Management of Data,* 207-216.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and regression trees,* Chapman & Hall/CRC, New York.

Cho, K. H. and Park, H. C. (2011). Study on the multi intervening relation in association rule. *Journal of the Korean Data Analysis Society*, **13**, 297-306.

Choi, S. B. and Kang, C. W. (2011). Analysis of department homepage using web mining technique. *Journal of the Korean Data Analysis Society*, **13**, 317-330.

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

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

Lee, Y. S., Kim, K. K. and Kang, C. W. (2010). Development of customer lifetime value model based on TRFM for customer segmentation. *Journal of the Korean Data Analysis Society*, **12**, 3271-3282.

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**(4), 671–678

한국데이터정보과학회지

A study on decision tree creation using intervening variable

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 19 May 2011, revised 17 June 2011, accepted 22 June 2011

Abstract

Data mining searches for interesting relationships among items in a given database. The methods of data mining are decision tree, association rules, clustering, neural network and so on. The decision tree approach is most useful in classification problems and to divide the search space into rectangular regions. Decision tree algorithms are used extensively for data mining in many domains such as retail target marketing, customer classification, etc. When create decision tree model, complicated model by standard of model creation and number of input variable is produced. Specially, there is difficulty in model creation and analysis in case of there are a lot of numbers of input variable. In this study, we study on decision tree using intervening variable. We apply to actuality data to suggest method that remove unnecessary input variable for created model and search the efficiency.

Keywords: Association rule, data mining, decision tree, intervening variable, multi intervening association rule.

[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