

베이지안 네트워크를 활용한 정신장애 질병 섬망(delirium)의 주요 요인 네트워크 규명

이제영¹ · 최영진²

¹영남대학교 통계학과, ²영남대학교 통계학과

(2010년 12월 접수, 2011년 2월 채택)

요약

정신장애 질병과 관련된 인자를 찾기 위해 쉽고 간단하게 위험인자를 얻을 수 있는 로지스틱 회귀모형을 주로 이용하였다. 본 논문에서는 데이터마이닝 기법인 로지스틱 회귀모형과 신경망, C5.0, Cart, 베이지안 네트워크를 기저질환과 밀접하게 연관된 가역적 기질성 정신장애인 섬망(delirium) 자료에 적용하여 베이지안 네트워크 기법을 최적의 모형으로 선택하였다. 또한 베이지안 네트워크 기법을 활용하여 정신장애 질병인 섬망과 관련된 주요 위험인자 간 네트워크를 규명하였다.

주요어: 데이터마이닝, 베이지안 네트워크, 정신장애 질병 섬망.

1. 서론

현대사회에는 자살과 우울증과 같은 정신장애와 관련된 질병들이 증가하고 있는 추세이다. 따라서 많은 사람들이 정신장애 및 질병에 대한 관심도 점차 높아지고 있다. 우리는 정신장애 질병 중에서도 가역적인 기질성 정신장애인 섬망(delirium)이라는 질병에 관심을 가졌다. 정신장애 질병인 섬망은 혼동 상태와 흥분 행동을 나타내는 다양한 질환 중 하나로 기저질환과 밀접하게 연관된 정신장애 질병이다. 특히 고령 입원 환자의 경우 60%에서 발병하며 기저질환의 사망률과 이환율에 밀접하게 연관되어 임상적 중요성이 있다 (Inouye, 1994; Cole과 Primeau, 1993; Inouye 등, 1999). 지금까지 정신장애 질병인 섬망에 대한 연구가 주로 외국에서 이루어졌다. 중환자실 입원 환자의 섬망 위험인자에 대한 전향적 연구 (Dubois 등, 2001), 중환자실 입원 환자의 섬망 위험인자 및 예후에 관한 대규모 코호트 연구, 중환자실 입원 환자의 섬망에 관련된 32가지 위험인자 확립한 연구 (Arend와 Christensen, 2009) 등이 있다. 그러나 중환자실이 아닌 응급실을 통한 입원환자의 섬망 위험인자에 대한 국내 연구는 희박한 실정이다 (Kwak 등, 2011). 정신장애 질병과 관련된 위험인자를 찾는 방법으로 주로 로지스틱 회귀모형 (Kwak 등, 2011)을 이용해왔다. 로지스틱 회귀모형은 질병과 관련된 위험인자들을 손쉽게 찾을 수 있다는 장점이 있으나 위험인자 간 상호작용이 높으면 실제로 의미 있는 위험인자들을 충분히 얻지 못하는 어려움도 있다. 따라서 우리는 이러한 점을 고려하여 좀 더 다양한 데이터마이닝 기법에 섬망 자료를 적용한다. 여기서 데이터마이닝이란 대량의 데이터나 복잡한 구조의 데이터들을 정교한 통계분석과 모델링 테크닉을 이용하여 정확히 식별되지 않는 패턴이나 자료간의 상관관계를 밝혀내어 여러 가지 결과를 예측해 내는 통계적 기법이다 (Tan 등, 2007). 본 논문에서는 데이터마이닝 기법인 로지스틱 회귀모형(logistic), 신경망(neural network), C5.0, Cart(Classification and Regression Tree), 베이지안 네

¹교신저자: (712-749) 경북 경산시 대동 214-1, 영남대학교 통계학과, 교수. E-mail: jlee@yu.ac.kr

트위크(Bayesian network)를 정신장애 질병인 섬망 자료에 적용하였을 때 가장 정확도가 높은 베이직한 네트워크를 최종모형으로 하여 섬망과 관련된 주요 위험인자를 선별하고 그들 간 네트워크를 규명한다. 본 논문의 2장에서는 데이터마이닝 기법들에 대해서 소개하고 3장에서는 섬망을 데이터마이닝 기법에 적용한 결과를 살펴본다. 4장에서는 최종 모형으로 선택된 베이직한 네트워크 기법을 활용하여 정신장애 질병 섬망과 관련된 주요 위험인자 간 네트워크를 규명한다.

2. 데이터마이닝 기법 소개

2장에서는 정신장애 질병인 섬망에 적용할 다양한 데이터마이닝 기법들에 대해서 소개한다 (이용원, 2004). 2.1절에서는 로지스틱 회귀모형 기법을 소개하고 2.2절에서는 신경망, 2.3절에서는 C5.0, 2.4절에서는 Cart, 마지막으로 2.5절에서는 베이직한 네트워크 기법을 소개한다.

2.1. 로지스틱 회귀모형 기법 소개

로지스틱 회귀모형은 병리학 연구에서 질병과 연관 있는 위험요인들을 식별하거나, 임상 연구 자료에서 중요한 요인들을 식별하는 탐색적 분석에 많이 적용된다 (이재원 등, 2005; 허명희와 이용구, 2008). 로지스틱 회귀모형에서 종속변수는 주로 이항반응이지만 여러 개의 수준을 갖는 다항 반응일 수도 있다. 설명변수는 연속형, 이산형인 경우 둘 다 가능하며 정규분포의 가정을 전제로 하지 않는다. 본 논문에서는 이항반응에 대해서만 살펴보기로 한다. 이항반응변수를 종종 베르누이 변수라고도 한다. 이 변수에 대한 분포는 성공에 대한 확률 $P(Y = 0) = 1 - p_x$ 와 실패에 대한 확률 p_x 로 명시된다. 여기서 성공확률 p_x 에 대해 다음과 같은 선형 확률모형을 생각할 수 있다.

$$p_x = \alpha + \beta x.$$

하지만 위 모형은 구조적 단점을 가진다. p_x 가 (0,1)의 값을 가지고 $\alpha + \beta x$ 는 실수 전체의 값을 가지므로 성공확률 p_x 가 범위 외의 값을 가질 수 있게 된다. 또한 최소제곱추정량(LSE)은 분산 $\text{Var}(Y) = p_x(1 - p_x)$ 에서 분산이 x 의 영향을 받기 때문에 분산이 일정하지 않아 더 이상 최소 분산을 가지지 못한다. 따라서 이런 식 대신에 성공확률 p_x 에 대해 k 개의 설명변수 X_1, X_2, \dots, X_k 와 비선형인 아래의 식

$$p_x = \frac{\exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k)}{1 + \exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k)}$$

과 같은 함수를 생각하게 된다. 위 식에서 성공확률에 대한 오즈(odds)는 아래 식

$$\frac{p_x}{1 - p_x} = \exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k)$$

와 같고 성공확률의 오즈에 로그를 씌운 로그 오즈는 아래의 선형식으로 나타낼 수 있다.

$$\ln \left(\frac{p_x}{1 - p_x} \right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k.$$

위 식을 로지스틱 회귀모형이라 한다.

2.2. 신경망 기법 소개

인간 두뇌에 관한 연구는 오래전부터 계속되어 왔다. 컴퓨터의 발달로 컴퓨터를 이용한 인공지능의 구현을 위하여 많은 노력을 하고 있으며 최근에는 패턴인식, 의사결정 등에서 컴퓨터 보다 뛰어난 인간의

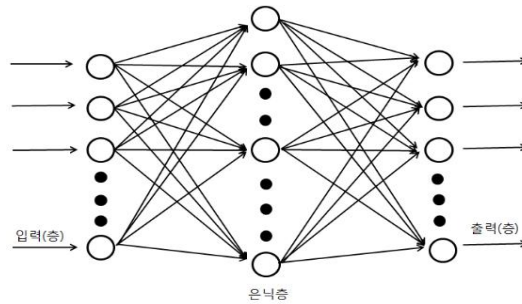


그림 2.1. 신경망의 기본 구조

두뇌를 모방한 신경망의 연구가 지속되고 있다 (Warren, 1994). 신경망은 몇 개의 뉴런과 이것들이 배열된 층(layer)으로 구성된다. 각 뉴런은 특정의 작업을 수행하고 신경망은 이들 뉴런을 연결함으로써 자극과 반응간의 관계를 학습하고, 새로운 데이터에 대한 예측 및 분류를 하게 된다. 예측은 목표변수가 실수값인 경우가 되며, 분류는 목표변수가 범주형인 경우를 일컫는다. 신경망은 인간의 신경-두뇌 시스템을 흉내 낸 것으로 그림 2.1과 같이 입력층(input layer), 은닉층(hidden layer), 출력층(output layer) 등 3개의 층으로 구성되어 있다. 여기에서 원으로 표시된 것을 노드라고 하며, 그림 2.1은 은닉노드들을 1개의 은닉층에 넣은 신경망의 경우이다.

신경망은 기본적으로 입력층의 각 입력노드로부터 전달된 신호들을 모아 선형결합을 하여 얻어지는 은닉층의 은닉노드의 신호를 출력층의 출력노드에서 전달받아 결합함으로써 최종반응을 보내게 되는 과정을 가지게 된다. 신경망은 특히 복잡한 비선형에 적합 시 좋은 결과를 얻을 수 있다.

2.3. C5.0 기법 소개

C5.0은 의사결정나무 알고리즘 중 하나로 Ross Quilan이라는 호주 학자에 의해 수년동안 연구하여 발표되어진 C4.5를 바탕으로 하여 개발된 데이터마이닝 이론 가운데 하나이다. 초기에는 ID3(Iterative Dichotomiser 3) 이름으로 하여 1986년에 몇가지 상업적 상품을 가지고 기계학습 분야에 적용하여 발표되어졌다. ID3는 C5.0은 정보이론에 따른 엔트로피(entropy) 개념을 이용하여 가장 간단한 결정나무를 구하고 있다. 정보이론 분야의 엔트로피의 개념은 미국의 응용 수학자이자 정보이론의 창시자인 C.E. Shannon에 의해 1948년 도입되었다. 엔트로피는 기호 배열상의 무질서의 정도, 즉 정보의 불확실성과 이에 관련된 정보량의 추정에 활용된다. 정보이론에서는 k 개의 사건이 p_1, p_2, \dots, p_k 의 확률로 발생한다면, 비트(이진) 단위의 정보의 엔트로피는

$$\text{Entropy} = - \sum_{l=1}^k p_l (\log_2 p_l)$$

이 된다 (허명희와 이용구, 2008). 여기서 Entropy ≥ 0 일 때 Entropy는 이진단위로 표시된 정보의 부족한 정도(불확실성), 즉 이제부터 얻어야 하는 정보량을 나타내며, Entropy = 0이면 부족한 것이 없는 완전한 정보를 얻었음을 뜻한다. C5.0은 엔트로피 기준에서 가장 엔트로피를 낮추는 분리 변수를 찾는다.

2.4. CART 기법 소개

Cart (Breiman 등, 1984)는 설명변수들과 목표변수로 이루어진 자료들에서 설명변수들의 특성에 따라 자료들을 이진분류하여, 2개의 하위노드를 생산하는 과정을 반복하여 자료들을 목표변수의 값이 유사

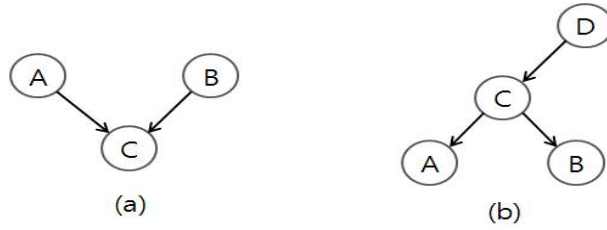


그림 2.2. DAG를 사용한 확률적 관련성 표현 (Tan 등, 2007)

한 부분집합으로 만드는 방법이다. 여기서 이지분류란 상위노드로부터 하위노드가 2개만 형성되는 것을 의미한다. 또한 목표변수에 가장 유의적인 설명변수를 찾는 것이 아니고, 노드의 불순도와 다양성을 가장 많이 줄여주는 설명변수의 최상의 절단점(연속형 설명변수)이나 최상의 그룹(범주형 설명변수)을 선택하는 방식을 채택 하고 있다. 이지분류를 수행하는 알고리즘으로 지니계수(gini index: 이산형 목표변수인 경우 적용) 또는 분산의 감소량(연속형 목표변수인 경우 적용)을 이용한다. 지니계수는 이산형 목표변수에 대해서 분류기준으로 사용되며, 이 계수는 각 마디에서의 불순도 또는 다양도를 측정하는 척도중의 하나이다. 만일 자료셋이 k 개의 범주로 분할되고 범주 비율이 p_1, p_2, \dots, p_k 일 때, 지니계수는 다음과 같이 정의된다.

$$G = 1 - \sum_{l=1}^k p_l^2 = 1 - \sum_{l=1}^k \left(\frac{n_l}{n}\right)^2,$$

여기서 n 은 그 마디에 포함된 관찰치 수, n_l 는 목표변수의 l 번째 범주에 속하는 관찰치의 수를 의미한다. Cart는 지니 계수를 기준으로 가장 성취도가 좋은 변수 및 수준을 찾는다.

2.5. 베이지안 네트워크 기법 소개

베이지안 네트워크는 변수 집합사이의 확률적인 관계를 네트워크 형태로 표현하는 방법이다 (Tan 등, 2007). 특정 분야의 영역 지식을 확률적으로 표현하는 대표적인 수단으로, 변수들 간의 확률적 의존 관계를 나타내는 그래프 (Heckerman, 1995; Jensen, 1996; 김경현, 2005)와 각 변수별 조건부 확률로 구성된다. 따라서 베이지안 네트워크는 각 노드마다 조건부 확률표를 갖는 방향성 비순환 그래프(Directed Acyclic Graph; DAG)로 정의할 수 있다 (Hwang과 Zhang, 2005). 그림 2.2(a)는 확률변수 A, B, C의 관계를 네트워크로 표현한 것이다. A와 B는 독립변수이고 각각은 변수 C에게 직접적인 영향을 가지고 있다. 그래프에서 노드는 변수를, 화살표는 변수 쌍 사이의 종속관계를 보여준다. 그림 2.2(b)를 살펴보면 C는 A와 B의 부모노드이고 A와 B는 C의 자식노드가 된다. D는 A와 B의 조상노드가 된다.

일반적으로 베이지안 네트워크는 다른 노드들의 확률값들을 기초로 특정 노드가 가질 값에 대한 조건부 확률을 계산하는데 이용할 수 있다. 따라서 하나의 베이지안 네트워크는 한 개체의 다른 속성들의 값이 주어졌을 때 분류 클래스 노드의 사후 확률 분포를 구해줌으로써 개체들에 대한 하나의 자동 분류기로 이용될 수 있다. 즉, 베이지안 네트워크를 기초로 분류 클래스를 확률적으로 예측할 수 있다. 베이지안 네트워크는 조건부 확률 계산에 베이즈 정리를 이용하여 다음의 수식을 얻을 수 있다.

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}.$$

베이즈 정리는 사후 확률을 구할 때 즉, 예측문제를 해결할 때 사용될 수 있다. 베이지안 네트워크는 화살표로 연결된 노드들과는 서로 의존적이고 자손이 아닌 노드들과는 조건부 독립적이라는 속성을 가지

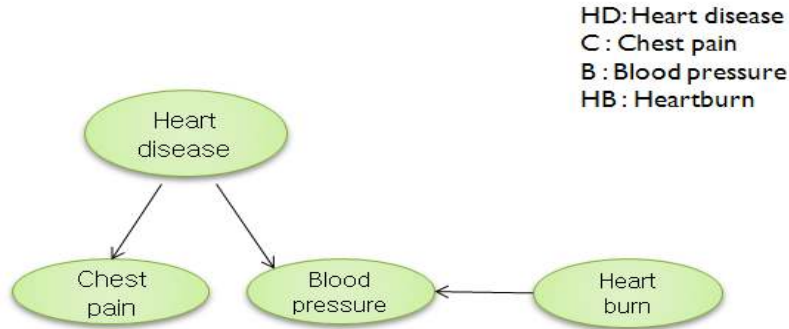


그림 2.3. 심장병과 위험인자의 관계 표현한 베이지안 네트워크

표 2.1. 심장병과 위험인자 관계의 조건부 확률

(a)		(b)		
	C = Yes		B = Yes	
HD = Yes	0.85	HD = Yes	HB = Yes	0.8
HD = No	0.2	HD = Yes	HB = No	0.6
		HD = No	HB = Yes	0.4
		HD = No	HB = No	0.1

고 있다. 이것을 아래의 식으로 표현할 수 있다. 먼저 조건부 확률의 성질에 의해서 다음이 성립하고,

$$P(A|B) = \frac{P(A, B)}{P(B)}$$

곱셈규칙에 의해서

$$P(A, B) = P(A|B)P(B) = P(B|A)P(A)$$

가 성립한다. 결과적으로 체인룰을 사용하면,

$$P(X_1, X_2, \dots, X_n) = P(X_1) \prod_{i=2}^n P(X_i | X_1, \dots, X_{i-1})$$

의 식이 만들어진다. 베이지안 네트워크는 가정의 단순함에도 불구하고 많은 연구를 통해 비교적 높은 분류 성능을 보여주는 것으로 알려져 있다. 그림 2.3은 심장병(Heart disease)과 연관된 위험인자인 속쓰림(Heartburn), 가슴앓이(Chest pain), 고혈압(Blood pressure)을 갖고 있는 환자를 모델링 한 베이지안 네트워크의 예를 보여준다 (Heckerman, 1997; Tan 등, 2007).

베이지안 네트워크는 그림 2.3과 같이 각 변수들에 대한 의존성을 그래프로 표현할 뿐 아니라 표 2.1과 같은 조건부 확률도 함께 표현할 수 있다. 예를 살펴보면 가슴앓이와 심장병에 대해서 종속성을 가지며 가슴앓이에 대한 확률값이 표 2.1(a)에 나타나있다. 또한 고혈압은 심장병과 속쓰림에 영향을 받으며 표 2.1(b)는 고혈압에 대한 확률값을 나타낸다. 즉, 표 2.1은 가슴앓이를 가질 확률과 고혈압을 가질 확률을 표현한 것이다. 따라서 여기서 알고 있는 확률을 바탕으로 가슴앓이가 있을 때, 고혈압이 있을 때 심장병이 걸릴 확률을 구할 수가 있다. 3장에서는 2.1절~2.5절에 소개된 데이터마이닝 기법을 정신장애 질병인 섬망에 적용하였을 때 가장 높은 정확도를 가지는 모형을 최적의 모형으로 선택하고 4장에서는 최종 모형으로 선택된 데이터마이닝 기법을 섬망 자료에 적용한 결과를 살펴본다.

표 3.1. 정신장애 질병 섬망과 관련된 14가지 위험인자 (Kwak 등, 2011)

위험인자	독립변수	독립변수 설명
질병에 취약한 인자 (Predisposing)	요양병원 (nursing home)	1: 요양병원 경험 유 0: 요양병원 경험 무
	신장수치 이상 (bun/cre)	1: Bun \geq 20 or Cre \geq 1.7 0: 정상
	뇌졸중/간질 (stroke/epilepsy)	1: 앓은 경험 있음 0: 앓은 경험 없음
	저산소증 (hyposia)	1: 70~80보다 낮은 경우 0: 정상(70~80)
급성질병 (Acute illness)	대사이상 (metabolic abnormality)	1: 대사성 산증이 있는 경우 0: 정상
	전해질 불균형 (electrolyte imbalance)	1: Na; 135~145, K; 3.5~5.5 벗어난 경우 0: 정상
	영양실조 (malnutrition)	1: 영양실조 0: 정상
	혈역동학적 이상 (hemodynamic instability)	1: Hb down, 혈압 \leq 90 0: 정상
약리학 (Pharmacology)	진통제 (analgesics)	1: 약물사용 0: 사용없음
	항생제 (antibiotics)	1: 약물사용 0: 사용없음
	기관지 확장제 (bronchodilator)	1: 약물사용 0: 사용없음
	심장약 (cardiacdrug)	1: 약물사용 0: 사용없음
	이뇨제 (diuretic)	1: 약물사용 0: 사용없음
	진정제 (sedative)	1: 약물사용 0: 사용없음

3. 데이터마이닝 기법 적용 결과 및 최종모형 선택

3.1절에서는 본 논문에 사용된 정신장애 질병인 섬망 자료를 살펴보고 3.2절에서는 2장에서 소개된 데이터마이닝 기법을 섬망 자료 적용하였을 때 정확도가 가장 높은 모형을 최종모형으로 선별한다.

3.1. 실험자료

2008년 1월부터 2009년 12월까지 24개월 동안 종합병원 응급실을 통해 내과 병동으로 입원한 70세 이상 고령 환자를 대상으로 구성되어졌다. 연구에 포함된 환자의 성별, 나이, 과거력, 사회력, 생체 징후는 의무기록을 바탕으로 조사된 내용이다. 정신장애 질병인 섬망 진단은 정신장애 진단통계 편람 제4판에 의해 정신과적 기저질환을 제외한 기저질환에 의해 섬망으로 진단받음을 기준으로 하였다 (Kwak 등, 2011). 연구에서 사용한 대상자 수는 414명이며 그 중 42명은 정신과 치료를 통해 기저질환으로 인한 섬망으로 진단받았으며 372명은 진단받지 않았다. 본 논문에서 사용한 위험인자는 2009년 Emma Arend 등이 보고한 32가지 위험인자 (Arend와 Christensen, 2009) 가운데 연령을 제외한 31가지 위험인자 중에서 섬망과 유의한 관련이 있다고 선별된 14가지 위험인자를 사용하였다. 표 3.1은 분석에 사용된 14가지 위험인자를 나타낸 것이다.

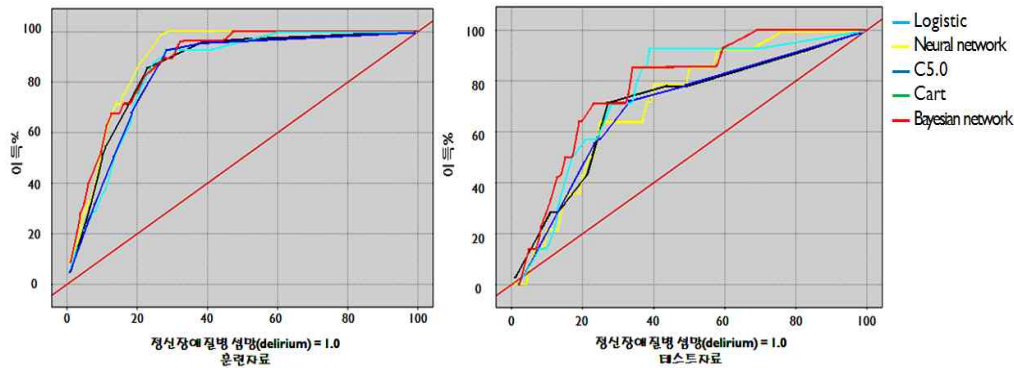


그림 3.1. 훈련자료와 테스트자료의 이득도표

표 3.2. 데이터마이닝 기법들의 적용결과 정확도 비교

모델	훈련자료 정확도	테스트자료 정확도
Logistic	0.807	0.738
Neural Network	0.855	0.767
C5.0	0.794	0.738
Cart	0.836	0.796
Baysian Network	0.852	0.845

3.2절에서는 2.1절~2.5절에서 소개된 데이터마이닝 기법인 로지스틱 회귀모형, 신경망, C5.0, Cart, 베이지안 네트워크를 정신장애 질병인 섬망 자료에 적용한 결과를 살펴본다.

3.2. 데이터마이닝 기법 적용 결과 및 최종모형 선택

본 논문에서 사용된 섬망 자료 414명 중에서 42명만이 섬망을 앓고 있는 환자들이다. 따라서 섬망이 있는 환자와 그렇지 않은 환자들의 비율을 맞추기 위해 부스팅을 통하여 균형화 작업을 실시해 섬망의 비율을 50:50으로 맞춘 자료를 사용하였다. 또한 훈련자료와 테스트자료의 비율을 각각 70%, 30%로 분할하여 분석하였다. 그림 3.1은 각각 훈련자료와 테스트자료에서의 이득도표를 나타낸 그림이다. 훈련자료에서는 신경망(0.855)과 베이지안 네트워크(0.852)가 높게 나타난 것을 볼 수 있다. 또한 테스트자료에서는 전반적으로 베이지안 네트워크(0.845)가 높게 나타났음을 볼 수 있다.

이득도표에 대한 내용을 자세하게 알아보기 위해 각 데이터마이닝 기법들의 정확도를 구하였다. 표 3.2는 데이터마이닝 기법을 섬망 자료에 적용한 결과의 정확도를 나타낸 것이다.

훈련자료에서의 정확도를 살펴보면 로지스틱 회귀모형은 80.7%, 신경망은 85.5%, C5.0은 79.4%, Cart 기법은 83.6%, 베이지안 네트워크는 85.2%로 나타났다. 훈련자료에서 만들어진 모형을 테스트자료에 적용한 결과를 살펴보면 로지스틱 회귀모형은 73.8%, 신경망은 76.7%, C5.0은 73.8%, Cart는 79.6%, 베이지안 네트워크의 정확도가 84.5%로 베이지안 네트워크 기법이 다른 기법보다 정확도가 매우 높음을 볼 수 있었다. 질병과 관련된 위험인자를 발견하는데 주로 사용되었던 로지스틱 회귀모형 기법보다 약 10% 정도 정확도가 높게 나타났다. 훈련자료와 테스트자료를 종합적으로 봤을 때 정확도가 가장 높게 나타난 베이지안 네트워크를 최종모형으로 선택하였다. 따라서 4장에서는 베이지안 네트워크를 정신장애 질병인 섬망 자료에 적용하여 섬망과 관련된 위험인자를 발견하고 섬망과 주요 위험인자 간 네트워크를 규명한다.

표 4.1. 베이저안 네트워크를 통해 선별된 7가지 주요 위험인자

위험인자	사후 확률
뇌졸중/간질	0.832
대사이상	0.424
신장수치 이상	0.299
혈역동학적 이상	0.184
기관지 확장제	0.132
전해질 불균형	0.111
항생제	0.093

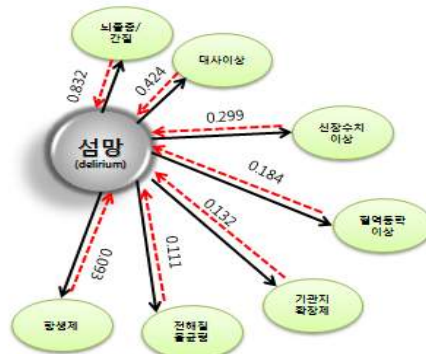


그림 4.1. 정신장애 질병 섬망에 영향을 미치는 주요 위험인자

4. 주요인자간 네트워크 규명을 위한 베이저안 네트워크 적용

3장에서는 다양한 데이터마이닝 기법을 정신장애 질병인 섬망 자료에 적용한 결과 정확도가 가장 높은 베이저안 네트워크가 최종 모형으로 선별되었다. 베이저안 네트워크를 섬망 자료에 적용한 결과 14가지 위험인자 중 뇌졸중/간질, 대사이상, 신장수치 이상, 혈역동학적 이상, 기관지 확장제, 전해질 불균형, 항생제가 정신장애 질병인 섬망에 영향을 미치는 7가지 위험인자로 선택되었다. 섬망에 영향을 많이 미치는 위험인자 순서대로 표 4.1에 나타내었다.

표 4.1은 위험인자가 단일로 존재할 때 섬망의 발병률에 대해서 살펴본 것이다. 먼저 뇌졸중/간질이 있을 때 정신장애 질병인 섬망이 나타날 확률은 약 83.2%로써 다른 기저질환보다도 매우 높은 발병률로 나타났다. 다음으로 대사이상이 있을 때 42.4%의 확률로 섬망의 발병률이 나타났다. 또한 신장수치 이상이 있을 때 29.9%, 혈역동학적 이상이 있는 경우 18.4%, 기관지 확장제를 복용한 경우 13.2%, 전해질 불균형일 때 11.1%, 항생제를 복용했을 때 9.3%로 정신장애 질병 섬망이 걸릴 확률이 나타났다. 그림 4.1은 표 4.1의 결과를 섬망에 영향을 많이 미치는 위험인자 순으로 네트워크로 구현한 것이다.

위 그림을 살펴보면 뇌졸중/간질, 대사이상이 있을 때 섬망에 미치는 영향이 큰 반면에 기관지 확장제, 전해질 불균형, 항생제는 그다지 큰 영향을 미치지 않는 것을 볼 수 있다.

표 4.2는 정신장애 질병 섬망이 걸릴 가능성을 높게 하는 위험인자인 뇌졸중/간질, 대사이상, 신장수치 이상과 다른 위험인자들이 더하여졌을 때의 섬망이 나타날 확률을 나타낸 것이다. 또한 그림 4.2는 정신장애 질병인 섬망에 영향을 미치는 7개의 위험인자 중에서 섬망에 영향을 많이 미치는 주요 3개 위험인자인 뇌졸중/간질, 대사이상, 신장수치 이상을 중심으로 하여 새롭게 네트워크를 구현한 것이다. 먼저 뇌졸중/간질이 있는 환자가 대사이상과 신장수치 이상이 있을 때 정신장애 질병 섬망이 나타날 확률은

표 4.2. 섬망의 발병률에 대한 주요 위험인자 간 상호작용 결과

종속변수	주요 위험인자	확률	위험인자	확률
정신장애 질병 섬망(delirium)	뇌졸중	0.832	대사이상	0.979
			신장수치 이상	0.964
			전해질 불균형	0.888
	대사이상	0.424	혈역동학적 이상	0.681
			기관지 확장제	0.581
			항생제	0.161
	신장수치 이상	0.299	전해질 불균형	0.715
혈역동학적 이상			0.553	
항생제			0.360	

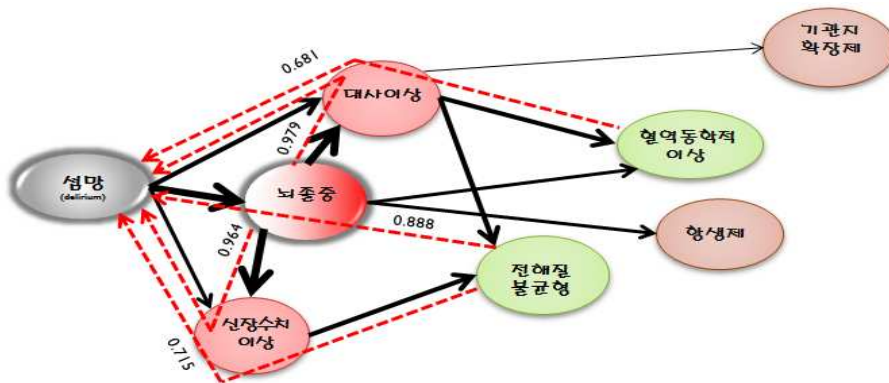


그림 4.2. 섬망과 위험인자들의 베이지안 네트워크 결과

각각 97.9%, 96.4%로 매우 높음을 볼 수 있다. 대사이상과 혈역동학적 이상을 가진 경우와 기관지 확장제를 같이 가지고 있는 환자의 경우에는 각각 68.1%, 58.1%로 대사이상만 가졌을 때 보다 섬망에 대한 발병률이 높게 나타났다. 신장수치 이상과 혈역동학적 이상이 있을 때 55.3%, 신장수치 이상과 전해질 불균형이 있을 때 71.5%로 신장수치 이상만 가졌을 때보다 약 2배 이상 섬망이 걸릴 가능성이 높아진 것을 볼 수 있다.

5. 결론 및 토의

본 논문은 베이지안 네트워크 모형을 적용하여 정신장애 질병인 섬망과 관련된 위험인자를 선별하고 위험인자간 네트워크를 규명한 것이다. 우리는 데이터마이닝 기법인 로지스틱 회귀모형, 신경망, C5.0, Cart, 베이지안 네트워크를 정신장애 질병 섬망 자료에 적용하였을 때 가장 모형의 정확도가 높은 기법인 베이지안 네트워크를 최종모형으로 선택하였다. 베이지안 네트워크를 최종모형으로 하였을 때 정신장애 질병인 섬망에 영향을 미치는 위험인자로 뇌졸중/간질, 대사이상, 신장수치 이상, 혈역동학적 이상, 기관지 확장제, 전해질 불균형, 항생제 7가지 주요 위험인자가 선택되었다. 단일의 위험인자를 가진 환자의 경우, 섬망이 나타날 확률이 높은 위험인자로는 뇌졸중/간질, 대사이상, 신장수치 이상을 가진 경우로 나타났다. 위험인자 간 상호작용의 경우, 뇌졸중/간질과 대사이상, 뇌졸중/간질과 신장수치 이상을 동시에 가진 환자의 경우에는 섬망의 발병률이 각각 97.9%, 96.4%로 뇌졸중/간질 단일로 가진 환자보다 발병률이 높아짐을 볼 수 있었다. 즉, 섬망에 높은 영향을 미치는 뇌졸중/간질, 대사이상, 신장

수치 이상과 다른 위험인자를 동시에 가진 환자는 정신장애 질병인 섬망의 발병률이 매우 높았으며 또한 단일의 위험인자를 가진 경우보다 위험인자 간 상호작용으로 나타날 때 섬망의 발병률이 훨씬 높아짐을 볼 수 있었다 (표 4.2, 그림 4.2). 따라서 우리는 최종모형으로 선택된 베이지안 네트워크를 활용하여 정신장애 질병 섬망에 영향을 미치는 7가지 위험인자를 찾았고 이들 간 네트워크를 규명하였다. 또한 베이지안 네트워크를 통해서 위험인자가 단독으로 정신장애 질병 섬망에 영향을 미칠 때 보다 상호작용으로 영향을 미칠 때 훨씬 더 섬망이 나타날 발병률이 커짐을 밝힐 수 있었다. 베이지안 네트워크를 통해 밝힌 주요 7개 위험인자를 지닌 환자들을 잘 살피고 특히 뇌졸중/간질과 대사이상, 뇌졸중/간질과 신장 수치 이상을 가진 환자를 주의깊게 살핀다면 사전에 큰 위험을 막을 수 있을 것이다.

참고문헌

- 김경현 (2005). 베이지안 네트워크에 기초한 백혈병 유전자데이터의 분석, *B.I. Thesis* 1-15.
- 이용원 (2004). 고혈압의 위험요인에 대한 데이터 마이닝 모형 분석 - 종합건강검진 데이터를 바탕으로, *M.S. Thesis* 1-52.
- 이재원, 박미라, 유한나 (2005). <생명과학연구를 위한 통계적 방법>, 자유 아카데미, 서울.
- 허명희, 이용구(2008). <데이터 마이닝 모델링과 사례>, 한나래, 서울.
- Arend, E. and Christensen, M. (2009). Delirium in the intensive care unit: A review, *British Association of Critical Care Nurses, Nursing in Critical Care*, **14**, 145-154.
- Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Decision Tree*, Chapman & Hall.
- Cole, M. and Primeau, F. (1993). Prognosis of delirium in elderly hospital patients, *Canadian Medical Association Journal*, **149**, 41-46.
- Dubois, M., Strobik, Y., Bergeron, N., Dumont, M. and Dial, S. (2001). Delirium in an intensive care unit: A study of risk factors, *Intensive Care Medicine*, **27**, 1297-1304.
- Heckerman, D. (1995). A tutorial on learning with Bayesian networks, *Technical Report MSR-TR-95-06, Microsoft Research*, 1-58.
- Heckerman, D. (1997). *Bayesian Networks for Data Mining*, Kluwer Academic Publishers, 79-119.
- Hwang, K. B. and Zhang, B. T. (2005). An Introduction to Bayesian Networks: Concepts and Learning from Data, "<http://bi.snu.ac.kr/Courses/4ai05f/introBN.pdf>", *SNU Biointelligence Lab*, 1-93.
- Inouye, S. (1994). The dilemma of delirium: Clinical and research controversies regarding diagnosis and evaluation of delirium in hospitalized elderly medical patients, *The American Journal of Medicine*, **97**, 278-288.
- Inouye, S., Schlesinger, M. and Lyndon, T. (1999). Delirium: A symptom of how hospital care is failing older persons and a window to improve quality of hospital care, *The American Journal of Medicine*, 565-573.
- Jensen, F. (1996). *An Introduction to Bayesian Networks*, Springer-verlag, New York.
- Kwak, K., Lee, S. B. and Do, B. S. (2011). Delirium in an emergency department: A study of risk factors, *Journal of the Korean Society of Emergency Medicine*, Submitted.
- Tan, P., Steinbach, M. and Kumar, V. (2007). *Introduction to Data Mining*, Addison Wesley Longman, California.
- Warren, S. (1994). Neural networks and statistical models, *Proceedings of the 19th Annual SAS Users Group International Conference*, 1-13.

Network Identification of Major Risk Factors Associated with Delirium by Bayesian Network

Jea-Young Lee¹ · Young-jin Choi²

¹Department of Statistics, Yeungnam University

²Department of Statistics, Yeungnam University

(Received December 2010; accepted February 2011)

Abstract

We analyzed using logistic to find factors with a mental disorder because logistic is the most efficient way to assess risk factors. In this paper, we applied data mining techniques that are logistic, neural network, c5.0, cart and Bayesian network to delirium data. The Bayesian network method was chosen as the best model. When delirium data were applied to the Bayesian network, we determined the risk factors associated with delirium as well as identified the network between the risk factors.

Keywords: Bayesian network, data mining, delirium mental disorder.

¹Corresponding author: Professor, Department of Statistics, Yeungnam University, Kyungsan 712-749, Korea. Email: jlee@yu.ac.kr