한국정보전자통신기술학회논문지(jkiiect)'18-04, Vol.11 No.2
논문 18-11-02-169 http://dx.doi.org/10.17661/jkiiect.2018.11.2.169

베이지안 네트워크를 이용한 다차원 범주형 분석

김용철*

Multi-dimension Categorical Data with Bayesian Network

Yong-Chul Kim*

요 약 일반적으로 자료의 효과 연속형인 경우 분산분석과 이산형인 경우 분할표 카이제곱 검정을 통계적 분석방법으로 사용한다. 다차원의 자료에서는 계층적 구조의 분석이 요구되어지며 자료간의 인과관계를 나타내기 위해 통계적 선형모형을 채택하여 분석한다. 선형모형의 구조에서는 자료의 정규성이 요구되어지며 일부 자료에서는 비 선형모형을 채택할 수도 있다. 특히, 설문조사 자료 구조는 문항의 특성상 이산형 자료의 형태가 많아 모형의 조건에 만족하지 않는 경우가 종종 발생한다. 자료구조의 차원이 높아질수록 인과관계, 교호작용, 연관성분석 등에 다차원 범주형 자료 분석 방법을 사용한다. 본 논문에서는 확률분포의 계산을 이용한 베이지안 네트워크 모형이 범주형 자료 분석에서 분석절차를 줄이고 교호작용 및 인과관계를 분석할 수 있다는 것을 제시하였다.

Abstracts In general, the methods of the analysis of variance(ANOVA) for the continuous data and the chi-square test for the discrete data are used for statistical analysis of the effect and the association. In multidimensional data, analysis of hierarchical structure is required and statistical linear model is adopted. The structure of the linear model requires the normality of the data. A multidimensional categorical data analysis methods are used for causal relations, interactions, and correlation analysis. In this paper, Bayesian network model using probability distribution is proposed to reduce analysis procedure and analyze interactions and causal relationships in categorical data analysis.

Key Wards : Bayesian network, categorical data, conditional probability, odds ratio, survey analysis

1. 서론

모집단의 자료구조를 파악하기 위해 널리 사용되어지는 자료 수집 방법은 설문조사이며 목적에 따라서 자료의 형태에 따라서 분석방법을 적정하게 선택하여 분석한다. 일반적으로 설문조사 후 정리된 자료를 평균분석, 상관분석, 분산분석 등으로 설문조사의 목적을 분석할 수 있으며 이러한 설문지 분석은 보편화되어 있다.

특히, 이차원 범주형 자료의 경우 카이제곱 검정으로 동질성과 독립성검정을 함으로서 자료의 연관성을 분석한다. 2×2 분할 표에서 연관성을 분석하는 도구로 오즈비를 이용한다. 두 개의 확률 변수 X, Y의 연관성을 분석하기 위하여 주변확률 X의 관심 대상의 확률을 $\pi = p(X=1)$라 하면 오즈비는 $\pi/1-\pi$로 정의되어지며 상대적으로 관심대상이 되는 기대확률을 분석한다. 그러나 다차원구조의 확률 분할표의 작성은 각각의 변수의 개수에 따서 차원이 결정되며 복잡한 구조를 가지고 있으며 계산상의 어려움을 초래한다.

다차원 결합분포 함수는 조건부 확률을 이용하며 결합 분포 함수는 세분화하여 낮은 차원의 조건부 확률의 곱으로 표현이 가능하다. 확률변수 Z가 주어진 조건하에서 두 확률변수 X, Y가 독립이라면 결합 분포함수는 식 1과 같다.

$$\begin{aligned} p(X,Y,X) &= p(Z)p(X|Z)p(Y|X,Z) \\ &= p(Z)p(X|Z)p(Y|Z). \end{aligned} \quad \cdots\cdots (1)$$

*Corresponding Author : Department of Logistic and Statistical Information, Yongin University(yckim@yongin.ac.kr)
Received March 15, 2018 Received March 29, 2018 Accepted March 29, 2018

즉, 식 1에서 $p(Y|X,Z)=p(Y|Z)$이다. 조건부 독립을 이용하면 다차원의 결합분포 함수의 계산식을 단순화 할 수 있다.

베이지안 네트워크 모형은 변수들 간의 인과관계를 나타낼 수 있다[1,2,3,5]. 황성철, 이일병[6]등은 MDL Principle을 적용한 점수 기반 베이지안 네트워크 학습 방법에 대하여 논의하였으며, 정성원, 이도헌, 이광형[7]등은 베이지안 네트워크에 존재할 수 있는 불확실성을 언급한 후, 베이지안 네트워크 내의 변수들이 갖는 확률분포의 분산을 이용해 베이지안 네트워크의 불확실성을 정의하는 방법을 제안하였다. 또한, 임성수, 조성배[8]는 스크립트로부터 베이지안 네트워크를 자동으로 생성하여 베이지안 네트워크를 이용한 대화형 에이전트의 확장성을 높이는 방법을 제안하여 제한된 조건하에서 주변 확률을 예측하였다.

설문조사 자료 분석에서 특정 문항들의 효과를 분석하기 위하여 분산분석 및 분할표 검정으로 일차적 분석을 반복적으로 하며 선형 모형의 분석에서는 자료의 정규성이 대체적으로 요구되어지며 모형 가정의 위배는 분석 결과에 심각한 오류를 발생시킨다[4,9,10]. 베이지안 네트워크는 확률을 가지고 분석을 하기 때문에 이러한 점을 보완할 수 있다. 설문조사 자료 분석은 이산형의 자료를 분석하는 경우가 많이 발생하며 특히 다차원 범주형 자료 분석은 베이지안 네트워크 모형을 적용하여 분석하면 인과관계에 관련하여 교호작용을 파악하거나 연관 분석이 가능하다.

본 논문에서는 설문지 분석에서 설문문항들 간의 계층적 모형을 설계하여 인과관계분석 및 효과분석을 할 수 있는 베이지안 네트워크 모형을 제시하고 효과성에 대하여 논의하였다.

2. 사전적 이론

2.1 2×2 분할표의 연관성 분석 방법

확률 변수 X, Y에 대한 분할 표에서 결합 활률을 π_{ij}라하면 주변확률 분포는 $\pi_{i+}=\sum_{j=1}^{2}\pi_{ij}$ 이고 $\pi_{+j}=\sum_{i=1}^{2}\pi_{ij}$ 이다. $X=i$로 주어졌을 때 $Y=j$일 조건부 확률은 $\pi_{j|i}$로 표현 할 수 있다. 두 확률이 모든 i,j에 대하여 $\pi_{j|i}=\pi_{ij}/\pi_{i+}$ 이면 확률적 독립이라 한다. 그리고 두 확률의 차이가 모든 $i=1,2$에 대하여 $\pi_{i|1}=\pi_{i|2}$이면 동질하다고 한다. 상대적 위험도는 $\pi_{1|1}/\pi_{1/|2}$이고 1이면 독립을 의미한다.

오즈비는 $\theta=\frac{\pi_{11}\times\pi_{22}}{\pi_{12}\times\pi_{21}}$로 표현하고 $\theta=1$이면 독립이고 $\theta>1$이면 상대적으로 첫 행이 두 번째 행 보다 자주 발생한다고 해석한다. 2×2 분할 표에서는 위의 식 결합분포, 주변분포, 조건부 분포, 오즈비등을 사용하여 연관성을 분석한다. 확률 변수의 수가 많아지면 다 차원의 분할 표에 대한 분석은 결합 확률분포의 복잡성 때문에 로그 선형모형으로 분석하기도 한다.

2.2 베이지안 네트워크

베이지안 네트워크는 변수들의 계층적 구조의 관계성을 이용하여 확률 변수들의 결합 분포를 조건부 확률 분포를 이용하여 함수식의 구조를 단순하게 바꾸는 분석 방법이다.

베이지안 네트워크의 구조는 방향성이 있는 비 순환 구조의 그래프 형태로 이루어져 있어야하며 확률 변수 $X_1,X_2,...,X_n$의 확률분포는 식 2와 같다.

$$\begin{aligned}&p(X_1,X_2,X_3,...,X_n)\\&=p(X_1)p(X_2|X_1)p(X_3|X_1,X_2)\cdots p(X_n|X_1,X_2,X_3,...,X_{n-1}).\end{aligned}\quad(2)$$

식 2에서 주어진 조건 X_i에 따라서 $X_1,X_2,...,X_{i-1},X_{i+1},...,X_n$의 조건부 독립인 경우에 식 3과 같다.

$$p(X_1,X_2,X_3,...,X_n)=\Pi p(X_j|X_{i\neq j}).\quad(3)$$

1) 순차적 그래프의 형태는 그림 1과 같고 전체 확률을 구하면 $p(x,y,z)=p(x)p(y|x)p(z|y)$이다.

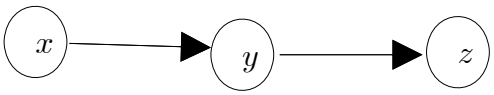

그림 1. 순차적 그래프
Fig. 1. Sequence Graph

그림 1에서 알 수 있듯이 y의 값이 중간에서 노드 역할을 함으로써 y의 값이 주어진 상태에서 x와 z가 조건부 독립이 된다.

2) 계층적 그래프에서의 전체 확률은
$p(x,y,z) = p(y)p(x|y)p(z|y)$이다.

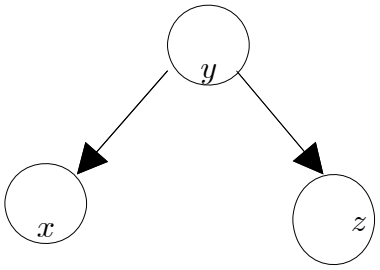

그림 2. 계층적 그래프
Fig. 2. Hierarchical Graph

그림 2는 그림 1과 같이 y의 값이 중간 노드 역할을 하고 있으나 x와 z에 모두 영향을 주고 있으며 y값이 주어진 상태에서 x와 z가 조건부 독립이다.

3) 베이지안 네트워크 모형의 유용성은 결합 확률밀도 함수에 대하여 주어진 상위 K 노드에 대해 하위 노드들의 조건부 독립이라는 조건하에서 분할표의 셀의 개수를 $O(2^N)$에서 $O(N\times 2^K)$로 줄 일수 있다. 또한 선형모형에서의 근사적 분석을 확률적인 분석으로 가능하게 한다. 그러므로 순차적 및 계층적 그래프를 혼합하여 분석의 목적에 맞도록 모형을 설계를 하면 된다.

2.3 설문조사에서의 베이지안 네트워크

설문조사의 구조적 형태를 보면 인구학적인 설문문항과 설문 조사목적이 담긴 문항으로 되어 있다. 일반적으로 설문분석은 기본적으로 빈도분석, 평균분석, 분산분석, 경로분석 등으로 조사의 목적에 따라 분석한다. 특히 다변량 분석에서 서로의 연관성에 대해서는 상관계수와 부분상관계수로 분석이 가능하지만 변수의 수가 많아지면 인과관계의 경로를 분석하기가 쉽지 않다.

본 논문에서는 설문조사의 다변량 분석에서 다음과 같은 비 순환형 베이지안 네트워크 모형을 적용하여 조사 목적에 합당한 결과를 유추할 수 있는 모형을 제시하고자 한다.

첫 번째 층에는 효과원인에 관련된 문항 $X_1, X_2, \cdots, X_M$을 배열하고, 두 번째 층에는 효과결과에 관련된 문항 Y_1, Y_2, Y_3, 세 번째 층에는 인구학적 문항 $Z_1, Z_2, \cdots, Z_N$을 배열하여 필요한 주변 확률 과 결합 확률, 조건부 확률의 각각의 분포를 계산하여 분석하는 모형은 다음과 같이 제시 하고자 한다.

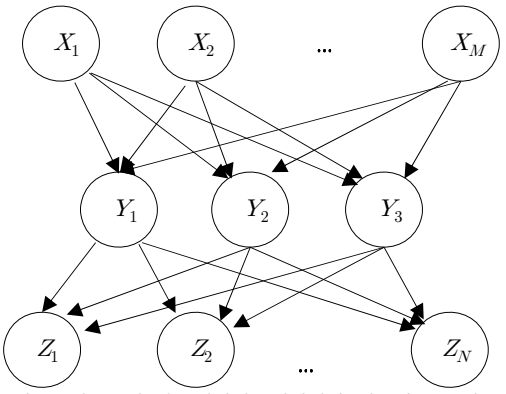

그림 3. 설문조사 자료에서의 베이지안 네트워크 모형
Fig. 3. Bayesian Network Model for the Data of Survey

설문조사에서 분석하고자 하는 문항에 적절한 베이지안 네트워크 모형을 선택한다면 다차원 범주형 자료분석과 대등하게 제한된 조건 없이 연관성 분석이 가능하다.

3. 베이지안 네트워크 모형 분석 결과

일반적으로 설문조사가 완료가 되면 일차원적인 분석으로 분할표 검정을 사용한다. 특히, 2×2 분할표 검정은 독립성 또는 동질성에 대하여 변수들의 연관성을 알아본다. 본 논문에서는 사례 분석 자료로 사회 건강관련 설문조사 6292명을 대상으로 하는 설문문항에서 베이지안 네트워크 모형에 적합하기 위하여 상위 층

의 원인변수로 식습관 관련 변수 X_1, X_2, X_3와 결과 분석을 목적으로 하는 변수인 고위험 변수 Y와 인구학적 변수로 나이와 성별 변수인 Z_1, Z_2로 하였다.

3.1 2×2 분할표 검정

자료는 Y와 X_1, X_2, X_3, Z_1, Z_2등의 변수들을 2×2 분할표 χ^2 독립성 검정 방법을 이용하여 분석하였다. 분석 결과는 다음과 같다.

표 1. $Y\times X_1$, $Y\times X_2$, $Y\times X_3$ 카이제곱 검정
Table 1. Chi-Square Test for $Y\times X_1$, $Y\times X_2$, $Y\times X_3$

	var-iable	Y		test statistic	p-value
var-iable	level	Yes	No		
X_1	Yes	42	4840	6.3012	0.0121
	No	23	1391		
X_2	Yes	18	2592	5.1257	0.0236
	No	47	3639		
X_3	Yes	1	52	0.3819	0.5366
	No	64	6179		

위의 표 1은 상위계층에 있는 변수들과 관심이 있는 변수와 독립성 검정 결과이며 Y와X_3은 검정통계량 0.3819이고 p-값은 0.5366으로 유의 수준 $\alpha=0.05$에서 서로 관련성이 없으며 Y와X_1 그리고 Y와X_2 는 관련성이 있음을 알 수 있다. 또한 Y=Yes 와 X_2=Yes 자료가 1 이면 일반적으로 범주형 자료의 각 셀의 크기가 5이상이 되는 조건이 위배되어 검정력이 작아진다.

표 2. $Y\times Z_1$, $Y\times Z_2$, 카이제곱 검정
Table 2. Chi-Square Test for $Y\times Z_1$, $Y\times Z_2$

	var-iable	Y		test statistic	p-value
var-iable	level	Yes	No		
Z_1	Yes	53	2870	32.5551	0.0001
	No	12	3361		
Z_2	Yes	13	52	1.2273	0.2679
	No	938	5293		

위의 표 2는 하위 계층에 있는 변수들과 관심이 있는 변수와 독립성 검정 결과이며 Y와Z_1은 검정통계량 32.5551이고 p-값은 0.001로 유의 수준 $\alpha=0.05$에서 매우 유의하며 서로 관련성이 있고 Y와Z_2는 관련성이 없음을 알 수 있다.

표 3. Y가 주어진 조건하에서의 Z_1, Z_2 카이제곱 검정
Table 3. For given Y, Chi-Square Test of Z_1, Z_2

	var-iable	Z_1			test statistic p-value
cond-itional var-iable	level		Yes	No	
Y	Yes	Z_2	13	0	3.6792
			40	12	0.055
	No	Z_2	488	450	15.7675
			2382	2911	0.0001

위의 표 3은 Y 변수가 Yes로 주어진 조건하에서 변수 Z_1와Z_2의 독립성 검정을 한 결과는 관련성이 있고 Y 변수가 No로 주어진 조건하에서 변수 Z_1와Z_2의 독립성 검정을 한 결과는 관련성이 없다고 할 수 있다. Y=Yes와 Z_1=No 의 셀의 개수가 0이므로 검정의 신뢰성이 떨어진다.

위의 분석절차에서와 같이 분할표 검정은 개별 변수들 간의 관련성을 파악하기 위해서는 유용하다. 분할표 검정은 검정 통계량 분포 조건과 셀의 관측치 개수 등 필요한 제한적 조건도 만족해야한다. 그러나 베이지안 네트워크 모형 분석은 확률을 이용한 분석이므로 제한 조건 없이 변수들 간의 관련성 분석이 가능하다.

3.2 베이지안 네트워크 모형

다차원 분할표의 모형을 베이지안 네트워크 모형으로 나타내면 첫 번째 층에는 효과에 관련된 식습관 관련 문항 X_1, X_2, X_3을 배열하고, 두 번째 층에는 결과에 관련된 고위험군 문항 Y, 세 번째 층에는 인구학적 나이와 성별 문항Z_1, Z_2를 배열하여 필요한 주변 확률 과 결합 확률, 조건부 확률의 각각의 분포에 대한 분석하는 모형은 그림 4와 같다.

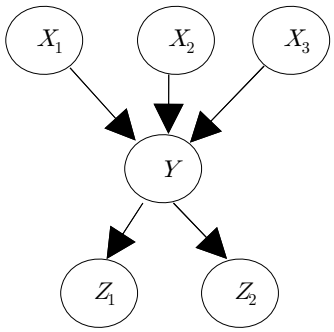

그림 4. 사례 자료에서의 베이지안 네트워크 모형
Fig. 4. Bayesian Network Model for the Simulation Data

베이지안 네트워크 모형에서의 전체 확률분포의 계산은 식 4와 같다.

$$p(X_1,X_2,X_3,Y,Z_1,Z_2) = p(X_1)p(X_2|X_1)p(X_3|X_1,X_2)p(Y|X_1,X_2,X_3)\, p(Z_1,|X_1,X_2,X_3,Y)p(Z_2|X_1,X_2,X_3,Z_1) = p(X_1)p(X_2)p(X_3)p(Y|X_1)p(Y|X_2)p(Y|X_3)\, p(Z_1|Y)p(Z_2|Y) \quad (4)$$

주변 확률 분포 $p(X_1),p(X_2),p(X_3)$는 표 4와 같다.

표 4. 주변 확률 분포 $p(X_1),p(X_2),p(X_3)$
Table 4. Marginal probability distribution for $p(X_1),p(X_2),p(X_3)$

$p(X_1)$	true	0.7754042
	false	0.2245958
$p(X_2)$	true	0.4145516
	false	0.5854484
$p(X_3)$	true	0.008433658
	false	0.991566342

조건부 확률은 $p(Y|X_1)$, $p(Y|X_2)$, $p(Y|X_3)$, $p(Z_1|Y)$, $p(Z_2|Y)$은 표 5와 같다.

표 5. 조건부 확률 분포 $p(Y|X_1),p(Y|X_2),p(Y|X_3),p(Z_1|Y),p(Z_2|Y)$
Table 5. Conditional probability distribution for $p(Y|X_1),p(Y|X_2),p(Y|X_3),p(Z_1|Y),p(Z_2|Y)$

variable/level		true	false
$p(Y\|X_1)$	true	0.6713358	0.3286642
	false	0.7764870	0.2235130
$p(Y\|X_2)$	true	0.3029224	0.6970776
	false	0.4157131	0.5842869
$p(Y\|X_3)$	true	0.018574097	0.9814259
	false	0.008328151	0.9916718
$p(Z_1\|Y)$	true	0.01806473	0.9819353
	false	0.00356700	0.9964330
$p(Z_2\|Y)$	true	0.013695696	0.9863043
	false	0.009692785	0.9903072

확률변수 X_1,X_2,X_3,Y들의 연관성 분석은 관측된 $Y=y$값에 대한 전체 확률 $p(X_1,X_2,X_3,Y=y,Z_1,Z_2)$로 나타난다. 특히 관측된 $Y=y$에 대한 조건부 확률 Z_1,Z_2는 식 5와 같다.

$$p(Z_1,Z_2|Y=y) = p(Z_1|Y=y)p(Z_2|Z_1,Y=y) = p(Z_1|Y=y)p(Z_2|Y=y). \quad (5)$$

확률변수들의 연관관계 또는 인과관계를 분석하는데 베이지안 네트워크의 모형은 특정 변수의 변화량에 따라 관심 있는 변수의 확률 변화량으로 분석이 가능하다. 또한 베이지안 네트워크의 모형은 확률 값에 대하여 분석하기 때문에 일반적으로 요구되어지는 자료의 정규성 및 범주의 작은 빈도수에 대해서도 고려하지 않아도 된다.

4. 결론

본 논문에서는 설문조사 분석에서 자주 사용되어지는 범주형 자료 분석을 베이지안 네트워크 모형으로 독립성 및 교호작용 분석이 가능하다는 것을 보여 주었다. 범주형 자료 분석에서 확률변수 X, Y의 카이제곱 독립성 검정 가설을 살펴보면 $H_0 : p(X|Y) = p(X)$이며 베이지안 네트워크 모형에서는 두 확률변수의 독립성을 식 5로 로 분석 할 수 있으며 덧붙여 Y의 조건에 대한 X의 조건부 확률에 대한 분석도 가능하다.

$$p(X,Y) = p(X)p(Y|X) = p(X)p(Y) \quad (5)$$

일반적으로 선형모형 분석은 자료의 정규분포를 가정하여 분석하며 자료의 선형성, 독립성, 등분산성 등을 만족하지 못하면 분석에서 심각한 오류를 발생할 수 있다. 또한 범주형 자료 분석에서는 범주의 빈도수가 작으면 검정통계량의 분표에서 오류를 발생 한다. 그러나 베이지안 네트워크 모형은 확률 변수들의 확률 값을 이용하여 분석하기 때문에 자료의 정규성 및 범주의 작은 빈도수에 대해서 고려하지 않아도 되는 장점이 있다. 그러므로 베이지안 네트워크 모형은 설문조사에서 문항들의 상호관계를 분석하는 데 매우 유용하게 사용이 가능하다. 향후 과제로는 베이지안 네트워크 모형을 비 순환구조인 경로 분석에 적용한다면 일반적인 경로 분석에서 필요한 제한조건에 관계없이 모형 적합이 가능 할 것이다.

REFERENCES

[1] A, Onisko, M. J. Druzdzel and H. Wasyluk, "Learning Bayesian network parameters from small data set: Application of Noisy-Or gates," Int. J of Approximate Reasoning, vol. 27, no. 2, pp. 165-182, 2001.

[2] Cooper, Gregory F and Herskovits Edward, "A Bayesian Method for the Induction of Probabilistic Networks from Data," Machine Learning, vol. 9, pp. 309-347, 1992.

[3] Heckerman, David, Geiger, Dan and Chickering, David M., "Leraning Bayesian Networks: The Combination of Knowledge and Statistical Data," Machine Learning, vol. 20, 197-203, 1995.

[4] Johnson, R.A and Wichern, D.W., "Applied Multivariate Statistical Analysis", Prentice Hall, 1992.

[5] R. Kohavi and G. H. John, "Wrappers for feature subset selection," Artificial Intelligence, vol. 97, pp. 273-324, 1997.

[6] S. Hwang, L. Lee, "A Score-Based bayesian network learning method by adopting Minimum Description Length principle", Proceeding of KIISE, vol. 33, no. 2, pp. 412-415, 2006.

[7] S. Jung, D. Lee, G. Lee, "Reducing Uncertainty of Bayesian Networks by Reducing Variances of Probability Distributions", Proceeding of KIISE, vol. 33, no. 2, pp. 238-243, 2006.

[8] S. lim, S. Cho, "Automatic Construction of Hierarchical Bayesian Networks for Topic Inference of Conversational Agent", KIISE, vol.33 no. 10, pp. 877-885, 2006.

[9] T.W. Anderson, "An Introduction to Multivariate Statistical Analysis", JohnWiley & Sons, 1971.

[10] Y. Sung,"Applied Multivariate Statistical Analysis", Tamjin Press, 1998.

저자약력

김 용 철(Yong_Chul Kim) **[정회원]**

- 1985년 2월 : 경희대학교 수학과(학사)
- 1991년 5월 : 미주리대학교 통계학과(석사)
- 1994년 5월 : 미주리대학교 통계학과(박사)
- 1996년 3월 ~현재 : 용인대학교 물류통계정보학과 교수

<관심분야>

IT융합기술, Data Mining, AI정보기술, 딥러닝