碩士學位 請求論文

指導敎授 金 贊 敏

개선된 트리 증강 나이브 베이즈 모형의 제안

成均館大學校 一般大學院

統計學科

林 賢 祐

碩士學位請求論文

개선된 트리 증강 나이브 베이즈 모형의 제안

2022

林賢祐

碩士學位 請求論文

指導敎授 金 贊 敏

개선된 트리 증강 나이브 베이즈 모형의 제안

Proposal of the improved tree augmented naive Bayes model

成均館大學校 一般大學院

統計學科

林 賢 祐

碩士學位 請求論文

指導敎授 金 贊 敏

개선된 트리 증강 나이브 베이즈 모형의 제안

Proposal of the improved tree augmented naive Bayes model

이 論文을 統計學 碩士學位請求論文으로 提出합니다.

2022 年 4 月 日

成均館大學校 一般大學院

統計學科

林 賢 祐

이 論文을 林 賢 祐 의 統計學

碩士學位 論文으로 認定함.

2022 年 6 月 日

審査委員長

審査委員

審査委員

목차

표목차

그림목차

논문요약

개선된 트리 증강 나이브 베이즈 모형의 제안

최근 많은 분야에서 빅데이터에 적용한 분류 기법의 사용이 증가하면서 다양한 머신러닝 기법의 사용이 증가하고있다. 그중 방향성 비순환 그래프라고도 불리는 베이지안 네트워크는 빅데이터에 내제되어 있는 정보를 사용하여 변수 간의 인과관계를 시각화하는 특징으로 인해 많은 분야에서 사용되고 있다. 나이브 베이즈와 TAN은 베이지안 네트워크 분류기의 대표적인 기법으로, 특히 TAN은 실생활에서 쉽게 위배되는 나이브 베이즈의 조건부 독립이라는 강한 가정을 완화시키고 변수간 인과관계를 시각화할 수 있는 장점으로 인해 주로 사용되고 있다. 그러나 기존 TAN의 네트워크는 구축 과정에서 모든 변수를 사용하기 때문에 인과관계에 포함되지 않아야 할 변수가 네트워크에 포함되는 문제점을 갖고 있고, 이로 인해 부자연스러운 네트워크가 형성되어 관계해석에 어려움을 주었다.

본 논문에서는 이러한 문제점을 해결하기 위해 네트워크의 구축과정에서 제외하고자 하는 변수 집합을 클래스 변수에 포함시키고 인과관계에서 배제하는 새로운 방식의 TAN(New Tree-Augmented Naive Bayes; NTAN)을 제안한다. 변수를 단순하게 제외시키는 것이 아닌, 클래스 변수에 포함시킴으로써 변수가 가지고 있는 정보를 사전정보로 사용하여 정보의 손실을 막고 자연스러운 네트워크를 구축하였다.

시뮬레이션 결과 NTAN이 TAN과 비교하여 유사한 분류성능을 나타내나 더 자연스러운 네트워크를 형성하여 변수간 관계해석을 용이하게 만듦을 확인하였다. 그리고 실제 건강검진 데이터와 UCI에서 제공하는 유방암 재발 여부 데이터에 NTAN을 적합하여 자연스러운 네트워크를 구축하였으며 이를 통해 TAN보다 적절한 변수간 관계해석을 진행하였다. 뿐만 아니라 NTAN의 분류 성능이 TAN보다 향상된 것을 확인하였다. 따라서 논문에서 제안하는 NTAN이 기존의 TAN과 비교하여 자연스러운 네트워크를 형성하고 분류성능 또한 향상시키는 것을 재차 확인하였다.

주제어: 베이지안 네트워크, 방향성 비순환 그래프, Tree Augmented Naïve Bayes, 머신러닝

제 1 장 서 론

현대 사회는 데이터의 양이 방대한 빅데이터의 시대이고, 많은 기업과 정부는 빅데이터로부터 정보를 추출하여 인사이트를 도출하고 예측 및 분류를 통해 가치를 창출해 내고 있다. 특히 분류기법의 사용이 많아지면서 분류를 위한 다양한 머신러닝 기법이 사용되고 있다.

분류기법 중 하나인 K-NN(Cover 등, 1967)은 간단한 모델로써 다양한 종류의 데이터에 쉽게 적용되어 사용되고 있다. Sharma 등 (2016)은 자궁경부암 Pap 이미지 데이터에 K-NN 분류기를 적합하여 자궁경부암의 진행 단계를 분류하였다. Pap 이미지에는 세포질과 세포핵이 포함되어 있어 암의 진행 단계를 파악할 수 있다. Islam 등 (2018)은 facebook 사용자의 우울증을 감지하기 위해 K-NN 분류기법을 사용하였다. 사용자의 댓글, 게시물 데이터를 통해 감정적 단어(우울, 슬픔, 분노 등)와 시간적 요소(과거, 현재, 미래), 그리고 언어 스타일(동사, 품사 스타일)을 이용하여 데이터를 추출하였고, 이에 K-NN 분류기법을 적용하여 우울증을 탐지하고 분류성능을 평가하였다. 또한 Moldagulova와 Sulaiman (2017)은 2015-2016년 정부 뉴스 및 전자 정부 뉴스를 수집하여 K-NN 분류기법을 통해 텍스트 분류를 진행하였다. 높은 정확도를 보임을 입증하며 K-NN 분류기법이 효과적임을 보였다.

의사결정나무(Quinian, 1986)또한 대표적인 분류기법 중 하나로 분류과정을 나무구조로 시각화하는 특징으로 인해 널리 사용되고 있다. Dana와 Alashqur

(2014)는 의사결정나무를 이용하여 알츠하이머 여부를 진단하는 모델을 생성하였다. 5가지 속성인 성별, 연령, 유전적 원인, 뇌 손상 및 혈관 질환을 이용하여 데이터를 통해 의사결정나무를 구축하여 새로운 데이터에 대해 분류 예측을 하는 모델을 제시하였고 정보점수를 기반으로 결과에 영향을 많이 주는 속성에 대해 서술하였다. Gakii와 Jepkoech (2019)는 PH 농도, 알칼리도, 전도도, 색상을 이용하여 수질을 평가하는 의사결정나무를 형성하였다. 형성된 의사결정나무를 토대로 새로운 데이터에 대해 수질의 좋음과 나쁨을 분류하는 예측 모델을 제시하였고, 의사결정나무를 그래프 형태로 시각화하여 노드가 나뉘는 과정과 기준을 제시하여 결과에 대한 해석의 용이성을 나타내었다. 그리고 Kumar 등 (2012)은 기존의 네트워크 탐지 침입 시스템(IDS)에 의사결정나무 모델을 적용하여 기존에 탐지하지 못한 네트워크 침입을 판별하고자 하였다. KDD99 데이터를 이용하여 기존 네트워크 탐지 기록 데이터를 통해 의사결정나무를 형성하였다. 그리고 형성된 의사결정나무를 이용하여 새로운 데이터에 대해 네트워크 정상 유무를 분류 예측하였으며, 악성 네트워크와 양성 네트워크 간 차이점에 대해 분석적 통찰력을 제공하였다.

최근에는 의사결정나무의 앙상블모델인 랜덤포레스트(Breiman, 2001)와 서포트 벡터 머신(Vapnik, 1999)을 이용한 분류예측 연구가 활발히 이뤄지고 있다. Xuan 등 (2018)은 랜덤포레스트를 이용하여 신용카드 사기사건을 탐지하는 분류 모델을 형성하였다. 중국 전자상 거래 데이터에서 구매금액, 구매빈도, 구매시간 간격 등을 이용하여 비정상적인 소비패턴을 파악하여 사기를 탐지하였으며 높은 분류성능을 나타내었다. 또한 Sun 등 (2002)은 웹 페이지의 텍스트와 컨텍스트인 제목, 하이퍼링크를 이용하여 서포트 벡터머신을 Cornell, Texas, Washington, 그리고

Wisconsin 4개대학의 1997년 웹 페이지 데이터로 이루어져 있는 WebKB 데이터에 적용해 웹 분류를 진행하였다.

이와 같이 정형데이터와 비정형데이터, 그리고 다양한 분야에서 분류기법이 사용되고 있으며 이를 위한 머신러닝 기법 또한 다양하게 사용되며 개발되고 있다. 그 중 수집된 데이터의 사전정보를 이용하여 사후분포를 도출함으로써 분류를 진행하는 베이지안 분류기가 많이 사용되고 있다. 가장 널리 사용되는 베이지안 분류기 중 하나는 Duda와 Hart (1973), Langley 등 (1992)에 의해 처음 소개된 나이브 베이즈 분류기(naive Bayes classifier; 이하 나이브 베이즈)이다. 나이브 베이즈에서는 클래스 변수가 주어졌을 때 각 속성들이 조건부 독립이라는 강한 가정을 가지며, 이로 인해 데이터의 결합확률분포는 클래스 변수의 사전확률과 각 속성의 조건부 확률 분포의 곱으로 표현되며 사후분포는 이에 비례한다. 그 후 사후분포의 값을 가장 크게 만드는 클래스값으로 분류예측을 진행한다. 조건부 독립이라는 강한 가정에 의해 모형은 단순화될 수 있으며, 데이터의 수가 많아도 빠른 예측이 가능하다. 그러나 조건부 독립이라는 강한 가정은 실제 데이터에서는 비현실적이다. 예를 들어 고객의 대출을 심사하는 분류문제에서 고객의 소득, 연령, 재산의 상관관계를 무시하는 것은 직관적으로도 어긋나게 된다.

이러한 문제점을 해결하기 위해 Pearl (1988)은 베이지안 네트워크(Bayesian network)를 소개하였다. 베이지안 네트워크는 각 변수들에 대한 결합확률분포를 효과적으로 표현하는 방향성 비순환 그래프(directed acyclic graph; DAG)이다. 그래프에서 변수들은 노드로 표현되며, 변수간의 상관관계는 간선으로 표시된다. 선의 화살표 방향에 따라 부모노드(원인노드)와 자식노드(결과노드)가 결정되고 자식노드는 부모노드를 조건부로 가지게 되며, 변수간 선이 없는 경우는 독립임을

나타낸다. 형성된 네트워크에서 결합확률분포함수를 각 변수의 조건부확률식의 곱으로 나타내고 사후분포를 계산하여 분류를 진행한다. 나이브 베이즈 분류기 또한 베이지안 네트워크로 표현되며 모든 변수가 클래스 변수로부터 화살표를 받는 형태로 네트워크가 형성된다. 베이지안 네트워크는 변수간의 인과관계를 시각화하는 특징으로 인해 화이트박스(white box) 모델이라고 불리며, 다양한 분야에서 사용되고 있다.

하나의 결합확률분포함수는 사슬 법칙(chain rule)에 의해 다양한 형태로 표현되며 그에 따라 네트워크의 형태도 다양하게 표현된다. 따라서 가장 적합한 형태의 네트워크를 찾는 것이 중요하다. 변수간 관계가 과하게 연결된 네트워크는 모델이 복잡해지고 적합 시간이 오래 걸리며 과적합의 위험이 있다. 반대로 너무 간단하게 연결된 경우 변수간 관계가 제대로 표현되지 않으며 분류 성능이 떨어질 수 있다. Lam와 Bacchus (1994), Suzuki (1993)는 최적의 베이지안 네트워크를 설정하는 기준으로 MDL(minimum description length)을 사용하였다. MDL은 베이지안 네트워크가 데이터를 설명하는데에 필요한 정보량이라는 의미를 가지고 있으며, 값이 낮을수록 데이터에 적합한 네트워크를 형성하는 것을 의미한다.

Friedman 등 (1997)은 나이브 베이즈의 강한 독립성 가정을 완화하기 위해 노드 사이에 트리형태 네트워크를 가정한 TAN(Tree-Augmented Naive Bayes)를 제안하였다. 각 변수가 클래스 변수 이외에 최대 하나의 변수를 추가로 부모노드로 가지는 것을 허락함으로써 조건부 독립성 가정을 완화하였다. TAN은 cl-algorithm(Chow와 Liu, 1968)을 이용하여 네트워크 구조를 학습하고, 그에 맞는 모수인 CPT(conditional probability table)를 학습하는 과정을 거쳐 형성된다. 특히 Friedman 등 (1997)은 cl-algorithm을 통해 형성된 TAN의 네트워크가

최소 MDL임을 보임으로써 TAN의 네트워크가 최적의 네트워크임을 입증하였다. 따라서 TAN은 나이브 베이즈보다 높은 정확도를 보이고, 변수간 관계를 시각화할 수 있다는 장점으로 인해 변수간 관계를 분석하는 데에 많이 사용되고 있다.

김인철 (2002)은 나이브 베이즈와 TAN, 그리고 네트워크 구축에 제약을 주지 않는 일반적인 베이지안 네트워크를 불임환자들의 임상데이터 분석에 적용하였다. 이를 통해 임신 여부에 영향을 주는 요인들간 상호관계를 분석하여 직접적인 영향을 미치는 요인을 식별할 수 있었으며 분류 정확도가 K-NN, 의사결정나무모델보다 좋다는 것을 보였다. 그리고 김현미와 정성환 (2013)은 망막 임상데이터에 베이지안 네트워크를 적용하여 망막 질환과 그 요인과의 상호연관성을 분석한 후 연관 정도를 테이블을 이용하여 수치로 나타내고 망막 질환 발생 가능성을 높이는 요인을 밝혀내었다. Najafi와 Afsharchi (2012)은 컴퓨터 네트워크의 공격을 탐지하는 Intrusion Detection Systems(IDS)에 TAN을 적용하여 침입을 탐지하였으며 이를 의사결정나무와 서포트 벡터 머신과 비교하여 더 나은 분류성능을 보임을 확인하였다.

그러나 TAN의 네트워크 구조 학습 과정에서 인과관계에 포함되지 말아야 하는 변수가 네트워크 구조에 속하는 문제점이 발생한다. 예를 들어, 당뇨병 임상데이터를 통해 네트워크를 구축한 결과 성별과 고혈압이 연결되는 형태가 나타난 경우, 고혈압이 성별에 영향을 받는다는 해석을 할 수 있다. 그러나 성별의 경우 제어할 수 있는 변수가 아니기 때문에 인과관계에 포함시키고 해석하는 것은 자연스럽지 못하다고 볼 수 있다. 본 논문에서는 기존의 TAN과는 달리 제어할 수 없는 변수에 제약을 걸어 네트워크에 포함되지 않으면서 정보를 잃지 않게 예측변수에 포함시키는 개선된 TAN(New tree-augmented naïve Bayes; 이하

NTAN)을 제안한다. 시뮬레이션과 실제 데이터를 통해 NTAN과 TAN의 네트워크를 비교하고, 변수간 관계 해석이 용이해짐을 보인다. 또한 다양한 분류 성능 평가지표를 이용하여 NTAN이 TAN과 다른 머신러닝 기법들에 비해 우수한 정확도와 F1을 가지는 것을 보인다.

본 논문에서는 분류 종류인 다중 분류(Multiclass Classification)와 이진 분류(Binary Classification) 중에서 이진 분류 문제에 대해 연구를 진행한다. 2장에서는 베이지안 네트워크의 개념과 모수 학습과정을 살펴본 후 3장에서 본 논문에서 새롭게 제시하는 NTAN에 대해 소개한다. 4장에서는 분류성능을 평가할 다양한 지표들의 개념을 소개한다. 5장에서는 두가지 상황의 시뮬레이션을 통해 기존 방식의 TAN과 NTAN의 네트워크와 분류성능을 비교하는 실험을 진행한다. 그리고 6장에서는 실제 건강검진 데이터와 UCI 데이터를 이용하여 모델의 분류성능을 평가하고 변수 간의 인과관계를 시각화 하여 논문에서 제시하는 NTAN의 네트워크와 기존의 TAN의 네트워크를 비교한다. 7장에서는 결론 및 향후 연구를 논의한다.

제 2 장 베이지안 네트워크

2.1 베이지안 네트워크 모형학습

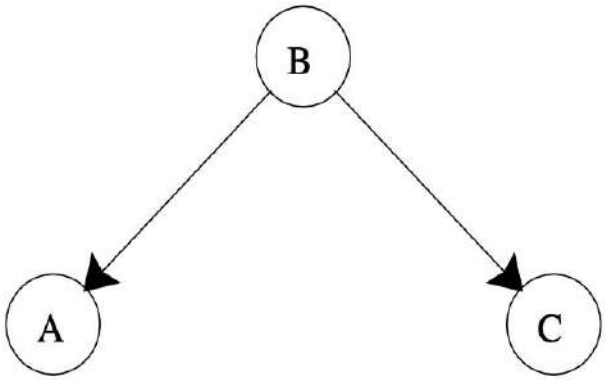


그림 1: 베이지안 네트워크

베이지안 네트워크는 방향성 비순환 그래프(DAG)라고도 불리며, 결합확률분포의 형태를 그래프로 나타낸 확률적 그래픽 모델이다. 베이지안 네트워크는 각 변수를 의미하는 노드(node)와 변수 사이를 잇는 호(arc)로 구성된다. 노드간 호가 연결되어 있는 경우 해당 노드간 인과관계가 존재하며, 호가 연결되어 있지 않은 경우 각 노드는 독립이다. 호에서 방향성은 화살표로 표시되며 화살표가 시작하는 쪽의 노드를 부모노드(parent node; 원인노드)라 하고 화살표가 끝나는 쪽의 노드를 자식노드(child node; 결과노드)라고 한다. 비순환성 그래프라는 것은 하나의 노드에서 출발하여 화살표의 방향을 따라갔을 때

자기자신으로 되돌아오지 않는 것을 의미한다. 일반적으로 베이지안 네트워크에서 결합확률분포는 사슬 법칙(chain rule)에 의해 다음과 같이 표현된다.

$$P(\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_\mathrm{N}) = \prod_{i=1}^{N} P\left(X_i \middle| Pa(X_i)\right).$$

이 때 $P(X_1, X_2, \ldots, X_N)$ 은 결합확률분포를 의미하고, $Pa(X_i)$ 는 X_i 의 부모노드 집합을 의미한다. $P(X_i|Pa(X_i))$는 X_i의 부모노드가 주어졌을 때의 X_i에 대한 조건부 확률분포이며 베이지안 네트워크에서 노드를 의미한다. 예를 들어, 그림 1에서 변수 B는 변수 A와 C의 부모노드이며, 반대로 A와 C는 B의 자식 노드가 된다. 그림 1에 나타난 네트워크를 통해 결합확률분포 $P(A, B, C) = P(A|B)P(B)P(C|B)$ 로 표현되며 $P(A|B), P(C|B)$는 조건부 확률분포를 의미한다.

베이지안 네트워크는 네트워크의 형태를 결정짓는 구조 학습과 그에 해당하는 모수학습 과정을 거쳐 학습된다. 특히 구조학습단계에서 네트워크는 다양한 형태로 표현되므로, 적합한 네트워크를 찾는 것이 중요하다. 본 논문에서는 베이지안 네트워크의 기본적인 형태인 나이브 베이즈와 나이브 베이즈의 가정을 완화한 TAN, 그리고 논문에서 새롭게 제안하는 NTAN의 구조와 모수 학습과정을 소개한다.

2.2 구조 학습

2.2.1 Naïve Bayes

나이브 베이즈는 베이지안 네트워크 모델 중 가장 간단한 모델로, 각 변수들이 조건부 독립이라는 강한 가정을 전제로 하는 모델이다. 즉 그림 2(왼쪽)와 같이 각 변수들은 클래스 변수를 부모노드로 가지며 다른 변수와는 연결되지 않는 형태이다.

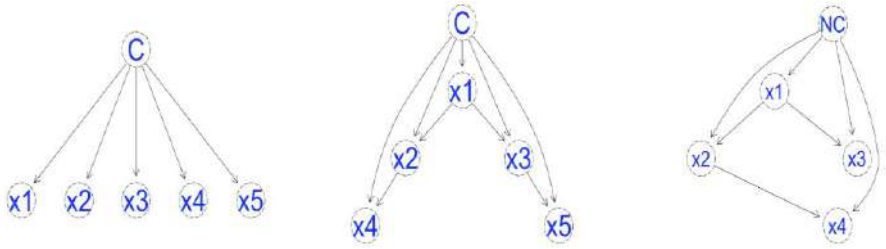


그림 2: 베이지안 네트워크 예시: 나이브 베이즈(왼쪽), TAN(가운데), NTAN(오른쪽)

따라서 변수 $(X_1, X_2, \ldots, X_N)$이 주어졌을 때 클래스 C의 사후 분포는 다음과 같이 표현된다.

$$P(C|X_1, X_2, \ldots, X_N) \propto P(C) \prod_{i=1}^{N} P(X_i|C).$$

나이브 베이즈는 강한 가정을 전제로 하는 모델임에도 불구하고 높은 정확도를 보여주기 때문에 많이 사용되는 모델이다. 그러나 실제 데이터에서 나이브 베이즈의 가정이 위배되는 경우가 많다는 문제점이 있다.

2.2.2 TAN (Tree-Augmented Naïve Bayes)

나이브 베이즈의 강한 가정을 보완하기 위해 각 변수들에 대해 클래스 변수를 제외한 최대 1개의 부모 노드를 더 허용하는 나무 구조 네트워크인 TAN (Friedman 등, 1997)이 고안되었다. 네트워크를 구축하는 방식에 따라 여러가지 모형이 존재하나 본 논문에서는 cl-algorithm(Chow와 Liu, 1968)을 이용하여

나무 구조 네트워크를 구성하는 TAN을 소개한다. cl-algorithm을 이용한 TAN 형성 과정은 다음과 같이 표현되는 **Algorithm 1**과 같다.

Algorithm 1: TAN algorithm(Chow 와 Liu, 1968)

1. 각 변수 쌍들에 대해 CMI(Conditional Mutual Information): $I_{\hat{P}_D}(X_i; X_j|C)$를 계산한다. $I_{\hat{P}_D}(X_i; X_j|C) = \sum_{x_i,x_j,c} \log \frac{P(x_i,x_j|c)}{P(x_i|c)P(x_j|c)}$, $i \neq j$
2. 모든 $X_1, \ldots, X_N$ 에 대해 완전한 무방향성 그래프를 형성한 후 계산한 $I_{\hat{P}_D}(X_i; X_j|C)$를 가중치로써 각 변수 쌍 사이 선(호)에 표시한다.
3. Maximum Weighted Spanning Tree를 형성한다.
4. 임의로 뿌리노드를 선택하고 모든 호의 방향을 바깥쪽으로 설정하여 무방향성 네트워크를 유방향성 네트워크로 변환한다.
5. 마지막으로 클래스 변수를 네트워크에 추가하여 각 변수로 방향이 향하도록 호를 추가한다.

Algorithm 1에서 Maximum Weighted Spanning Tree는 가중치의 합이 가장 큰 비순환 그래프이다. 완전한 무방향성 그래프로부터 가중치가 큰 순서대로 호를 선택하며, 이 때 선택된 호가 방향성 그래프를 유발한다면 선택하지 않고 그 다음으로 가중치가 높은 호를 선택한다. 노드의 개수가 N일때 N-1개의 호가 선택될 때까지 진행한다. Fridman 등 (1997)은 cl-algorithm을 이용하여 형성한 TAN의 네트워크가 MDL이 최소인 네트워크이며 일반적으로 나이브 베이즈보다 높은 정확도를 보임을 입증하였다. TAN은 그림 2(가운데)와 같이 변수간

인과관계를 네트워크 형태로 나타낼 수 있다는 장점이 있으며 이에 따른 사후 분포는 다음과 같이 표현된다.

$$P(C|X_1, X_2, \dots, X_N) \propto P(C) \prod_{i=1}^{N} P(X_i|X_j, C), \quad i \neq j.$$

이 때 X_i의 부모 노드가 클래스 변수 뿐이라면 $X_j = \emptyset$ 이다.

2.3 모수 학습

베이지안 네트워크에서 각 노드는 부모노드를 조건부로 하는 CPT(Conditional Probability Table)을 가진다. CPT는 구성요소로 조건부확률을 가지며 이를 모수로 가정한다. 모수 학습방법에는 최대가능도추정법과 베이지안방법 두 가지가 존재한다(Stephenson, 2000). 본 논문에서는 베이지안 네트워크 모수 학습에 일반적으로 사용되는 베이지안방법을 사용한다. 하나의 변수 X_i의 부모노드가 j의 값을 가지는 경우 X_i가 클래스 k의 값을 가지는 확률 모수 θ_{ijk}를 다음과 같이 나타낸다.

$$P(X_i = k|Pa(X_i) = j) = \theta_{ijk},\ i = 1, \dots, N,\ j = 1, \dots, q_i,\ k = 1, \dots, r_i.$$

이 때 N은 변수 X의 개수, q_i은 부모노드의, r_i는 클래스 변수의 카디널리티(cardinality)를 의미한다. 베이지안 모수 추정방법을 사용하기 위해 θ_{ijk}에 사전분포를 가정하여 추정을 진행한다. TAN은 기본적으로 이산형 변수를 입력 변수로 받으므로, 입력 변수의 확률분포를 다항분포로 가정한다. 또한 사전확률 θ_{ijk}의 사전분포를 공액사전분포인 Dirichlet분포로 가정하여 사후분포와

사전분포가 같은 분포족이 되도록 한다. 즉, 네트워크 구조 S가 주어져있을 때 모수의 사전확률분포를 다음과 같이 정의한다.

$$(\theta_{ij1}, \theta_{ij2}, \ldots, \theta_{ijr_i})|S \sim \text{Dirichlet}(\alpha_{ij1}, \alpha_{ij2}, \ldots, \alpha_{ijr_i}),\ \alpha_{ijk} > 0.$$

이 때 α_{ijk}는 Dirichlet hyperparameter를 나타낸다. 또한 주어진 데이터 D에 대해 다항분포를 가정하였으며 사전분포가 공액사전분포인 Dirichlet분포이기 때문에 사후분포 또한 Dirichlet분포를 따른다. 즉 데이터 D와 네트워크 구조 S가 주어진 경우 모수의 사후분포는,

$$(\theta_{ij1}, \theta_{ij2}, \ldots, \theta_{ijr_i})|S, D \sim \text{Dirichlet}(\alpha_{ij1} + N_{ij1}, \alpha_{ij2} + N_{ij2}, \ldots, \alpha_{ijr_i} + N_{ijr_i}).$$

로 표현된다. 이 때 N_{ijk}는 X_i의 부모노드가 j의 값을 가지는 경우 X_i가 클래스 k의 값을 가지는 표본의 개수이다. 따라서 모수 추정치인 사후평균 $\hat{\theta}_{ijk}$는,

$$\hat{\theta}_{ijk} = \frac{N_{ijk} + \alpha_{ijk}}{N_{ij} + \alpha_{ij}}\ .$$

로 표현되며 이 때 $N_{ij} = \sum_k N_{ijk}$, $\alpha_{ij} = \sum_k \alpha_{ijk}$이다. 만일 모든 $\alpha_{ijk} = 0$인 경우, 최대가능도추정량과 같은 값을 가지게 된다. 본 논문에서는 사전분포의 non-informative prior를 가정하기 위해 모든 α_{ijk}의 값을 1로 설정하여 실험을 진행한다. 또한 변수 간의 관계를 구성하고 네트워크로 시각화하기 위해 R 패키지 'bnclassify'와 'Rgraphviz'를 사용한다(Mihaljevic 등, 2018; Hansen 등, 2022).

2.4 머신러닝 기법

본 논문에서 모델의 분류성능을 비교하기 위해 통상적으로 많이 사용하는 머신러닝 모델인 K-NN기법, 의사결정나무, 랜덤포레스트 모델을 사용한다.

2.4.1 K-NN (K-Nearest Neighbor)

K-NN은 분류할 데이터 근처 K개의 데이터 중 가장 많은 데이터가 속해 있는 클래스로 분류하는 기법이다(Cover 등, 1967). 데이터에 대한 분포 가정이 필요하지 않은 비모수적 방법으로 사용이 간단하고 직관적임에도 높은 정확도를 보이기 때문에 널리 사용되고 있다. 그러나 K값에 대한 종속성이 높고 최적의 K값을 찾는 것은 어렵다. 또한 표본의 수가 많은 경우 계산량이 증가하여 모델 적합 시간이 오래 걸린다는 단점이 있다. 또한 학습 후 모델이 생성되지 않아 새로운 데이터가 추가될 때마다 새롭게 모델을 적합 시켜야 하는 lazy model이라 불린다.

K-NN기법을 사용하기 위해 데이터간 거리를 측정해야 한다. 일반적으로 유클리드 거리(Euclidean Distance), 맨해튼 거리(Manhattan Distance)를 사용하지만, 연속형 자료에 적용되는 척도들이기 때문에 이산형 자료를 취급하는 본 논문에서는 사용될 수 없다. 따라서 이산형 자료의 거리 계산을 위해 다음과 같이 정의된 해밍 거리(Hamming Distance)를 이용한다.

$$D_H = \sum_{j=1}^{J} I(x_j \neq y_j).$$

이 때 J는 클래스 변수를 제외한 변수의 총 개수이다. 즉 해밍 거리 D_H는 표본 x와 y를 비교하였을 때 다른 값을 가지는 변수의 총 개수를 의미한다. 새로운 표본 z의 클래스는 표본 집합에서 z와 해밍 거리 D_H가 가장 가까운 k개의 표본 중 가장 많은 표본이 속한 클래스로 분류예측을 진행한다.

2.4.2 의사결정나무 (Decision Tree)

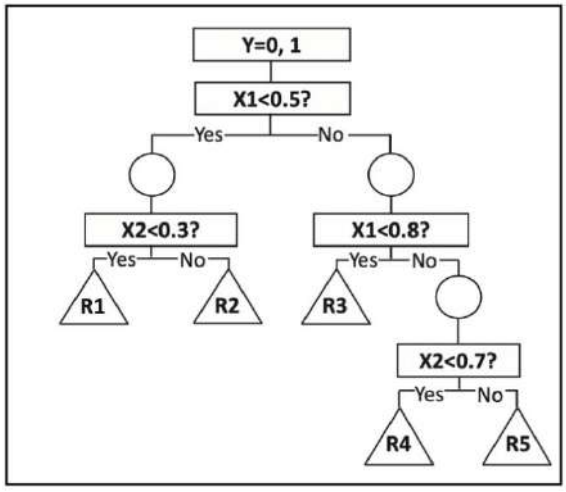


그림 3: 이진분류 의사결정나무의 간단한 예시

의사결정나무 (Decision Tree)는 의사결정 규칙을 나무 구조로 나타내어 전체 자료를 몇 개의 소집단으로 분류하거나 예측하기 위해 일반적으로 사용하는 데이터마이닝 기법이다(Quinian, 1986). 그림 3과 같이 형성되며 적합 속도가 빠르고 이산형 자료와 연속형 자료 모두 다룰 수 있으며 나무 구조를 통해 결과에 대한 해석이 용이하다는 장점으로 인해 많이 사용되지만 모델분산이 높아 과적합(over fitting)의 위험이 높다는 단점이 있다. 의사결정나무를 학습할 때 일정한 정지규칙(stopping)을 통해 나무의 과성장을 방지하고 이미 학습된 나무 구조를 가지치기(pruning)과정을 거쳐 과적합이 되는 것을 방지한다.

CART(Classification And Regression Trees)는 대표적인 의사결정나무 모델로 각 노드에서 다음 노드로 분리될 때 지니계수(Gini index)를 기준으로 이진으로 분류되는 모델이다(Breiman 등, 1984). 입력값으로 이산형 자료와 연속형 자료 모두 다룰 수 있으며, 예측과 분류작업 모두 가능하다. 지니계수는 다음과 같이

표현되며, 지니계수가 낮을수록 불순도가 낮음을 의미하므로 지니계수가 낮은 변수를 분리 기준으로 잡아 노드를 분리시킨다.

$$G = 1 - \sum_{i}^{k} p_i^2.$$

이 때 k는 그룹의 클래스의 개수를 나타낸다. 예를 들어 나이를 기준으로 나누었을 때 Yes: 3, No: 4로 나누어지고, 성별을 기준으로 Yes: 2, No: 5로 나누어질 때 지니계수는 각각 $G(\text{age}) = 1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2 = \frac{24}{49}$, $G(\text{sex}) = 1 - \left(\frac{2}{7}\right)^2 - \left(\frac{5}{7}\right)^2 = \frac{20}{49}$ 이다. 성별을 기준으로 하였을 때 지니계수가 더 낮으므로 성별을 기준으로 노드를 나눈다.

또 다른 의사결정나무 모델로는 노드 분류시 엔트로피, information gain을 사용하여 노드를 다중으로 분류하는 C4.5(Quinlan, 1992), 이산형 자료를 취급하며 노드 분류시 카이제곱 통계량을 기준으로 다중 분류를 시행하는 CHAID(Kass, 1980) 등이 있다. 본 논문에서는 각 노드에서 분리될 때 지니계수(Gini index)를 기준으로 이진으로 분리되는 CART기법을 사용하며, R 패키지 'rpart'를 사용한다(Therneau 등, 2015).

2.4.3 랜덤포레스트 (Random Forest)

의사결정나무는 적합이 빠르고 성능이 좋을 뿐 아니라 결과 도출 과정을 나무구조로 표현하기 때문에 결과 해석이 가능하다. 그러나 의사결정나무는 모델분산이 높아 과적합의 위험이 매우 높다는 것이 가장 큰 단점이다. 랜덤포레스트(Random forest)는 이와 같은 의사결정나무의 과적합문제를 해결하기

위해 고안된 앙상블 기법이다(Breiman, 2001). 랜덤포레스트는 결측치가 있는 경우에도 성능이 좋으며 과적합의 위험이 적고 변수의 중요도를 알 수 있다는 장점이 있다. 하지만 데이터 수가 많은 경우 모델 적합에 오랜 시간이 걸리며 결과 해석이 어렵다는 단점이 있다. 랜덤포레스트 알고리즘은 다음 Algorithm 2와 같다.

Algorithm 2: Random forest algorithm (Liaw 와 Wiener, 2002)

1. 기존 데이터에서 n_{tree}개의 부트스트랩 표본을 생성한다.
2. 각 부트스트랩 표본에서 가지치기를 시행하지 않은 의사결정나무를 생성한다. 이 때, 노드에서 분리기준으로 사용되는 변수는 랜덤하게 정하며 개수는 분류의 경우 $\sqrt{p}$개로, 회귀의 경우 $\frac{p}{3}$개로 설정한다. 이 때 p는 설명변수의 개수이다.
3. 생성된 n_{tree}개의 의사결정 나무를 이용하여 새로운 데이터에 대해 예측을 진행한다. 분류의 경우에는 각 의사결정나무에서 예측한 값중 가장 많이 나온 클래스 값으로 예측하고, 회귀의 경우 예측값의 평균을 사용한다.

기존 데이터에서 부트스트랩 샘플을 생성할 때, 데이터의 약 $\frac{1}{3}$이 부트스트랩 샘플에 포함되지 않는다. 이를 OOB(Out-Of-Bagging)라고 하며 이 데이터를 이용하여 에러율을 계산할 수 있다. 각 부트스트랩 샘플을 생성하여 의사결정나무를 만들고 이에 OOB를 이용하여 예측을 진행하여 Algorithm 2에서와 같은 방식으로 예측을 진행한 뒤 에러율을 계산한다. OOB를 이용한 예측은 꽤나 정확하다고 알려져 있으며, 이를 평가 데이터로 사용하는 경우도 있다. 본 논문에서

랜덤포레스트를 구현하기위해 R 패키지 'randomForest'를 사용하였다(Liaw와 Wiener, 2018).

제 3 장 개선된 TAN (New TAN; NTAN)

기존의 TAN에서 네트워크를 구축할 때 모든 변수를 이용하기 때문에 네트워크에 포함되지 않아야 할 변수가 들어가고 부적절한 관계해석을 유발하는 문제점이 발생한다. 본 논문에서는 이와 같은 문제점을 해결하기 위해 새로운 방식의 TAN(New TAN; NTAN)을 제시한다. TAN의 경우 모든 변수 쌍에 대해 CMI(Conditional Mutual Information)를 계산하여 네트워크를 형성하기 때문에 인과관계에 포함되지 않아야 할 변수들이 네트워크에 포함되는 구조로 학습이 되는 문제점이 발생한다. 이러한 현상을 막기위해 먼저 구조 학습에 포함시키지 않을 변수들을 선정하고 CMI를 계산할 때 해당 변수들이 포함되는 경우 네트워크에 포함되지 않도록 제약을 부여한다. 그러나 이 경우 해당 변수들과 다른 변수와의 상관관계를 사용할 수 없다는 단점이 존재한다. 따라서 해당 변수들을 클래스 변수에 포함시켜 정보를 보존하는 방식을 사용한다. 이 경우 2개이던 클래스 변수의 카테고리가 더 많은 카테고리로 확장된다. 예를 들어 클래스 변수 C에 2개의 카테고리가 있고 제외할 변수집합 $\boldsymbol{E}$에 6개의 카테고리가 있는 경우, 새로운 클래스 변수에는 $2 \times 6 = 12$개의 카테고리가 존재한다.

또한 기존의 TAN에서는 뿌리노드를 임의로 선정하였으나 NTAN에서는 Jiang 등 (2005)이 제안한 방법인 클래스 변수와 MI(Mutual Information)가 가장 높은 변수를 뿌리노드로 선정하는 방식을 응용한다. 새로운 클래스 변수에는 제외할 변수집합 $\mathbf{E}$가 포함되어 있으므로, 변수집합 $\boldsymbol{E}$와 관계가 높은 변수가 뿌리노드로

선정되는 것을 방지하고 적합한 네트워크를 형성하기 위해 기존의 클래스 변수와의 MI가 높은 변수를 뿌리노드로 선정한다. 그 이후의 네트워크 형성과정은 기존의 TAN과 유사하다. NTAN을 형성하는 과정은 다음과 같다.

Algorithm 3: NTAN algorithm

1. 네트워크에서 제외시킬 변수집합 E를 선택한다.
2. 해당 변수들을 클래스 변수에 포함시켜 새로운 클래스 변수(NC)를 생성한다.
3. 각 변수 쌍들에 대해 CMI(Conditional Mutual Information): $I_{\hat{P}_D}(X_i; X_j|NC)$를 계산한다. 이 때 변수가 E에 포함되는 경우 CMI를 0으로 설정한다.
4. 모든 $X_1, \dots, X_N$ 에 대해 완전한 무방향성 그래프를 형성한 후 계산한 $I_{\hat{P}_D}(X_i; X_j|NC)$를 가중치로써 각 변수 쌍 사이 선(호)에 표시한다.
5. Maximum Weighted Spanning Tree를 형성한다.
6. 기존의 클래스 변수와의 MI: $I_{\hat{P}_D}(X_i; C)$ 가 가장 높은 변수를 뿌리 노드로 선택하고 모든 호의 방향을 바깥쪽으로 설정하여 무방향성 네트워크를 유방향성 네트워크로 변환한다. $\left(I_{\hat{P}_D}(X_i; C) = \sum_{x_i,c} P(x_i, c)\log\frac{P(x_i,c)}{P(x_i)p(c)}\right)$
7. 마지막으로 새로운 클래스 변수를 네트워크에 추가하고 각 변수로 방향이 향하도록 호를 추가한다.

그림 2에서 제외시킬 변수를 X_5라고 선정한 경우, NTAN은 그림 2의 오른쪽과 같이 표현된다. NTAN의 사후분포는 다음과 같이 표현된다.

$$P(NC|X_1, X_2, \dots, X_N) \propto \mathrm{P}(NC) \prod_{X_1,\dots,X_N \,\not\subset\, E} P\left(X_i|X_j, NC\right),\ i \neq j.$$

이 때 NC는 제외할 변수집합 E 가 포함된 새로운 클래스 변수이며, X_i 의 부모노드가 새로운 클래스 변수(NC)뿐이라면 $X_j = \emptyset$이다.

Algorithm 3 과정을 통해 학습된 네트워크 구조를 이용하여 분류를 진행하는 경우 이진 분류 문제에서 다중 분류(Multiclass Classification) 문제로 확장되게 된다. 이 경우 클래스 변수를 알맞게 예측해도 제외할 변수집합 E 를 알맞게 예측하지 못하면 분류성능이 떨어지게 된다. 기존에 예측하고자 하는 변수는 클래스 변수이기 때문에, 이 경우 분류성능을 정확하게 측정할 수 없다. 이러한 문제를 해결하기 위해 다중 분류로 예측한 값에서 클래스 값만 추출하여 사용한다. 분류과정은 다음과 같다.

Algorithm 4: NTAN classification process

1. 기존의 클래스 변수 C에 제외할 변수집합 E를 포함시킨 새로운 클래스 변수 NC를 생성하여 이진 분류 문제를 다중 분류 문제로 확장시킨다.
2. Algorithm 3을 통해 학습한 네트워크를 기반으로 다중 분류를 실시한다.
3. 예측한 다중 분류값을 이진 분류값으로 축소한다.

예를 들어 클래스 변수가 당뇨병 유무이고, 제외할 변수집합 E에 나이와 성별이 포함된다고 하면, 나이/성별/당뇨병 유무가 합쳐진 새로운 클래스 변수의 경우 예측값 중 나이와 성별은 무시하고 당뇨병 유무를 사용하여 당뇨병 유무만 판별한다. 즉 다중 분류값이 '0'/'0'/'1' 이라고 예측된 경우 '1'로 취급한다. 이후의 성능 평가 과정은 이진 분류 문제와 동일하게 진행한다.

제 4 장 분류모델 성능 평가지표

4.1 혼동행렬 (Confusion Matrix)

혼동행렬은 분류모델을 적합하였을 때 생성되는 예측값과 실제값의 교차표로 분류모델의 성능을 확인할 수 있는 지표이며 다음과 같이 표현된다.

		Predict	
		Positive (1)	Negative (0)
Actual	Positive (1)	True Positive (TP)	False Negative (FN)
	Negative (0)	False Positive (FP)	True Negative (TN)

표 1: 혼동행렬

표 1에서 각 열은 예측된 클래스의 인스턴스를 나타내며 각 행은 실제 클래스의 인스턴스를 나타낸다. 또한 1은 Positive를 나타내며 0은 Negative를 나타낸다. 혼동행렬에서 1이라고 예측하였을 때 실제로 1인 경우를 True Positive(TP), 0이라 예측하였을 때 실제로 0인 경우를 True Negative(TN), 1이라고 예측하였으나 실제로는 0인 경우를 False Positive(FP), 0이라고 예측하였으나 실제로는 1인 경우를 False Negative(FN) 라고 한다. 이 4가지 지표를 이용하여

분류모델의 성능을 평가하는데 사용되는 다양한 지표를 계산할 수 있으며 상황에 따라 필요한 분류성능 평가지표를 사용한다(홍종선 등, 2022).

4.2 분류성능 평가지표

혼동행렬을 통해 계산할 수 있는 분류성능 평가지표중 본 논문에서는 최근 통상적으로 많이 사용되는 분류성능 평가지표인 정확도, 특이도, 정밀도, 재현율, F1 점수, AUC, 매튜 상관계수(Matthews Correlation Coefficient; MCC)를 사용한다. 각 분류성능 평가지표는 1에 가까울수록 분류모델의 성능이 좋음을 의미하고 0에 가까울수록 성능이 좋지 않음을 의미한다. MCC의 경우 1은 완전일치, 0은 랜덤일치, −1은 완전불일치를 의미한다. 또한 네트워크가 실제 변수간 인과관계를 잘 표현하는지 평가하기 위해 네트워크 유사도(Similarity Measure; SM)를 사용한다. 평가지표들의 산출방식과 의미는 아래 표 2와 같다. 표 2에 표기된 분류성능 평가지표들을 이용하여 앞서 설명한 NTAN과 기존방식의 TAN, 나이브 베이즈, K−NN, 의사결정나무, 랜덤포레스트의 분류성능을 비교한다. 먼저 표본의 수와 변수의 개수에 따른 두가지 시뮬레이션을 통해 NTAN과 기존방식인 TAN의 분류성능과 네트워크 유사도를 평가하고 실제 데이터를 통해 머신러닝 기법들과의 분류성능을 비교한다.

성능평가지표	산출식	설명
정확도 (ACC)	$\frac{TP+TN}{TP+TN+FP+FN}$	발생할 수 있는 전체 경우에서 올바르게 예측한 비율.
정밀도 (Precision)	$\frac{TP}{TP+FP}$	분류모델이 Positive 라고 분류한 것 중에서 실제로 Positive 인 것의 비율.
재현율 (Recall)	$\frac{TP}{TP+FN}$	실제값이 Positive 인 것 중에서 분류모델이 Positive 라고 예측한 것의 비율로 민감도와 같다.
특이도 (Specificity)	$\frac{TN}{TN+FP}$	실제값이 Negative 인 것 중 분류모델이 Negative 라고 예측한 것의 비율.
F1 점수	$\frac{2\times Precision\times Recall}{Precision+Recall}$	정밀도와 재현율의 조화평균으로 불균형 데이터의 분류성능을 평가하는데에 주로 사용된다.
AUC	ROC 커브의 밑면적	ROC 커브는 x축이 (1−특이도), y축이 민감도로 구성된 그래프로 분류기준인 임계값을 연속적으로 바꾸며 그래프 값의 변화를 나타낸다.
매튜 상관계수 (MCC)	$\frac{\frac{TP}{N}-S\times P}{\sqrt{SP(1-S)(1-P)}}$ *	실제값 벡터와 예측값 벡터의 유사도를 나타내며 -1 에서 1 의 값을 가진다.
SM	$\frac{\text{correct \# of arcs}^{\ddagger}}{\text{Model's Total \# of arcs}}$	네트워크에서 변수간 실제 인과관계를 표현한 비율.

표 2: 성능평가지표

* $N=TP+TN+FP+FN$, $S=\frac{TP+FN}{N}$, $P=\frac{TP+FP}{N}$

‡ arcs는 호를 의미하고, #은 number를 의미한다.

제 5 장 시뮬레이션

2장에서 소개한 베이지안 네트워크 기법들과 3장에서 소개한 NTAN의 성능을 평가하기 위해서 2가지 상황에 따른 시뮬레이션을 진행한다. 첫 번째로, 변수의 개수가 고정되어 있는 경우 표본의 수에 따른 분류성능을 평가한다. 표본의 수가 적은 경우부터 수를 점점 늘려 표본의 수가 많은 경우까지 모델이 잘 적합하는지 확인한다. 두 번째로, 표본의 수가 충분할 때 변수의 개수에 따른 모델의 분류성능을 확인하여 변수의 개수가 많은 경우에도 모델이 잘 적합하는지 확인하고자 한다.

구체적으로 1) 기존 방식의 TAN; 2) 선택된 변수를 제외하고 기존의 클래스 변수는 그대로 사용하는 TAN-II; 그리고 3) 본 논문에서 새롭게 제시하는 NTAN을 포함하여 총 3가지 종류의 TAN을 비교한다. 또한 기존 방식의 NB(Naive Bayes)와 본 논문에서 제시하는 방식을 적용시킨 NNB(New Naive Bayes)를 추가로 비교한다. NNB의 구축 과정은 NTAN의 네트워크 구축 과정인 **Algorithm 3**에서 과정 2까지 동일하게 진행하며 그 이후의 네트워크 구축 과정은 기존의 나이브 베이즈 구축 과정과 동일하다. NNB의 네트워크에는 제외할 변수집합 E가 클래스 변수에 포함된 나이브 베이즈 네트워크 형태를 가진다. 이후 분류 과정은 NTAN과 동일하게 진행한다.

시뮬레이션을 진행하는동안 전체 데이터를 10-fold하여 훈련 데이터와 평가 데이터를 9:1의 비율로 분할하여 진행한다. 변수 Y는 예측하고자 하는 클래스

변수이고 제외할 변수집합 E는 (sex, age)로 지정하며 변수 NC는 변수 Y에 E가 포함된 새로운 클래스 변수를 의미한다. 변수 생성 시 연속형 변수는 정규분포를, 이산형 변수는 이항분포를 이용하였으며 logit link와 probit link를 이용하여 변수를 생성하였다.

$$\text{probit}(x): \quad P(Y = 1|X) = \Phi(X^T\beta),$$

$$\text{logit}(x): \quad P(Y = 1|X) = \frac{e^{X^T\beta}}{1 + e^{X^T\beta}}.$$

그리고 TAN은 기본적으로 이산형 변수를 입력값으로 받기 때문에 연속형 변수는 4분위수를 기준으로 4개의 카테고리로 범주화하였다. 또한 TAN과 NB를 구현하기위해 R 패키지 'bnclassify'의 bnc함수를 사용하였다. 또한 사전분포의 non-informative prior를 가정하기 위해 smoothing 값(α)를 1로 설정하여 분석을 진행하였다. 그리고 구축된 네트워크를 그래프로 구현하기 위해 R 패키지 'Rgraphviz'의 plot함수를 사용하였다. 4장에서 소개한 분류성능 평가지표를 이용하여 모델들의 분류성능을 비교하였으며 시뮬레이션 II에서 변수의 개수가 40개인 경우와 70개인 경우를 제외한 각 상황마다 300번 반복 시행 후 나온 분류성능 평가지표의 평균을 내어 결과를 산출하였다. 시뮬레이션 II에서 변수의 개수가 40개인 경우와 70개인 경우에는 100번 반복 시행 후 나온 분류성능 평가지표의 평균을 내어 결과를 산출하였다. 또한 2가지 상황에 따른 시뮬레이션을 진행하면서 네트워크 유사도(Similarity Measure; SM)를 통해 TAN, TAN-II, NTAN을 적합하여 생성된 네트워크가 실제 변수간의 인과관계를 얼마나 잘 나타내는지 평가한다

5.1 시뮬레이션 I

첫 번째 시뮬레이션은 표본의 수에 따른 성능을 비교하기 위해 진행한다. 표본의 수가 N=100, 300, 500, 1,000, 5,000, 10,000, 50,000인 경우까지 총 7개의 상황을 비교한다. 변수의 개수는 10개로 고정하였고 아래와 같이 생성된 데이터를 이용하여 TAN, TAN-II, NTAN, NB, NNB를 구축하였다. 이 때 B는 베르누이 분포를 의미하고, P 는 사건이 발생할 확률을 의미한다. 또한 N은 정규분포를 의미하며 μ는 평균, sd는 표준편차를 의미한다. 제외할 변수집합 E는 (sex, age) 총 두 가지 변수로 설정하였다. 그리고 Y 변 수에 제외할 변수집합 E를 포함시켜 NC 변수를 생성한 후 NTAN과 NNB를 구축하고 분류예측을 진행하였다. TAN과 NB는 변수집합 E를 네트워크에서 제외하지 않고 Y 변수를 예측할 클래스 변수로 사용하여 분류예측을 진행하였다. 그리고 TAN-II는 변수집합 E 를 네트워크에서 제외하고 Y 변수를 예측할 클래스 변수로 사용하여 분류예측을 진행하였다. 그림 4는 데이터를 생성할 때 의도한 변수집합 E 가 제외된 실제 인과관계를 네트워크 형태로 시각화한 것이다. 변수간 선(arc)은 변수들간 관계가 있음을 의미하고 화살표가 출발하는 쪽이 원인 노드이고 방향이 도착하는 쪽이 결과노드가 된다. 예를 들어 x_3의 경우 x_1의 결과 노드이며, x_6의 원인노드이다. 따라서 x_1, x_3 그리고 x_6 간의 관계는 $x_3|x_1$, $x_6|x_3$로 표현된다. 그림 4와 구축된 TAN, TAN-II, NTAN의 네트워크를 비교하여 네트워크가 변수간의 실제 인과관계를 잘 표현하는지 유사도를 이용하여 평가한다.

$$Y \sim B(P = 0.5), \quad \text{sex} \sim B(P = 0.5),$$

$$x_1 \left| Y \sim B\left(P = \frac{\exp(-1.5 + 3 \times Y)}{1 + \exp(-1.5 + 3 \times Y)}\right), \quad x_2 \right| x_1 \sim B\big(P = \Phi(-1 + 2 \times x_1)\big),$$

$$x_3 | x_1 \sim N(\mu = 10 + 5 \times x_1,\ \text{sd} = 2), \quad \text{age} | x_2 \sim B\left(P = \frac{\exp(-1 + 2 \times x_2)}{1 + \exp(-1 + 2 \times x_2)}\right),$$

$$x_4 | x_2 \sim N(\mu = 50 + 15 \times x_2,\ \text{sd} = 3),$$

$$x_5 | \text{age}, x_2 \sim B\left(P = \frac{\exp(1.5 - 3 \times \text{age} - x_2)}{1 + \exp(1.5 - 3 \times \text{age} - x_2)}\right),$$

$$x_6 | x_3 \sim N(\mu = 100 + 1.5 \times x_3,\ \text{sd} = 5), x_7 | \text{age}, x_2 \sim B\big(P = \Phi(1 - 2 \times \text{age} - 0.5 \times x_2)\big).$$

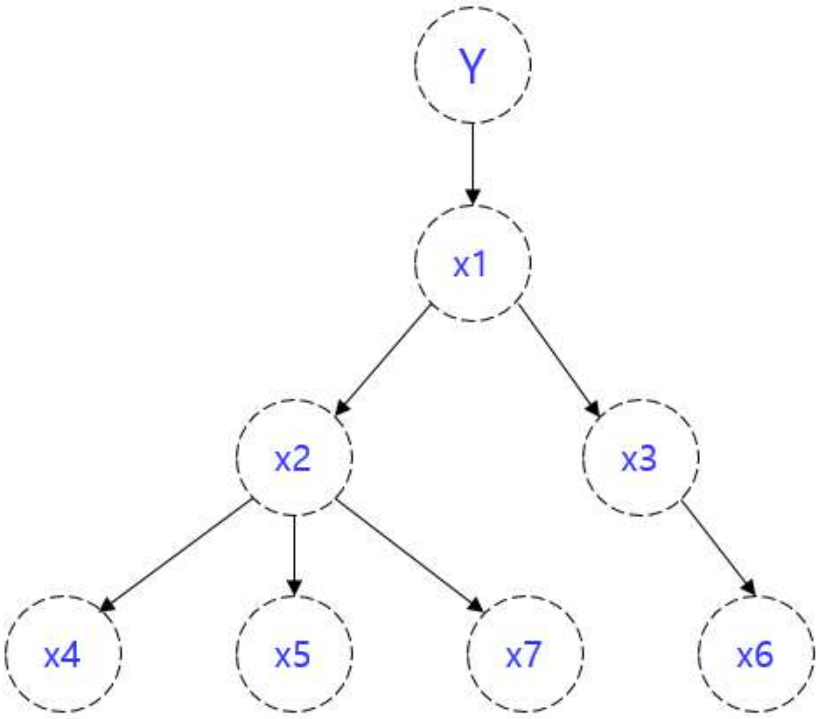


그림 4: 시뮬레이션 I 변수관계[§]

[§] 변수집합 E에 해당하는 sex와 age가 제외된 실제 인과관계를 표현한 네트워크.

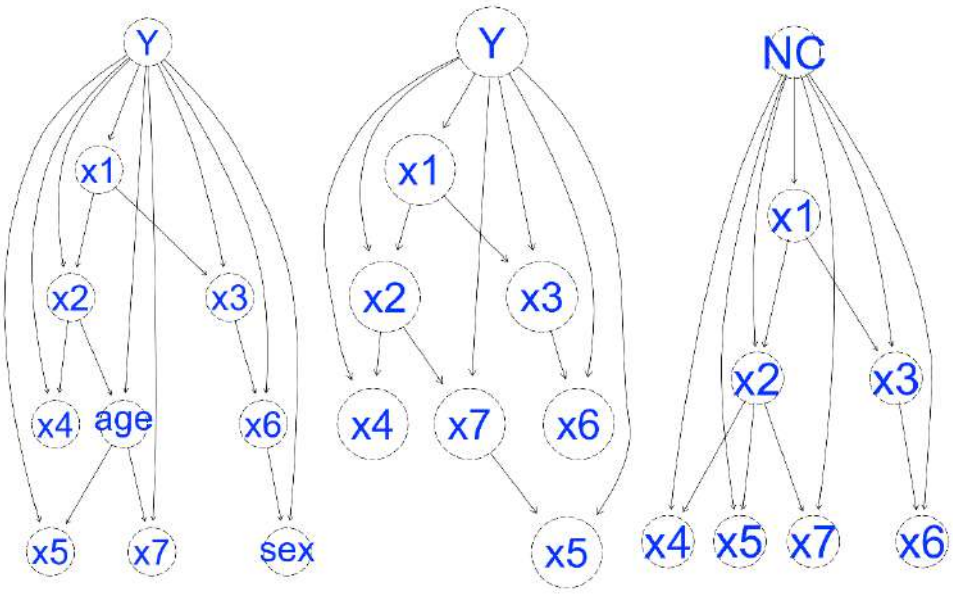


그림 5: TAN(왼쪽), TAN-II(가운데), NTAN(오른쪽)*

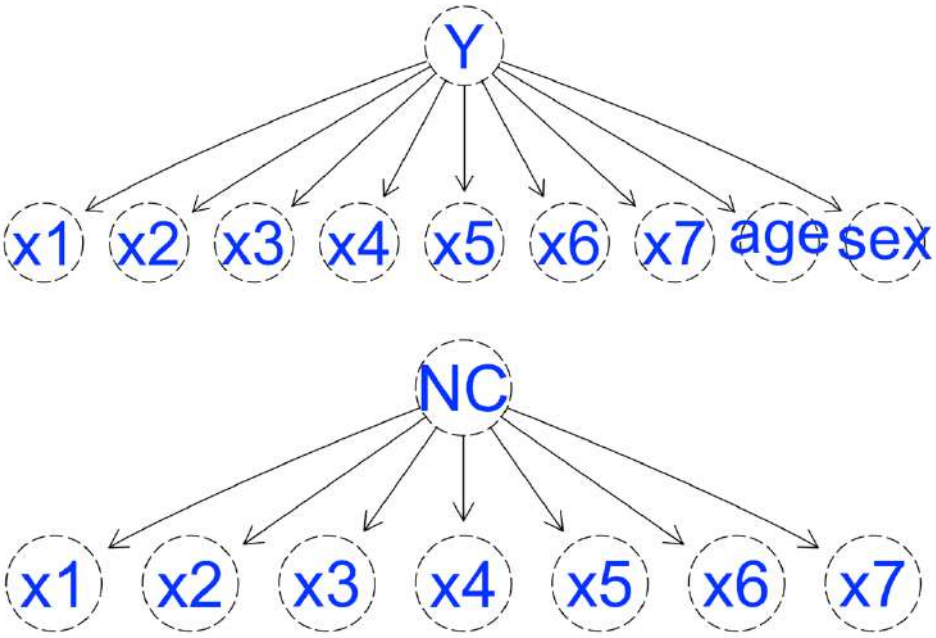


그림 6: NB(위), NNB(아래)

* 표본의 수 N=50,000인 경우의 네트워크.

N = 100								
	ACC	AUC	Recall	Spec	Prec	F1	MCC	SM
TAN	0.757 (0.08)*	0.788 (0.09)	0.75 (0.12)	0.763 (0.12)	0.762 (0.11)	0.749 (0.09)	0.52 (0.16)	0.24 (0.07)
TAN-II	0.761 (0.08)	0.797 (0.08)	0.753 (0.11)	0.769 (0.12)	0.766 (0.12)	0.753 (0.09)	0.53 (0.16)	0.326 (0.01)
NTAN	0.763† (0.09)	0.807 (0.09)	0.757 (0.13)	0.769 (0.13)	0.767 (0.12)	0.756 (0.1)	0.523 (0.17)	0.264 (0.11)
NB	0.778 (0.07)	0.808 (0.08)	0.775 (0.11)	0.779 (0.11)	0.779 (0.11)	0.771 (0.09)	0.549 (0.15)	0
NNB	0.781‡ (0.08)	0.813 (0.09)	0.781 (0.18)	0.78 (0.18)	0.781 (0.18)	0.775 (0.09)	0.549 (0.15)	0
N = 300								
	ACC	AUC	Recall	Spec	Prec	F1	MCC	SM
TAN	0.791 (0.08)	0.81 (0.09)	0.796 (0.11)	0.785 (0.11)	0.785 (0.1)	0.788 (0.08)	0.602 (0.15)	0.324 (0.12)
TAN-II	0.795 (0.07)	0.811 (0.09)	0.8 (0.1)	0.791 (0.11)	0.79 (0.1)	0.793 (0.08)	0.609 (0.15)	0.46 (0.18)
NTAN	0.787 (0.07)	0.812 (0.08)	0.789 (0.1)	0.786 (0.11)	0.784 (0.1)	0.784 (0.08)	0.591 (0.15)	0.333 (0)
NB	0.782 (0.07)	0.81 (0.08)	0.783 (0.11)	0.781 (0.1)	0.778 (0.1)	0.778 (0.08)	0.576 (0.15)	0
NNB	0.792 (0.08)	0.811 (0.08)	0.794 (0.1)	0.79 (0.11)	0.788 (0.1)	0.789 (0.08)	0.597 (0.15)	0

표 3: 표본의 수에 따른 분류성능 비교

* 괄호() 안에 표시된 수는 표준편차를 의미한다.

† 빨간색으로 표시된 수는 TAN, TAN-II, NTAN중 해당 평가지표가 가장 높은 수를 의미한다.

‡ 파란색으로 표시된 수는 NB, NNB중 해당 평가지표가 가장 높은 수를 의미한다.

N = 500								
	ACC	AUC	Recall	Spec	Prec	F1	MCC	SM
TAN	0.802 (0.05)*	0.815 (0.07)	0.802 (0.08)	0.803 (0.08)	0.802 (0.08)	0.801 (0.06)	0.62 (0.11)	0.370 (0.13)
TAN-II	0.805† (0.05)	0.815 (0.06)	0.806 (0.08)	0.806 (0.08)	0.805 (0.08)	0.804 (0.06)	0.623 (0.11)	0.529 (0.19)
NTAN	0.794 (0.05)	0.813 (0.06)	0.793 (0.08)	0.796 (0.08)	0.795 (0.08)	0.793 (0.06)	0.606 (0.11)	0.337 (0.04)
NB	0.785 (0.06)	0.811 (0.06)	0.786 (0.08)	0.785 (0.08)	0.785 (0.08)	0.785 (0.06)	0.584 (0.11)	0
NNB	0.796‡ (0.06)	0.814 (0.06)	0.796 (0.08)	0.795 (0.08)	0.795 (0.08)	0.795 (0.06)	0.604 (0.11)	0
N = 1,000								
	ACC	AUC	Recall	Spec	Prec	F1	MCC	SM
TAN	0.81 (0.04)	0.81 (0.05)	0.807 (0.06)	0.813 (0.06)	0.81 (0.05)	0.807 (0.04)	0.625 (0.08)	0.450 (0.12)
TAN-II	0.811 (0.04)	0.81 (0.05)	0.807 (0.06)	0.814 (0.06)	0.811 (0.05)	0.808 (0.04)	0.625 (0.08)	0.667 (0.17)
NTAN	0.806 (0.04)	0.81 (0.05)	0.803 (0.06)	0.809 (0.06)	0.807 (0.06)	0.803 (0.04)	0.616 (0.08)	0.377 (0.11)
NB	0.785 (0.04)	0.805 (0.05)	0.784 (0.06)	0.787 (0.06)	0.785 (0.06)	0.782 (0.05)	0.574 (0.08)	0
NNB	0.795 (0.04)	0.809 (0.05)	0.792 (0.06)	0.798 (0.06)	0.795 (0.06)	0.792 (0.04)	0.598 (0.08)	0

표 4: 표본의 수에 따른 분류성능 비교

* 괄호() 안에 표시된 수는 표준편차를 의미한다.

† 빨간색으로 표시된 수는 TAN, TAN-II, NTAN중 해당 평가지표가 가장 높은 수를 의미한다.

‡ 파란색으로 표시된 수는 NB, NNB중 해당 평가지표가 가장 높은 수를 의미한다.

N = 5,000								
	ACC	AUC	Recall	Spec	Prec	F1	MCC	SM
TAN	0.82 (0.02)*	0.821 (0.02)	0.822 (0.02)	0.818 (0.02)	0.816 (0.03)	0.819 (0.02)	0.636 (0.04)	0.499 (0.02)
TAN-II	0.82† (0.02)	0.82 (0.02)	0.822 (0.02)	0.818 (0.02)	0.816 (0.03)	0.819 (0.02)	0.636 (0.04)	0.825 (0.04)
NTAN	0.819 (0.02)	0.82 (0.02)	0.821 (0.02)	0.817 (0.02)	0.816 (0.03)	0.818 (0.02)	0.635 (0.04)	0.65 (0.1)
NB	0.79 (0.02)	0.816 (0.02)	0.793 (0.03)	0.788 (0.03)	0.786 (0.03)	0.789 (0.02)	0.575 (0.04)	0
NNB	0.804‡ (0.02)	0.818 (0.02)	0.805 (0.03)	0.803 (0.02)	0.801 (0.03)	0.803 (0.02)	0.603 (0.04)	0
N = 10,000								
	ACC	AUC	Recall	Spec	Prec	F1	MCC	SM
TAN	0.817 (0.01)	0.816 (0.01)	0.816 (0.02)	0.817 (0.02)	0.816 (0.02)	0.816 (0.01)	0.636 (0.02)	0.5 (0)
TAN-II	0.817 (0.01)	0.817 (0.01)	0.816 (0.02)	0.817 (0.02)	0.816 (0.02)	0.816 (0.01)	0.636 (0.02)	0.833 (0)
NTAN	0.817 (0.01)	0.816 (0.01)	0.816 (0.02)	0.817 (0.02)	0.816 (0.02)	0.816 (0.01)	0.636 (0.02)	0.7 (0.07)
NB	0.786 (0.01)	0.813 (0.01)	0.785 (0.02)	0.787 (0.02)	0.786 (0.02)	0.785 (0.01)	0.578 (0.03)	0
NNB	0.8 (0.01)	0.815 (0.01)	0.799 (0.02)	0.801 (0.02)	0.8 (0.02)	0.799 (0.01)	0.604 (0.03)	0

표 5: 표본의 수에 따른 분류성능 비교

* 괄호() 안에 표시된 수는 표준편차를 의미한다.

† 빨간색으로 표시된 수는 TAN, TAN-II, NTAN중 해당 평가지표가 가장 높은 수를 의미한다.

‡ 파란색으로 표시된 수는 NB, NNB중 해당 평가지표가 가장 높은 수를 의미한다.

N = 50,000								
	ACC	AUC	Recall	Spec	Prec	F1	MCC	SM
TAN	0.818 (0.01)*	0.818 (0.01)	0.818 (0.01)	0.818 (0.01)	0.818 (0.01)	0.818 (0.01)	0.634 (0.01)	0.5 (0)
TAN-II	0.818† (0.01)	0.817 (0.01)	0.818 (0.01)	0.818 (0.01)	0.818 (0.01)	0.818 (0.01)	0.634 (0.01)	0.833 (0)
NTAN	0.818 (0.01)	0.818 (0.01)	0.818 (0.01)	0.818 (0.01)	0.818 (0.01)	0.818 (0.01)	0.634 (0.01)	0.966 (0.07)
NB	0.788 (0.01)	0.814 (0.01)	0.788 (0.01)	0.788 (0.01)	0.788 (0.01)	0.788 (0.01)	0.573 (0.01)	0
NNB	0.802‡ (0.01)	0.817 (0.01)	0.803 (0.01)	0.802 (0.01)	0.803 (0.01)	0.803 (0.01)	0.603 (0.01)	0

표 6: 표본의 수에 따른 분류성능 비교

표 3부터 6에서 ACC, AUC, Recall, Spec, Prec, F1, MCC, SM은 각각 정확도, ROC 밑면적, 재현율, 특이도, 정밀도, F1 score, 매튜 상관계수, 네트워크 유사도를 의미한다. 표 3에서 표본의 수의 개수가 현저히 적은 상황인 N=100인 경우, TAN의 정확도는 NB의 정확도보다 0.021, F1은 0.023, AUC는 0.02, 매튜 상관계수는 0.029만큼 낮아 조금 떨어지는 성능을 보인다. 이는 학습할 모수의 수에 비해 표본의 수가 부족하여 CPT가 희소 테이블(sparse table)로 형성되면서 발생하는 문제점이다. CPT가 희소 테이블이 되어 모수 추정이 제대로 이루어지지 않아 정확한 분류예측을 하지 못하게 된다. 하지만 TAN의 분류성능은 표 3과 표

* 괄호() 안에 표시된 수는 표준편차를 의미한다.

† 빨간색으로 표시된 수는 TAN, TAN-II, NTAN중 해당 평가지표가 가장 높은 수를 의미한다.

‡ 파란색으로 표시된 수는 NB, NNB중 해당 평가지표가 가장 높은 수를 의미한다.

4, 5, 6을 비교하였을 때 표본의 수가 증가함에 따라 정확도의 경우 0.757에서 0.791, 0.802, 0.81, 0.82로 증가하고 F1의 경우 0.749, 0.788, 0.801, 0.807, 0.819로 증가하였으며 매튜 상관계수는 0.52에서 0.636까지 증가하였다. 표본의 수가 5,000개 이상인 표 5와 6에서 정확도는 0.817 – 0.82, AUC는 0.816 – 0.821, 재현율은 0.818 – 0.822, 특이도는 0.818, 정밀도는 0.816 – 0.818, F1은 0.816 – 0.819, 매튜 상관계수는 0.634 – 0.636 사이의 값을 보여주며 표 3, 4와 비교하였을 때 표본의 수에 따른 분류성능의 변동이 상대적으로 적다. 이는 표본의 수가 5,000개가 넘는 시점부터 모수 학습이 충분히 된 상태로 분류예측을 진행하는 것을 의미한다. 그에 비해 NB의 경우 TAN에 비해 학습할 모수의 수가 적어 표본의 수의 영향을 덜 받는 것으로 보인다. 표본의 수가 매우 적은 경우인 표 3의 N=100인 경우와 표본의 수가 많은 경우인 표 6의 N=50,000인 경우를 비교하였을 때 NB의 정확도는 0.788로 동일한 것을 확인할 수 있다. 그러나 이 외에 AUC, 재현율, 특이도, 정밀도, F1, 매튜 상관계수는 표본의 수가 많은 경우인 표 5와 6에서 조금씩 높은 성능을 보이는 것을 볼 수 있다. NB의 경우에도 표본의 수에 영향을 받긴 하나 TAN 만큼 많은 영향을 받지 않는 것으로 보인다.

본 논문에서 제시하는 NTAN의 경우 표 3에서 N=100인 경우 TAN보다 모든 분류성능 평가지표에서 더 좋은 성능을 보인다. 특히 AUC는 0.807로 0.788인 TAN보다 약 0.02가량 높고, 정확도의 경우 0.763으로 0.757인 TAN보다 0.006만큼 높아 표본의 수가 현저히 적은 경우 NTAN이 TAN보다 나은 분류성능을 보이고 있다. 또한 표본의 수가 5,000개 이상인 표 5와 6에서 NTAN과 TAN의 분류성능이 거의 유사함을 확인할 수 있다. 그러나 NTAN의 경우 변수간 실제 인과관계를 평가하는 네트워크 유사도가 TAN의 네트워크

유사도보다 높고, 특히 표본의 수가 50,000개인 경우에 NTAN의 네트워크 유사도가 0.966으로 변수간 실제 인과관계를 거의 완벽하게 구현해 더 적절한 네트워크가 형성되어 결과해석이 용이하다는 장점이 있다. 그리고 NNB의 경우 표 3, 4, 5, 6 모두에서 표본의 수에 상관없이 NB보다 정확도, AUC, 재현율, 특이도, 정밀도, F1, 매튜 상관계수에서 항상 같거나 좋은 성능을 보이고 있다. 특히 표 6에서 N=50,000인 경우 NNB의 정확도는 0.802로 0.788인 NB보다 0.014만큼 높고, F1의 경우 0.803으로 0.788인 NB보다 0.015만큼 높으며 매튜 상관계수는 0.603으로 0.03만큼 높다. NNB의 경우에도 표본의 수가 많을수록 분류성능이 상승하는 것을 확인할 수 있다. 표 3에 N=100인 경우와 표 6에 N=50,000인 경우를 비교하였을 때 정확도는 0.781에서 0.802로, F1은 0.775에서 0.803로, 매튜 상관계수는 0.549에서 0.603으로 상승하였다. 그에 반해 NB의 경우 표 3과 표 6을 비교하여 정확도는 0.778에서 0.788로, F1은 0.771에서 0.788로 상승하긴 하였으나 NNB와 비교하였을 때 상승 폭이 적어 성능간 차이가 벌어지게 되었다.

그림 4는 시뮬레이션 I에서 생성하고자 하는 변수간 실제 인과관계를 네트워크로 표현한 것이다. 그림 5의 TAN 네트워크와 그림 4를 비교하였을 때 TAN의 네트워크는 생성한 데이터 분포에 맞는 네트워크가 형성되어 변수간 관계를 표현하지만 제외하고자 하는 변수집합 E가 포함되어 있어 인과관계 해석에 어려움을 준다. 그에 반해 NTAN의 네트워크에는 변수집합 E가 제외되면서 그림 4와 동일한 네트워크가 형성되었다. 이는 NTAN의 네트워크가 변수간 인과관계를 정확히 표현하여 적절한 인과관계 해석이 가능함을 나타낸다. 이는 표 5와 6에서 표본의 개수가 충분할 때 NTAN의 네트워크 유사도가 TAN의 네트워크 유사도보다 높으므로 더 적절한 네트워크가 형성됨을 수치적으로도 확인할 수 있다.

특히 표본의 수가 50,000개인 경우 NTAN의 네트워크 유사도는 0.966으로 변수간 실제 인과관계를 거의 완벽하게 재현하는 것을 확인하였다. 하지만 NB와 NNB의 경우 변수들이 모두 조건부 독립이기 때문에 변수간 관계가 형성되지 않으므로 네트워크 유사도가 0으로 계산된다. 그림 6에서 NB의 네트워크와 NNB의 네트워크를 비교하면 NNB의 네트워크에 변수집합 E 가 제외된 것을 확인할 수 있다.

5.2 시뮬레이션 II

시뮬레이션 I을 통해 표본의 수에 따른 TAN, TAN-II, NTAN, NB, NNB의 분류성능과 네트워크 유사도를 비교하였다. 이번에는 표본의 수 N=50,000으로 고정시키고, 변수의 개수 p=10, 25, 40, 70개인 총 4가지 상황에 대하여 시뮬레이션을 진행한다. 시뮬레이션 I과 마찬가지로 Y 변수는 예측하고자 하는 변수이며 제외할 변수집합 E는 (sex, age)로 설정하였다. 또한 Y 변수에 제외할 변수집합 E를 포함시켜 NC 변수를 생성하였다. 변수의 개수 p=10인 경우의 변수 생성 방식과 네트워크 형태는 시뮬레이션 I과 같다. p=25인 경우에는 다음과 같이 생성하였다. 변수의 개수 p=40인 경우와 p=70인 경우의 시뮬레이션 모델과 네트워크 구조는 부록에 첨부하였다. 또한 TAN, TAN-II, NTAN, NB, NNB 구축 방식은 시뮬레이션 I과 같은 방식으로 구축하고 분류예측을 진행하였다. 시뮬레이션 I과 마찬가지로 네트워크 유사도를 이용하여 TAN, TAN-II, NTAN의 네트워크가 실제 변수간 인과관계를 잘 표현하는지 평가하였다.

$Y \sim B(P = 0.5), \quad \text{sex} \sim B(P = 0.5),$

$x_1 \left| Y \sim B\left(P = \frac{\exp(-1.5 + 3 \times Y)}{1 + \exp(-1.5 + 3 \times Y)}\right), \quad x_2 \right| x_1 \sim B\big(\mathrm{P} = \Phi(-1 + 2 \times \mathrm{x}_1)\big),$

$x_3 | x_1 \sim N(\mu = 10 + 5 \times x_1,\ \text{sd} = 2), \quad \text{age} | x_2 \sim B\left(P = \frac{\exp(-1 + 2 \times x_2)}{1 + \exp(-1 + 2 \times x_2)}\right),$

$x_4 | x_2 \sim N(\mu = 50 + 15 \times x_2,\ \text{sd} = 3),$

$x_5 | \text{age}, x_2 \sim B\left(P = \frac{\exp(1.5 - 3 \times \text{age} - x_2\ )}{1 + \exp(1.5 - 3 \times \text{age} - x_2\ )}\right),$

$x_6 | x_3 \sim N(\mu = 100 + 1.5 \times x_3,\ \text{sd} = 5), x_7 | \text{age}, x_2 \sim B\big(P = \Phi(1 - 2 \times \text{age} - 0.5 \times x_2)\big),$

$x_8 \left| Y \sim B\left(P = \frac{\exp(-1 + 2 \times Y)}{1 + \exp(-1 + 2 \times Y)}\right), \quad x_9 \right| \text{age}, x_2 \sim B\big(P = \Phi(-1.5 + 3 \times \text{age} - x_2)\big),$

$x_{10} | x_8 \sim N(\mu = 30 + 3 \times x_8,\ \text{sd} = 1.5), \quad x_{11} | x_9 \sim N(\mu = 15 + 2 \times x_9,\ \text{sd} = 1),$

$x_{12} | x_{10} \sim N(\mu = 50 + 1.5 \times x_{10},\ \text{sd} = 2), \quad x_{13} | x_9 \sim B\big(P = \Phi(-1 + 2 \times x_9)\big),$

$x_{14} \left| x_{13} \sim B\left(P = \frac{\exp(-1 + 2 \times x_{13})}{1 + \exp(-1 + 2 \times x_{13})}\right), \quad x_{15} \right| Y \sim B\left(P = \frac{\exp(0.5 - Y)}{1 + \exp(0.5 - Y)}\right),$

$x_{16} \big| \text{age}, x_2 \sim B\big(P = \Phi(2 - 4 \times \text{age} - x_2)\big), \quad x_{17} \big| x_{15} \sim N(\mu = 50 - 3 \times x_{15},\ \text{sd} = 2),$

$x_{18} \big| x_{16} \sim B\big(P = \Phi(1 - 2 \times x_{16})\big), \quad x_{19} \big| x_{16} \sim B\left(P = \frac{\exp(-1 + 2 \times x_{16})}{1 + \exp(-1 + 2 \times x_{16})}\right),$

$x_{20} \left| x_{19} \sim B\left(P = \frac{\exp(-1.5 + 3 \times x_{19})}{1 + \exp(-1.5 + 3 \times x_{19})}\right), \quad x_{21} \right| x_{17} \sim N(\mu = 10 + x_{17},\ \text{sd} = 1),$

$x_{22} | x_{18} \ \sim B\left(P = \frac{\exp(-0.5 + x_{18})}{1 + \exp(-0.5 + x_{18})}\right).$

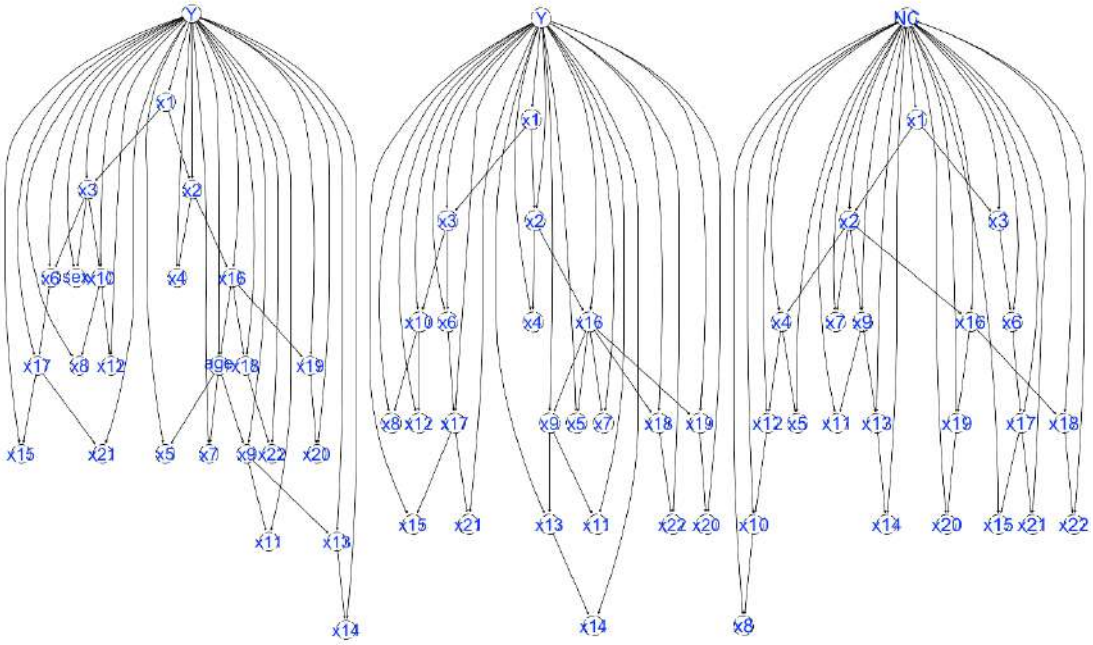


그림 7: p=25인 경우의 TAN(왼쪽), TAN-II(가운데), NTAN(오른쪽)

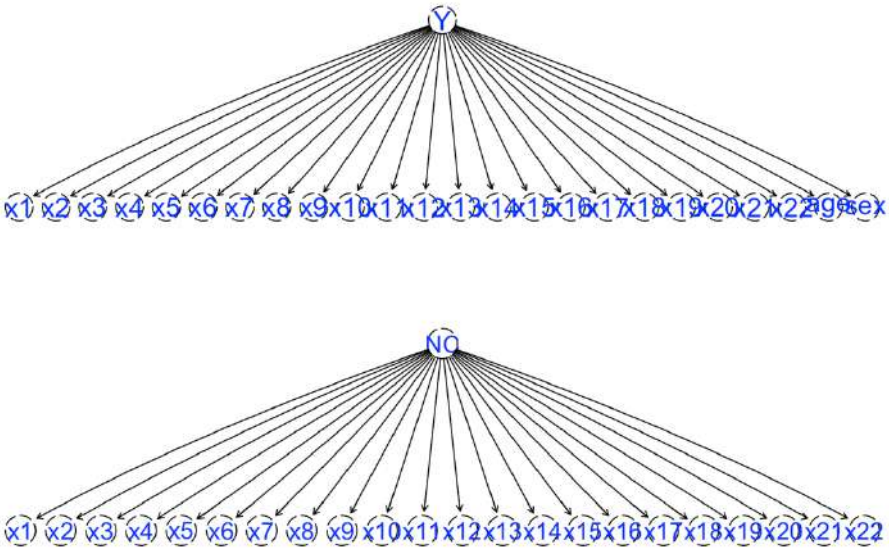


그림 8: p=25인 경우의 NB(위), NNB(아래)

p = 10								
	ACC	AUC	Recall	Spec	Prec	F1	MCC	SM
TAN	0.818 (0.01)*	0.818 (0.01)	0.818 (0.01)	0.818 (0.01)	0.818 (0.01)	0.818 (0.01)	0.634 (0.01)	0.5 (0)
TAN-II	0.818† (0.01)	0.817 (0.01)	0.818 (0.01)	0.818 (0.01)	0.818 (0.01)	0.818 (0.01)	0.634 (0.01)	0.833 (0)
NTAN	0.818 (0.01)	0.818 (0.01)	0.818 (0.01)	0.818 (0.01)	0.818 (0.01)	0.818 (0.01)	0.634 (0.01)	0.966 (0.07)
NB	0.788 (0.01)	0.814 (0.01)	0.788 (0.01)	0.788 (0.01)	0.788 (0.01)	0.788 (0.01)	0.573 (0.01)	0
NNB	0.802‡ (0.01)	0.817 (0.01)	0.803 (0.01)	0.802 (0.01)	0.803 (0.01)	0.803 (0.01)	0.603 (0.01)	0
p = 25								
	ACC	AUC	Recall	Spec	Prec	F1	MCC	SM
TAN	0.817 (0.01)	0.901 (0.01)	0.816 (0.01)	0.818 (0.01)	0.817 (0.01)	0.817 (0.01)	0.635 (0.01)	0.567 (0.03)
TAN-II	0.817 (0.01)	0.901 (0.01)	0.816 (0.01)	0.818 (0.01)	0.817 (0.01)	0.817 (0.01)	0.636 (0.01)	0.621 (0.03)
NTAN	0.817 (0.01)	0.901 (0.01)	0.816 (0.01)	0.818 (0.01)	0.818 (0.01)	0.817 (0.01)	0.635 (0.01)	0.75 (0.04)
NB	0.786 (0.01)	0.855 (0.01)	0.786 (0.01)	0.787 (0.01)	0.786 (0.01)	0.786 (0.01)	0.574 (0.01)	0
NNB	0.805 (0.01)	0.878 (0.01)	0.804 (0.01)	0.805 (0.01)	0.805 (0.01)	0.804 (0.01)	0.611 (0.01)	0

표 7: 변수의 개수에 따른 분류성능 비교

* 괄호() 안에 표시된 수는 표준편차를 의미한다.

† 빨간색으로 표시된 수는 TAN, TAN-II, NTAN중 해당 평가지표가 가장 높은 수를 의미한다.

‡ 파란색으로 표시된 수는 NB, NNB중 해당 평가지표가 가장 높은 수를 의미한다.

p = 40								
	ACC	AUC	Recall	Spec	Prec	F1	MCC	SM
TAN	0.862 (0.01)*	0.932 (0.01)	0.86 (0.01)	0.864 (0.01)	0.863 (0.01)	0.861 (0.01)	0.724 (0.01)	0.554 (0.02)
TAN-II	0.863 (0.01)	0.933 (0.01)	0.862 (0.01)	0.864 (0.01)	0.863 (0.01)	0.862 (0.01)	0.725 (0.01)	0.584 (0.03)
NTAN	0.864† (0.01)	0.934 (0.01)	0.865 (0.01)	0.864 (0.01)	0.864 (0.01)	0.864 (0.01)	0.729 (0.01)	0.712 (0.03)
NB	0.796 (0.01)	0.882 (0.01)	0.796 (0.01)	0.797 (0.01)	0.797 (0.01)	0.796 (0.01)	0.592 (0.01)	0
NNB	0.842‡ (0.01)	0.918 (0.01)	0.842 (0.01)	0.842 (0.01)	0.842 (0.01)	0.842 (0.01)	0.686 (0.01)	0
p = 70								
	ACC	AUC	Recall	Spec	Prec	F1	MCC	SM
TAN	0.9 (0.01)	0.964 (0.01)	0.9 (0.01)	0.9 (0.01)	0.9 (0.01)	0.9 (0.01)	0.8 (0.01)	0.546 (0.02)
TAN-II	0.901 (0.01)	0.964 (0.01)	0.901 (0.01)	0.901 (0.01)	0.902 (0.01)	0.901 (0.01)	0.802 (0.01)	0.563 (0.02)
NTAN	0.902 (0.01)	0.965 (0.01)	0.902 (0.01)	0.903 (0.01)	0.903 (0.01)	0.903 (0.01)	0.805 (0.01)	0.682 (0.03)
NB	0.808 (0.01)	0.901 (0.01)	0.807 (0.01)	0.808 (0.01)	0.807 (0.01)	0.807 (0.01)	0.614 (0.02)	0
NNB	0.885 (0.01)	0.954 (0.01)	0.885 (0.01)	0.885 (0.01)	0.884 (0.01)	0.885 (0.01)	0.771 (0.01)	0

표 8: 변수의 개수에 따른 분류성능 비교

* 괄호() 안에 표시된 수는 표준편차를 의미한다.

† 빨간색으로 표시된 수는 TAN, TAN-II, NTAN중 해당 평가지표가 가장 높은 수를 의미한다.

‡ 파란색으로 표시된 수는 NB, NNB중 해당 평가지표가 가장 높은 수를 의미한다.

그림 7에서 TAN의 네트워크의 형태를 보면 생성한 변수의 분포 형태에 맞는 네트워크가 형성되어 있다. 마찬가지로 그림 부록1-1의 TAN 네트워크와 그림 부록2-1의 TAN 네트워크 역시 생성하고자 하는 데이터 분포에 맞는 네트워크가 형성되어 있다(부록 참조). 그러나 네트워크에 변수집합 E가 포함되어 변수간 실제 인과관계를 적절히 나타내지 못해 인과관계 해석에 어려움을 주고 있다. 그에 비해 그림 7, 그림 부록1-3, 그림 부록 2-3에서 NTAN의 네트워크는 변수집합 E가 네트워크에서 제외되어 실제 인과관계를 적절하게 나타내는 자연스러운 네트워크가 형성되었다(부록 참조). 시뮬레이션 II의 경우 표본의 수를 50,000으로 설정하였기 때문에 시뮬레이션 I에서 살펴보았듯이 TAN의 성능이 NB보다 항상 좋은 것을 볼 수 있다. 표 7에서 변수의 개수가 10개인 경우 TAN의 정확도는 0.818로 NB의 0.788보다 0.03만큼 높고, F1은 0.818로 NB의 0.788보다 0.03만큼 높으며 매튜 상관계수는 0.634로 NB의 0.573보다 0.061만큼 높게 나타났다. 변수의 개수가 증가하는 만큼 TAN의 성능 또한 향상되었는데, 특히 표 8에서 변수의 개수가 70개인 경우 TAN의 정확도는 0.9로 0.808인 NB보다 0.092 높으며 AUC의 경우 0.964로 0.901인 NB보다 0.063만큼 높고 매튜 상관계수는 0.8로 NB보다 0.186만큼 높게 나타났다. F1 또한 0.9로 0.807인 NB보다 0.093 높은 것으로 나타났다. 본 논문에서 제시하는 NTAN의 경우 분류성능이 TAN, TAN-II와 비교하여 동일하거나 우수함을 확인하였다. 표 7에서 변수의 개수가 10개인 경우 모든 분류성능 평가지표에서 NTAN과 TAN이 같았으며, TAN-II의 경우 AUC가 0.817로 NTAN이 0.001 더 높은 것을 확인하였다. 그리고 표 8에서 변수의 개수가 40개인 경우 NTAN의 특이도는 TAN, TAN-II와 같았으나 이 외에 평가지표에서 가장 높은 성능을 보이는 것을 확인하였다. 변수의 개수가 70개인

경우 NTAN의 정확도, AUC, 재현율, 특이도, 정밀도, F1, 매튜 상관계수가 TAN, TAN-II보다 높은 것을 볼 수 있다. NB의 경우 변수의 개수가 증가할수록 AUC가 높게 나타났지만 정확도, 재현율, 특이도, 정밀도, F1, 매튜 상관계수는 AUC가 상승한 것에 비해 많이 상승하지 않았다. 표 7에서 변수의 개수가 10개인 경우에 0.814이던 AUC가 표 8에서 변수의 개수가 70개일 때 0.901까지 대폭 상승하였으나, 정확도와 F1의 경우 각각 0.788에서 0.808, 0.788에서 0.807로 소폭 상승하였다. 그에 비해 본 논문에서 제시하는 방식을 적용시킨 NNB의 경우 변수의 개수가 증가할수록 분류성능이 많이 향상되었으며, 특히 NB보다 항상 좋은 분류성능을 보이고 있다. 표 7과 표 8을 비교하였을 때 변수의 개수가 10개인 경우 NNB의 정확도가 0.802로 0.788인 NB보다 0.014만큼 높았고 변수의 개수가 70개인 경우 NNB의 정확도는 0.885로 0.808인 NB보다 0.078 높은 성능을 보여주고 있다. 또한 변수의 수가 10개일 때 NNB의 AUC는 0.817로 0.814인 NB와 비슷한 성능을 보였으나, 변수의 수가 70개로 증가한 경우 NNB의 AUC는 0.954로 0.901인 NB보다 0.053 높은 성능을 보여주고 있으며 매튜 상관계수는 0.771로 NB의 0.614보다 0.157 높은 성능을 보여준다. 이처럼 NNB는 NB보다 항상 높은 정확도, AUC, 재현율, 특이도, 정밀도, F1, 매튜 상관계수를 보이고 변수의 개수가 증가할수록 각 분류성능의 차이가 더 심하게 벌어지는 것을 확인할 수 있다.

NTAN의 네트워크 유사도는 표 7과 8에서 변수의 개수가 10개일 때 0.966, 25개일 때 0.75, 40개일 때 0.712, 70개일 때 0.682로 TAN, TAN-II의 네트워크 유사도보다 높은 것을 확인할 수 있다. 이는 NTAN의 네트워크가 변수간 실제 인과관계를 나타내는데에 있어 TAN과 TAN-II에 비해 우수함을 나타낸다. 또한

시뮬레이션 I과 마찬가지로 NB와 NNB의 경우 조건부 독립으로 인해 변수간의 관계가 형성되지 않아 네트워크 유사도가 0으로 계산된다.

시뮬레이션 I과 II를 통해 TAN, TAN-II, NTAN, NB, NNB의 표본의 수에 따른 분류성능과 네트워크 유사도, 변수의 개수에 따른 분류성능과 네트워크 유사도를 비교하였다. 시뮬레이션 I에서 TAN과 NTAN이 표본의 수가 너무 적은 경우 모수가 제대로 학습되지 못하는 문제가 발생하여 NB와 NNB에 비해 전체적으로 낮은 분류성능을 보여주지만, 반대로 NB와 NNB의 경우 표본의 수가 적음에도 좋은 성능을 보여주었다. 또한 표본의 수가 5,000개를 넘는 시점부터 NTAN의 모수 학습이 충분히 이루어져 TAN과 모든 분류성능이 비슷해졌다. 그리고 표본의 개수와 상관없이 NNB의 경우 NB보다 모든 평가지표에서 좋은 성능을 나타냄을 확인하였다. NTAN의 경우 TAN과 유사한 분류성능을 보이지만, 인과관계에 포함되지 말아야 할 변수가 포함되어 부자연스러운 해석을 야기하는 TAN과 비교하여 더 단순하고 자연스러운 네트워크를 형성함으로써 변수간 관계 해석이 용이하다는 장점이 있다. 이는 표본의 개수가 5,000개 이상인 경우에 NTAN의 네트워크 유사도가 TAN의 네트워크 유사도보다 높으므로 수치적으로도 입증할 수 있다. 또한 표본의 수가 적은 경우에는 NTAN의 정확도, AUC, F1, 매튜 상관계수 등이 TAN보다 좋음을 표 3에서 표본의 수가 100개인 경우를 통해 확인하였다. 시뮬레이션 II에서는 TAN, TAN-II, NTAN, NB, NNB의 분류성능이 변수의 개수가 증가할수록 상승하였다. 특히 NNB의 경우 변수의 개수가 증가함에 따라 기존의 NB보다 분류성능이 많이 향상되었다. 시뮬레이션 II를 통해 TAN, TAN-II, NTAN, NNB가 변수의 개수가 많을수록, 관계가 많을수록 더 좋은 분류성능을 내는 것을 확인하였다. 특히 본 논문에서 제시하는 NTAN의 분류성능이 TAN,

TAN-II와 비교하여 변수의 개수가 10개인 경우와 25개인 경우에는 거의 같았으며, 변수의 개수가 40개인 경우와 70개인 경우에는 우수함을 확인하였다. 게다가 네트워크 유사도 또한 NTAN이 TAN과 TAN-II와 비교하여 높은것을 확인하여 NTAN이 분류성능면에서도, 변수간 실제 인과관계를 표현하는데에 있어서도 TAN과 TAN-II보다 우수함을 확인하였다.

이번 장에서는 시뮬레이션을 통해 가상의 데이터에서 NTAN과 NNB가 잘 적합하는 것을 확인하였다. 다음 6장에서는 실제 데이터를 이용하여 NTAN과 TAN의 네트워크를 구축하고 형태를 비교한다. 그 후 네트워크를 통한 변수간 인과관계 해석을 비교한다. 또한 다른 머신러닝 기법들과 함께 NTAN과 NNB의 분류성능을 비교한다.

제 6 장 실제 데이터 적용

이번 장에서는 앞서 시뮬레이션을 진행한 TAN, TAN-II, NTAN, NB, NNB에 더해 K-NN, 의사결정나무, 랜덤포레스트의 분류성능을 실제 데이터를 통해 비교하고자 한다. 본 논문에서는 국민건강보험(NHISS)에서 제공하는 국가건강검진 데이터 중 시력검사 데이터와 혈액검사 데이터를 사용하고, 거기에 더해 UCI에서 제공하는 유방암 재발 여부 데이터를 사용한다.

국가건강검진 데이터의 경우 표본의 수가 많으므로 실험을 진행할 때 전체 데이터 셋을 10-fold하여 훈련 데이터와 평가 데이터를 9:1의 비율로 분할하여 진행하였다. 하지만 UCI에서 제공하는 데이터는 표본의 수가 적기 때문에 데이터 셋을 3-fold하여 훈련 데이터와 평가 데이터를 2:1의 비율로 분할하여 진행하였다. 일반적으로 K-NN를 구현하기 위해 R 패키지 'class'의 knn함수를 사용한다(Ripley와 Venables, 2015). 그러나 이 함수의 경우 연속형 데이터간 거리를 측정하는 척도만 제공하며 이산형 데이터간 거리를 측정하는 척도는 제공하지 않는다. 따라서 이산형 데이터간 거리를 측정하기 위해 해밍 거리를 계산하여 K-NN을 직접 구현하였다. K-NN의 하이퍼 파라미터인 최적의 k값을 찾기 위해 10겹 교차 검증(10-fold-cross validation)을 진행하였다. 교차 검증 과정에서 분류성능 중 정확도가 가장 높게 나오는 k값을 최적의 k값으로 결정하였다. 의사결정나무는 R 패키지 'rpart'의 rpart함수를 이용하였다. 그리고 prune함수로 형성된 나무에 가지치기 작업을 진행하여 최적의 나무구조를

형성하였다. 또한 랜덤포레스트의 경우 R 패키지 'randomForest'의 randomForest함수를 사용하여 구현하였다. 각 노드에서 의사결정에 사용되는 변수의 개수는 $\sqrt{p}$ 로 설정하며 생성할 나무 개수인 tree는 K-NN의 경우와 마찬가지로 10겹 교차 검증 과정을 통해 선정된 최적의 tree값으로 설정한다. 이 때 p 는 설명변수의 개수이다. TAN, TAN-II, NTAN, NB, NNB의 경우 시뮬레이션에서 설정한 것과 동일하게 smoothing 값(α)는 1로 하여 진행한다.

시력검사 데이터의 경우 전체 데이터 셋을 10-fold하여 훈련 데이터와 평가 데이터로 분할하고, 10겹 교차 검증 과정을 통해 K-NN의 최적의 k값과 랜덤포레스트의 최적의 tree값을 찾는 작업을 거치고, TAN, TAN-II, NTAN, NB, NNB, K-NN, 의사결정나무, 랜덤포레스트를 적합하는 과정을 100번 반복 시행한 후 분류성능을 평균 내어 결과를 비교하였다. 같은 과정을 혈액검사 데이터의 경우에는 50번을, UCI데이터의 경우에는 300번을 반복 시행한 후 분류성능을 평균 내어 결과를 비교하였다. 이번 장에서는 분류성능 평가지표 중 매튜 상관계수 대신 모델 적합시간을 비교하여 모델의 효율성도 함께 비교하였다.

6.1 불균형 데이터 처리

본 논문에서 다루는 국가건강검진 데이터와 UCI 데이터 모두 예측하고자 하는 클래스 변수 내에 클래스 비율이 불균형하게 분포하는 불균형 데이터이다. 데이터 처리 작업 없이 불균형 데이터를 이용하여 분석을 진행하는 경우 성능평가에 문제가 생긴다. 예를 들어 10,000개의 부품 중 실제 1,000개가 불량이라고 하였을 때, 모든 부품이 정상이라고 예측하면 정확도가 매우 높게 나타나지만 불량에 대한

민감도와 정밀도는 매우 떨어지는 현상이 발생한다. 따라서 정확한 분류성능 측정이 힘들기 때문에 이를 해결하기 위해 클래스의 비율을 조정하는 과정이 필요하다.

불균형 데이터 처리 방식에는 크게 under-sampling방법과 over-sampling 방법이 존재한다. 클래스 내에 비율이 높은 클래스를 다수 클래스, 비율이 낮은 클래스를 소수 클래스라고 할 때 under-sampling방법은 다수 클래스에 속하는 데이터를 제거하여 소수 클래스의 비율에 맞추는 방식이며 over-sampling 방법은 소수 클래스에 속하는 데이터를 복제하여 다수 클래스의 비율에 맞추는 방식이다. 두 방법 모두 단점이 존재하는데, 먼저 under-sampling방법의 경우 다수 클래스에 속하는 데이터를 제거하기 때문에 유용한 데이터를 잃는다는 위험성이 있다. Over-sampling방법의 경우 소수 클래스에 속하는 데이터를 복제하기 때문에, 데이터의 중복이 심해져 과적합이 발생할 위험성이 존재한다(Gu 등, 2008). 불균형 데이터 처리방법에는 다양한 방법이 존재하는데, 본 논문에서는 통상적으로 많이 사용되는 SMOTE(Synthetic Minority Over-sampling Technique)기법을 사용한다(Chawla 등, 2002). SMOTE기법은 기존 over-sampling기법과는 다르게 단순하게 데이터를 복제하는 것이 아닌 같은 클래스에 속하는 주위 k개의 데이터를 이용하여 데이터를 일반적으로 생성한다. 따라서 over-sampling기법의 단점인 과적합의 위험을 줄여 일반적으로 많이 사용되고 있다. 데이터에 SMOTE기법을 적용하기 위해 R 패키지 'unbalanced'의 ubSMOTE함수를 사용한다(Dal Pozzolo 등, 2015). 또한 주위에 참고할 데이터 개수 k는 5개로 설정하였으며 클래스 비율을 1:1로 맞춰 주었다.

6.2 시력검사 데이터

시력검사 데이터는 2015-2016년 일반검진 및 생애전환기 건강검진 데이터 1,000,000건으로 이루어져 있으며 항목은 다음과 같다. 변수는 아래 표 9와 같이 2개의 연속형 변수와 6개의 이산형 변수 총 8개의 변수로 이루어져 있다. TAN은 이산형 변수를 입력값으로 받으므로, 연속형 변수들에 대해 범주화 작업을 진행한다. 좌측 시력과 우측 시력은 0.1이하와 0.2-2.0 그리고 실명 총 21개의 카테고리로 범주화 하였고 나이 그룹은 생애 주기인 20-29세(청년), 30-49세(중년), 50-64세(장년), 65-(노년) 총 4개의 범주로 축소하여 진행하였다.

변수명	타입	의미	값
SEX	이산형	성별	여성: 1 남성: 0
AGE_G	이산형	나이 그룹	20 세 ~ 75 세 이상 (27 개 그룹)
HTN	이산형	고혈압	있음: 1 없음: 0
DM	이산형	당뇨병	있음: 1 없음: 0
GLAUCOMA	이산형	녹내장	있음: 1 없음: 0
VA_LT	연속형 (범주화 필요)	좌측 시력	0.1 이하 ~ 2.0, 실명
VA_RT	연속형 (범주화 필요)	우측 시력	0.1 이하 ~ 2.0, 실명
DR	이산형	당뇨망막병증	있음: 1 없음: 0

표 9: 시력검사 데이터 변수

총 8개의 변수 중 당뇨망막병증(DR)을 예측하고자 하는 클래스 변수로 선정하여 분류를 진행하였다. 변수 중 나이와 성별은 제어가 불가능한 요인들이기 때문에, 인과관계에 포함시킬 수 없다. 또한 인과관계 해석에서도 부자연스러운 해석을 초래하는 요인들이기 때문에 제약이 필요하다. 따라서 네트워크에 제외할 변수집합 E 는(SEX, AGE_G)로 선정하였고 변수집합 E 를 당뇨망막병증(DR) 변수에 포함시킨 class 변수를 생성하여 NTAN과 NNB를 구현하고 분류예측을 진행하였다. TAN과 NB는 변수집합 E 를 네트워크에서 제외하지 않고 기존의 당뇨망막병증을 예측할 클래스 변수로 사용하여 분류예측을 진행한다. TAN-II는 변수집합 E 를 네트워크에서 제외하고 기존의 당뇨망막병증을 예측할 클래스 변수로 사용하여 분류예측을 진행한다.

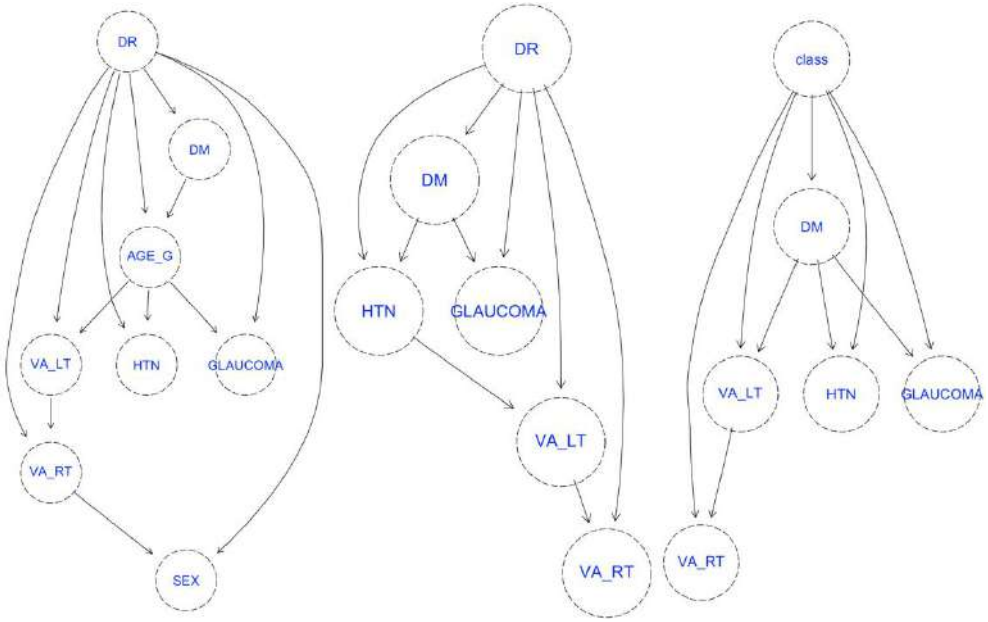


그림 9: 시력검사 데이터의 네트워크: TAN(왼쪽), TAN-II(가운데), NTAN(오른쪽)

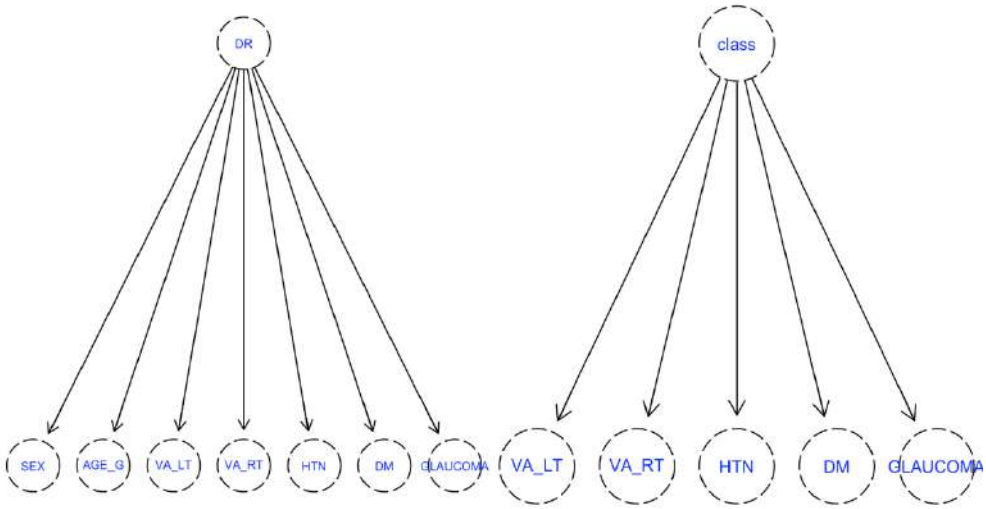


그림 10: 시력검사 데이터의 네트워크: NB(왼쪽), NNB(오른쪽)

	ACC	AUC	Recall	Spec	Prec	F1	Time
TAN	0.826	0.919	0.823	0.927	0.997	0.902	0.89(s)
TAN-II	0.813	0.919	0.809	0.945	0.998	0.894	0.57(s)
NTAN	0.847†	0.923	0.845	0.931	0.998	0.915	2.37(s)
NB	0.787	0.914	0.784	0.914‡	0.997	0.877	0.16(s)
NNB	0.853*	0.924	0.852	0.909	0.997	0.919	0.22(s)
K-NN	0.816	0.908	0.814	0.914	0.997	0.896	5.47(h)
DT	0.823	0.906	0.82	0.945	0.998	0.9	0.79(s)
RF	0.819	0.918	0.816	0.944	0.998	0.898	17.28(m)

표 10: 시력검사 데이터 분류성능 평가지표

표 10에서 K-NN은 K-NN기법(K-Nearest Neighbor), DT는 의사결정나무(Decision Tree), RF는 랜덤포레스트(Random Forest)를 의미한다. 그리고 ACC, AUC, Recall, Spec, Prec, F1, Time은 각각 정확도, ROC 밑면적, 재현율, 특이도, 정밀도, F1 score, 모델 적합시간을 의미하며 Time에서 s는 초, m은 분, h는 시간을 의미한다. 표 10에서 NTAN의 정확도는 0.847로 0.826인 TAN과 비교하여 0.021 높은 정확도를 보이고 있고, 특히 0.813인 TAN-II와

† 빨간색으로 표시된 수는 TAN, TAN-II, NTAN중 해당 평가지표가 가장 높은 수를 의미한다.

‡ 파란색으로 표시된 수는 NB, NNB중 해당 평가지표가 가장 높은 수를 의미한다.

* 밑줄친 수는 해당 평가지표에서 가장 높은 수를 의미한다.

비교하였을 때 0.034 높은 정확도를 보여 단순히 변수집합 E를 제외하는 것보다 높은 정확도를 보이고 있다. 또한 K-NN, 의사결정나무(DT), 랜덤포레스트(RF)와 비교하였을 때도 더 높은 정확도를 보인다. NNB의 경우 정확도가 0.853으로 0.787인 NB보다 0.066만큼 높은 수치를 보이고 모델 중 가장 높은 정확도를 나타내고 있다. 그리고 AUC 측면에서도 NTAN은 0.923으로 0.919인 TAN과 TAN-II에 비해 성능이 향상되었으며, NNB는 0.924로 0.914인 NB보다 향상되었다. 다른 머신러닝 기법보다도 높은 수치를 보이며, 특히 0.906인 의사결정나무에 비해 NTAN과 NNB는 각각 0.017, 0.018씩 더 높다. 특이도는 의사결정나무와 TAN-II이 0.945로 가장 높은 반면, NTAN은 0.931로 조금 떨어지는 성능을 보이고 특히 NNB는 0.909로 가장 낮게 나타났다. 그리고 정밀도와 재현율의 조화평균으로 표현되는 F1은 NTAN이 0.915로 0.902, 0.894인 TAN, TAN-II보다 높았으며, NNB 또한 0.919로 0.877인 NB보다 좋은 성능을 나타낸다. 머신러닝 기법들과 F1을 비교할 때 K-NN은 0.896, 의사결정나무는 0.9, 랜덤포레스트는 0.898로 NTAN과 NNB가 더 높은 수치를 나타내고 있다. NTAN과 NNB의 모델 적합 시간은 2.4초와 0.22초로 약 5시간인 K-NN과 약 17분인 랜덤포레스트에 비해 매우 빠르게 적합하는 것을 볼 수 있다. 게다가 정확도, AUC, 재현율, F1 면에서의 분류성능 또한 더 높아 NTAN과 NNB가 적합시간 대비 분류성능이 좋은 효율적인 모델이라고 할 수 있다. 그러나 TAN이 약 0.9초, TAN-II가 약 0.6초인 것에 비해 NTAN이 적합되는 데에 시간이 더 소요되는데, 이는 모수학습 단계에서 NTAN이 TAN, TAN-II보다 더 많은 모수를 학습해야 하기 때문이다. NNB의 경우 모델 적합하는 데에 소요되는 시간이 0.22초로 매우 짧은 시간을 소요하면서 정확도, AUC, 재현율, F1에서 가장

좋은 성능을 보여 매우 효율적인 모델이라고 볼 수 있다. 결론적으로 본 논문에서 제시하는 방식인 NTAN과 NNB는 기존의 TAN, NB보다 모든 평가 지표에서 향상된 분류성능을 보이며 특히 NTAN은 단순히 네트워크에서 변수집합 E 를 제외하는 TAN-II보다 특이도를 제외한 모든 분류성능 평가지표가 우수함을 확인하였다. 또한 머신러닝 기법들과 비교해도 NTAN과 NNB가 높은 정확도와 AUC, F1을 나타내며 좋은 분류성능을 보이는 것을 확인할 수 있다.

그림 9에서 TAN의 네트워크는 좌측 시력(VA_LT), 고혈압(HTN), 녹내장(GLAUCOMA)과 나이(AGE_G)간, 그리고 우측 시력(VA_RT)과 성별(SEX)간 관계가 성립되는 네트워크가 형성되었다. 이는 좌측 시력, 고혈압 그리고 녹내장이 나이에 의해 영향을 받는 부적절한 관계 해석을 초래하므로 부자연스러운 네트워크라고 볼 수 있다. 그에 비해 NTAN의 네트워크에는 기존 DM → AGE_G → VA_LT, DM → AGE_G → HTN, DM → AGE_G → GLAUCOMA로 구성되던 관계에서 나이 그룹이 빠지고 DM → VA_LT, DM → HTN, DM → GLAUCOMA로 구성되어 자연스러운 관계가 형성되었다. NTAN의 네트워크를 통해 좌측 시력, 고혈압, 녹내장이 당뇨병에게 영향을 받고 우측 시력은 좌측 시력에게 영향을 받는다는 관계 해석을 도출해낼 수 있다.

6.3 혈액검사 데이터

혈액검사 데이터는 2014-2015년 일반검진 및 생애전환기 건강검진 데이터 1,000,000건으로 이루어진 데이터이다. 변수는 4개와 이산형 변수와 5개의 연속형 변수 총 9개의 변수로 이루어져 있고 각 변수의 의미와 값은 아래 표 11과 같다.

변수명	타입	의미	값
SEX	이산형	성별	여성: 1 남성: 0
AGE_G	이산형	나이 그룹	20 세 ~ 75 세 이상 (27 개 그룹)
ANE	이산형	빈혈	있음: 1 없음: 0
IHD	이산형	허혈심장질환	있음: 1 없음: 0
HGB	수치형	혈색소(g/dL) (헤모글로빈)	여성: 6.3 ~ 16.6 남성: 7.9 ~ 19
TCHOL	수치형	총 콜레스테롤(mg/dL)	여성: 98 ~ 372 남성: 90 ~ 380
TG	수치형	중성지방(mg/dL)	여성: 20 ~ 834 남성: 22 ~ 1506
HDL	수치형	HDL 콜레스테롤(mg/dL)	여성: 22 ~ 131 남성: 19 ~ 127
STK	이산형	뇌혈관 질환	있음: 1 없음: 0

표 11: 혈액검사 데이터 변수

변수명	의미	남성 정상범위	여성 정상범위
HGB	혈색소 (g/dL)	12 미만: 빈혈 12 ~ 12.9: 경미한 빈혈 13 ~ 16.5: 정상 16.5 초과: 혈색소 과다	10 미만: 빈혈 10 ~ 11.9: 경미한 빈혈 12 ~ 15.5: 정상 15.5 초과: 혈색소 과다
TCHOL	총 콜레스테롤 (mg/dL)	200 미만: 바람직 200 ~ 239: 약간 높음 240 이상: 높음	
TG	중성지방 (mg/dL)	150 미만: 바람직 150 ~ 199: 약간 높음 200 ~ 499: 높음 500 이상: 매우높음	
HDL	HDL 콜레스테롤 (mg/dL)	40 미만: 낮음 40 ~ 59: 보통 60 이상: 높음	

표 12: 각 변수별 정상 수치(서울대학교병원 건강칼럼)

또한 연속형 변수를 범주화 하기 위해 아래 서울대병원 건강칼럼에서 제공하는 정상 수치(표 12)를 참고하여 범주화 작업을 진행하였다.

또한 나이 그룹은 시력검사 데이터 때와 동일하게 생애 주기를 기준으로 20-29세(청년), 30-49세(중년), 50-64세(장년), 65-(노년) 4개의 범주로 축소하여 진행하였다. 9개의 변수 중 뇌혈관질환(STK)를 예측하고자 하는 클래스 변수로 선정하여 분류예측을 진행하였으며 시력검사 데이터에서와 마찬가지로 네트워크에서 제외할 변수집합 E는 (SEX, AGE_G)로 선정하였다. 그리고 변수집합 E를 뇌혈관질환(STK)변수에 포함시킨 class 변수를 생성한 뒤 NTAN과 NNB를 구현하고 분류예측을 진행하였다. 또한 TAN과 NB는 변수를 모두 사용하고 뇌혈관질환을 예측할 클래스 변수로 사용하여 분류예측을 진행하였다. TAN-II는 변수 집합 E를 네트워크에서 제외한 후 뇌혈관질환 유무를 예측한다.

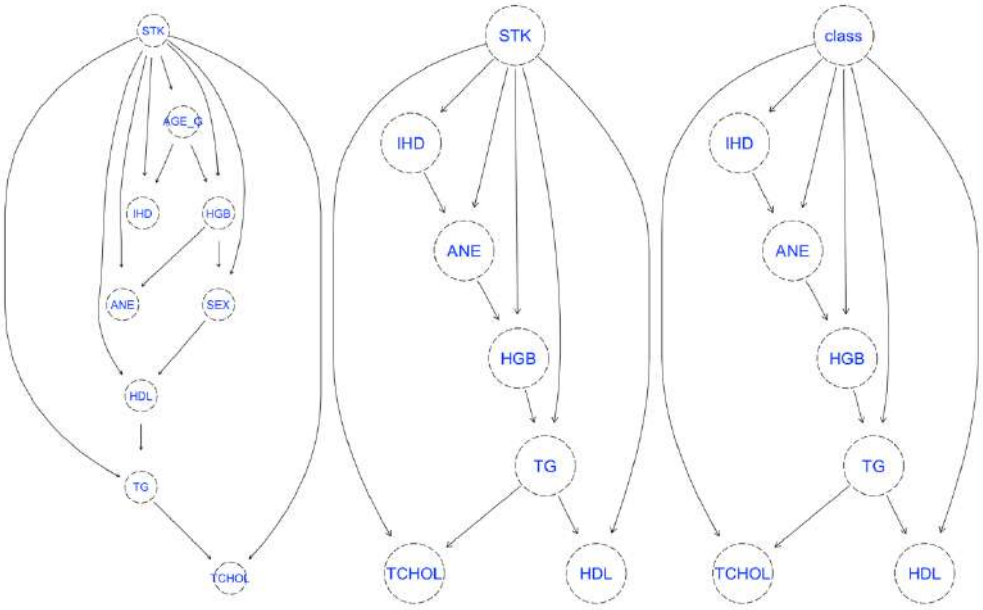


그림 11: 혈액검사 데이터의 네트워크: TAN(왼쪽), TAN-II(가운데), NTAN(오른쪽)

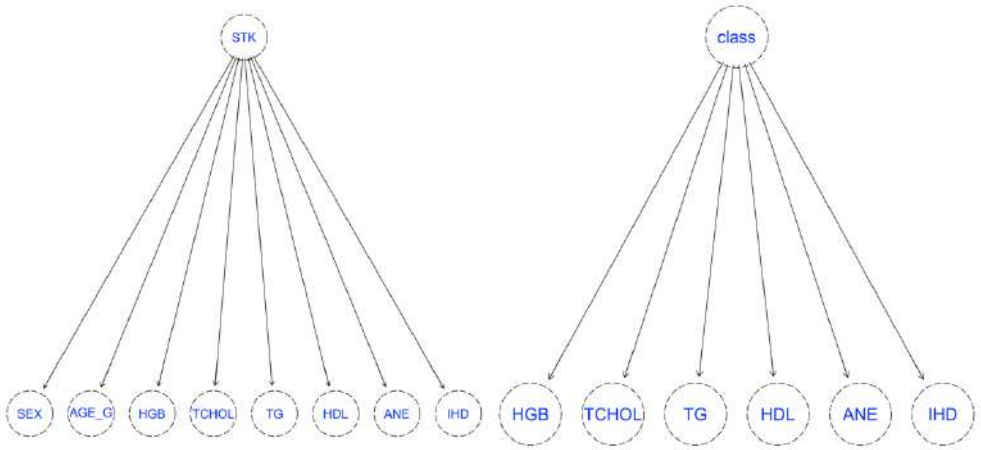


그림 12: 혈액검사 데이터의 네트워크: NB(왼쪽), NNB(오른쪽)

	ACC	AUC	Recall	Spec	Prec	F1	Time
TAN	0.769	0.775	0.779	0.609	0.968	0.863	1.79(s)
TAN-II	0.798	0.633	0.826	0.372	0.952	0.885	0.98(s)
NTAN	0.842†	0.634	0.876	0.329	0.952	0.913	2.28(s)
NB	0.783	0.768‡	0.8	0.529	0.962	0.874	0.18(s)
NNB	0.845*	0.63	0.88	0.324	0.952	0.914	0.74(s)
K-NN	0.751	0.735	0.761	0.603	0.966	0.851	76.98(h)
DT	0.779	0.779	0.787	0.664	0.972	0.87	2.74(s)
RF	0.77	0.785	0.777	0.665	0.972	0.864	1.02(h)

표 13: 혈액검사 데이터 분류성능 평가지표

표 13에서 NTAN의 정확도는 0.842로 0.769인 TAN과 0.798인 TAN-II에 비해 상당히 향상된 정확도를 보이고 있다. 이는 0.751인 K-NN, 0.779인 의사결정나무, 0.77인 랜덤포레스트와 비교하여도 높은 정확도를 보이고 있다. 그러나 NTAN의 AUC는 0.634로 0.775인 TAN에 비해 상당히 떨어지는 성능을 보이고 있다. 이는 특이도가 0.609인 TAN에 비해 NTAN의 특이도는 0.329까지 떨어진 것과 연관이 있으며 이는 K-NN, 의사결정나무, 랜덤포레스트에 비해 약

† 빨간색으로 표시된 수는 TAN, TAN-II, NTAN중 해당 평가지표가 가장 높은 수를 의미한다.

‡ 파란색으로 표시된 수는 NB, NNB중 해당 평가지표가 가장 높은 수를 의미한다.

* 밑줄친 수는 해당 평가지표에서 가장 높은 수를 의미한다.

0.3 떨어진 수치이다. 그러나 NTAN의 재현율은 0.876으로 TAN과 TAN-II에 비해 각각 약 0.1, 0.05가량 상승하였으며 0.761인 K-NN, 0.787인 의사결정나무, 0.777인 랜덤포레스트에 비해서도 매우 향상된 수치이다. NTAN의 정밀도는 0.952로 TAN-II, NNB와 더불어 모델 중 가장 낮은 수치를 기록하고 있으며, 0.972로 가장 높은 랜덤포레스트와 0.02 차이이다. 하지만 NTAN의 F1은 0.913으로 0.863인 TAN과 0.885인 TAN-II와 비교하여 향상된 성능을 보이고 있으며, 0.851인 K-NN과 비교하여 0.062 높은 수치이다. NNB의 경우 정확도 0.845를 기록하며 모델 중 가장 높은 정확도를 보이고 특히 0.783인 NB에 비해서 0.062 높은 성능을 보이고 있다. 그러나 AUC는 NTAN과 마찬가지로 기존의 방법인 NB보다 0.158 떨어지는 성능을 보이며 모델들 중 가장 낮은 수치를 기록하고 있다. 이와 더불어 NNB의 특이도 역시 NB에 비해 0.205 하락한 0.324를 보였으며 이는 모델 중 가장 낮은 수치이다. 하지만 재현율은 0.88로 NB에 비해 0.08, K-NN에 비해 0.119만큼 높으며 모델 중에 가장 높은 수치를 보이고 있다. 정밀도의 경우 0.952로 모델들 중 가장 낮았지만, 재현율과 정밀도의 조화평균인 F1은 0.914로 모델들 중 가장 높은 수치를 보이고 있다. 또한 NTAN과 NNB의 모델 적합시간은 약 2.28초, 0.74초로 K-NN과 랜덤포레스트에 비해 매우 빠르게 적합된다. 그러나 1.79초인 기존의 TAN, 0.18초인 NB와 비교하여 더 오래 걸리게 되는데, 이는 시력검사 데이터 때와 마찬가지로 학습할 모수의 수가 증가하기 때문에 나타나는 현상이다.

그림 11에서 TAN의 네트워크는 허혈성 심장질환(IHD), 헤모글로빈 수치(HGB)와 나이(AGE_G)간, 그리고 빈혈(ANE), 성별(SEX)과 헤모글로빈 수치간, 그리고 HDL 콜레스테롤 수치(HDL)와 성별간, 중성지방(TG)과 HDL

콜레스테롤 수치간, 그리고 총 콜레스테롤 수치(TCHOL)와 중성지방 수치간에 관계가 형성되었다. 이는 나이가 허혈성 심장질환과 헤모글로빈 수치에 영향을 주고, 성별이 HDL 콜레스테롤 수치에 영향을 주는 잘못된 해석을 초래한다. 그에 비해 NTAN의 네트워크는 기존 AGE_G → HGB → SEX → HDL로 구성되던 관계에서 성별과 나이 그룹이 빠지고 HGB → TG → HDL로 구성되어 부자연스러운 관계가 사라졌다. NTAN의 네트워크를 통해 빈혈과 헤모글로빈 수치간, 그리고 총 콜레스테롤 수치, HDL 콜레스테롤 수치와 중성지방간 관계가 성립함을 알 수 있다. 마찬가지로 그림 12에서 NNB의 네트워크는 NB의 네트워크에서의 나이 그룹과 성별이 사라지며 자연스러운 네트워크가 형성되었다.

6.4 UCI 유방암 재발 여부 데이터

UCI에서 제공하는 유방암 재발 여부 데이터는 총 286개의 데이터로 이루어져 있으며, 그 중 9개의 데이터에 결측값이 존재한다. 변수는 총 10개로 이루어져 있으며 각 변수명과 의미는 아래 표 14와 같다. UCI 데이터는 모든 변수가 범주형 변수이기 때문에 범주화 작업은 따로 진행하지 않는다. 나이 그룹의 경우 생애 주기를 기준으로 20-29세(청년), 30-49세(중년), 50-64세(장년), 65-(노년) 4개의 범주로 나누려 했으나 20대에 속하는 데이터가 1개 뿐이고 나이 그룹이 10세를 기준으로 나누어져 있어 생애 주기를 기준으로 범주화가 불가능 하였다. 따라서 UCI 데이터의 경우 20세-49세(청년+중년), 50대, 60대 이상 총 3개의 그룹으로 나누어 진행하였다. 또한 결측값이 포함된 데이터의 양이 전체 데이터 수에 비해 적으므로, 결측값이 포함된 데이터는 제거하였다. 데이터의 수가

충분하지 않기 때문에 전체 데이터를 3-fold로 분할하여 2:1의 비율로 훈련 데이터와 평가 데이터로 나눈다. 예측 변수는 Cancer 변수이며 나머지 변수들의 관계를 이용하여 분류예측을 진행한다. 또한 네트워크에서 제외할 변수집합 E는 age로 선정하였으며 변수집합 E를 유방암 재발 여부(Cancer)변수에 포함시킨 class 변수를 생성하여 분류예측을 진행하였다.

변수명	타입	의미	값
Cancer	이산형	유방암 재발여부	있음: 1 없음: 0
age	이산형	연령대	20 대 ~ 70 대 (6 개 그룹)
Menopause	이산형	진단시 환자의 폐경시기	0: 40 세 미만 1: 40 세 이상 2: premeno
tumor.size	이산형	최대 절제 암 크기 직경(mm)	0 ~ 59 (12 개 그룹)
inv.nodes	이산형	림프 노드 개수	0 ~ 39 (13 개 그룹)
node.caps	이산형	림프 노드 캡슐 전이성 암이 포함되었는지 여부	있음: 1 없음: 0
deg.malig	이산형	암의 조직학적 등급 분류 (높을수록 악성)	0: 1 등급 1: 2 등급 2: 3 등급
breast	이산형	유방암 발생 위치	0: 왼쪽 1: 오른쪽
breast.quad	이산형	유두를 중심으로 4 분면하여 유방암이 발생한 위치	0: 좌측 상단 1: 좌측 하단 2: 중앙 3: 우측 상단 4: 우측 하단
Irradiat	이산형	환자의 방사선 치료 여부	0: 받지 않음 1: 받음

표 14: UCI 데이터 변수

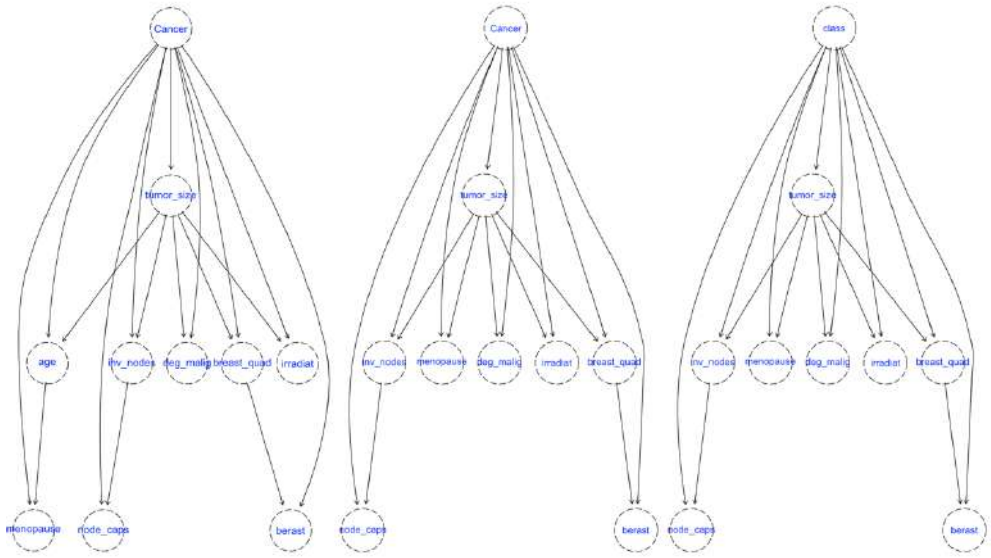


그림 13: UCI 데이터의 네트워크: TAN(왼쪽), TAN-II(가운데), NTAN(오른쪽)

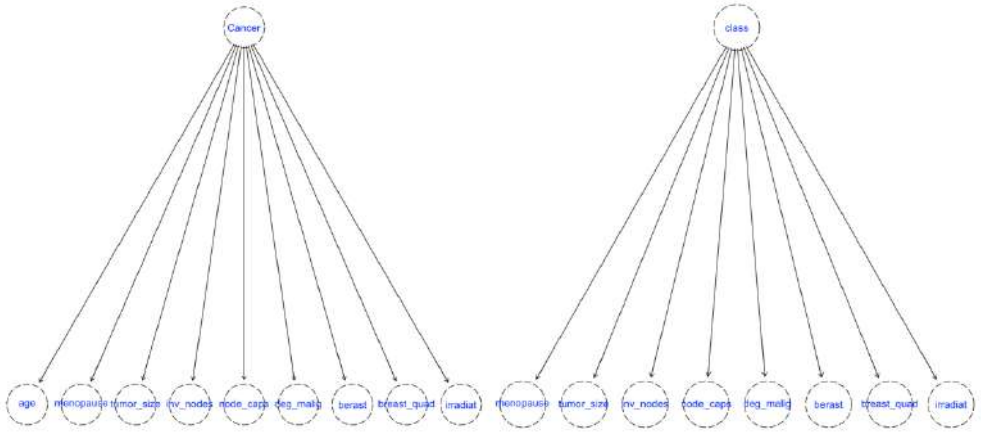


그림 14: UCI 데이터의 네트워크: NB(왼쪽), NNB(오른쪽)

	ACC	AUC	Recall	Spec	Prec	F1	Time(s)
TAN	0.638	0.664	0.665	0.573	0.791	0.72	0.17
TAN-II	0.641	0.679	0.656	0.607	0.801	0.719	0.15
NTAN	0.648†	0.687	0.677	0.579	0.796	0.73	0.26
NB	0.703	0.724	0.732	0.632‡	0.828	0.775	0.01
NNB	0.707*	0.736	0.753	0.595	0.818	0.783	0.02
K-NN	0.679	0.676	0.736	0.538	0.799	0.764	3.02
DT	0.648	0.636	0.689	0.546	0.791	0.732	0.01
RF	0.616	0.695	0.587	0.691	0.826	0.684	4.48

표 15: UCI 데이터 분류성능 평가지표

표 15에서 NTAN의 분류 성능이 TAN, TAN-II와 비교하여 상승되는 것을 볼 수 있다. NTAN의 정확도는 0.648로 0.638인 TAN과 0.641인 TAN-II보다 좋은 성능을 보인다. 의사결정나무와 동일한 정확도를 보이고 랜덤포레스트보다 0.032높은 정확도를 보이지만, K-NN에 비해서는 0.031만큼 떨어지는 정확도를 보이고 있다. 그리고 AUC는 0.687로 TAN과 TAN-II에 비해 각각 0.023,

† 빨간색으로 표시된 수는 TAN, TAN-II, NTAN중 해당 평가지표가 가장 높은 수를 의미한다.

‡ 파란색으로 표시된 수는 NB, NNB중 해당 평가지표가 가장 높은 수를 의미한다.

* 밑줄친 수는 해당 평가지표에서 가장 높은 수를 의미한다.

0.008만큼 향상된 성능을 보이며 0.676인 K-NN과 0.636인 의사결정나무에 비해서도 더 높은 성능을 보이지만, 0.695인 랜덤포레스트에 비해서는 조금 낮은 성능을 보이고 있다. 또한 재현율은 NTAN이 0.677, TAN이 0.665, TAN-II이 0.656으로 NTAN이 가장 높은 성능을 보여주고 있으나, K-NN과 의사결정나무에 비해서는 떨어지는 성능을 보여주고 있다. 그러나 NTAN의 특이도는 0.579로 0.573인 TAN보다 높지만 0.607인 TAN-II보다 떨어지는 값을 보여주고 있다. 또한 K-NN이 0.538로 가장 낮은 값을 보이고, 랜덤포레스트는 0.691로 가장 높은 값을 보이고 있다. 또한 정밀도의 경우에도 NTAN은 0.796으로 TAN보다 0.005 높지만 TAN-II보다는 0.005 낮은 성능을 보이고 있으나 0.826인 랜덤포레스트보다는 0.03 낮은 정밀도를 보이고 있다. NTAN의 경우 정밀도와 재현율의 조화 평균인 F1이 0.73으로 0.764인 K-NN, 0.732인 의사결정나무보다 떨어지는 성능을 보이지만, 랜덤포레스트보다 0.046만큼 높으며 TAN, TAN-II보다 약 0.01가량 높은 성능을 보이고 있다. NNB의 경우 정확도가 0.707으로 모델 중 가장 높은 정확도를 보이고 특히 랜덤포레스트에 비해 0.091 높은 정확도를 보이고 있다. AUC도 마찬가지로 0.736으로 모델 중 가장 높고 NB보다 0.012, 의사결정나무보다 0.088 높은 성능을 보이고 있다. 또한 재현율은 0.753으로 모델 중 가장 높은 재현율을 보이며 랜덤포레스트에 비해 0.149 높은 재현율을 보이고 있다. 그러나 NNB의 특이도는 0.595로 0.632인 NB보다 낮은 성능을 보이며 특히 랜덤포레스트보다 약 0.1만큼 낮은 성능을 보인다. 정밀도의 경우 NB가 0.828로 모델 중 가장 높은 정밀도를 보이고 있다. NNB는 0.818로 K-NN과 의사결정나무보다 높으나 0.826인 랜덤포레스트보다 낮고 모델 중 3번째로 높은 정밀도를 보인다. 하지만 NNB의 F1은 0.783으로 모델 중 가장

높은 성능을 보이고 있다. NTAN과 NNB는 기존 방식의 TAN, TAN-II, 그리고 NB와 비교하였을 때 특이도와 정밀도를 제외한 모든 지표에서 성능이 향상되었다. 하지만 NTAN은 NNB과 비교하였을 때 큰 성능 차이를 보인다. 특히 정확도는 약 0.06, AUC는 약 0.05, F1은 약 0.06만큼 떨어지는 성능을 보이고 있다. 이는 시뮬레이션 I에서 살펴보았듯 표본의 수가 적은 경우에 나타나는 NTAN의 단점으로 인해 발생하는 차이이다. 그에 비해 NB와 NNB는 표본의 수가 적은 경우에도 다른 기법들과 비교하여 높은 정확도, AUC 그리고 F1을 보이고 있다.

그림 13에서 TAN의 네트워크에 최대 절제 암 크기 직경(tumor_size)와 나이(age)간, 그리고 나이와 폐경시기(menopause)간에 관계가 성립되는 네트워크가 형성되었다. 이는 나이가 최대 절제 암 크기 직경에 영향을 받으며, 폐경 시기가 나이에 영향을 받는다는 잘못된 해석을 초래한다. 그에 비해 NTAN의 네트워크는 기존에 tumor_size → age → menopause로 형성되던 인과관계에서 나이(age)가 빠지고 tumor_size → menopause로 자연스러운 관계가 형성되었다. NTAN의 네트워크를 통해 림프 노드 개수(inv_nodes), 폐경시기(menopause), 암의 조직학적 등급(deg_malig), 방사선 치료 여부(irradiat), 유방암 발생 위치(breast_quad)는 최대 절제 암 크기 직경(tumor_size)에 영향을 받으며, 림프 노드의 전이성 암 여부(nodes_caps)는 림프 노드 개수(inv_nodes)에 영향을 받는다고 해석 할 수 있다.

제 7 장 결론

본 논문에서는 최근 다양한 분야에서 사용되고 있는 분류기법인 베이지안 네트워크 분류기 중 나이브 베이즈(NB)와 TAN(Tree-Augmented Naive Bayes)을 소개하였다. 그 중 TAN은 나이브 베이즈의 강한 가정을 완화하고 변수간 관계를 시각화하는 특징으로 인해 많이 사용되지만, 네트워크 구축시 인과관계에 포함되지 말아야 할 변수가 포함되어 부자연스러운 네트워크가 형성되는 문제점이 있다. 이러한 문제점을 해결하기 위해 본 논문에서 NTAN과 NNB을 제안하고 시뮬레이션과 실제 데이터에 적용하여 분류성능을 평가하고 형성된 네트워크를 이용하여 변수간 관계를 해석하였다.

먼저 시뮬레이션 I을 통해 표본의 수가 적은 경우인 N = 100인 경우 TAN, TAN-II, NTAN의 정확도, AUC, 재현율, 특이도, 정밀도, F1, 매튜 상관계수가 NB와 NNB에 비해 모두 낮아 분류성능이 NB와 NNB에 비해 떨어지는 것을 확인하였다. 이는 모수의 수에 비해 표본의 수가 적어 모수 학습이 제대로 이루어지지 않아 발생하는 문제이다. 그러나 이 때 NTAN의 정확도는 TAN에 비해 0.006, AUC는 0.019, 재현율은 0.07, 특이도는 0.06, 정밀도는 0.05, F1은 0.007, 매튜 상관계수는 0.003 높아 표본의 수가 적은 경우에 TAN보다 분류성능이 좋음을 확인하였다. 그에 비해 NB와 NNB는 학습할 모수의 수가 상대적으로 적어 표본이 적은 경우에도 좋은 분류성능을 보여주었다. 표본의 수가 증가하면서 모수 학습이 충분히 이루어지며 표본의 수가 1,000개를 넘는 시점부터

TAN, TAN-II, NTAN의 모든 평가 지표가 NB, NNB보다 높아져 분류성능이 더 좋아짐을 확인하였다. N=100인 경우와 N=50,000인 경우를 비교하여 NTAN의 정확도는 0.056, AUC는 0.011, 재현율은 0.061, 특이도는 0.049, 정밀도는 0.051, F1은 0.062, 매튜 상관계수는 0.111만큼 상승하였다. 그리고 TAN의 경우 정확도는 0.061, AUC는 0.03, 재현율은 0.068, 특이도는 0.055, 정밀도는 0.056, F1은 0.069, 매튜 상관계수는 0.114만큼 상승하였다. 이를 통해 TAN과 NTAN이 표본의 수에 영향을 받는다는 것을 알 수 있다. 그에 비해 NB의 정확도는 0.01, AUC는 0.006, F1은 0.017 상승하였으며 NNB의 경우에는 정확도는 0.021, AUC는 0.004, F1은 0.028 상승하였다. NB와 NNB는 TAN과 NTAN에 비해 표본의 수에 덜 민감하다는 것을 알 수 있다. 그리고 NNB의 경우 표본의 수에 관계없이 항상 모든 평가지표가 NB보다 높아 분류성능이 더 좋음을 알 수 있다. 표본의 수가 5,000개 이상인 경우 NTAN과 TAN의 분류성능은 비슷하지만, NTAN의 네트워크 유사도가 TAN의 네트워크 유사도보다 높은 것을 확인하였으며 특히 표본의 수가 50,000개인 경우에 NTAN의 네트워크가 실제 변수간 인과관계를 거의 완벽히 재현하는 것을 확인하였다. 이는 NTAN의 네트워크가 기존의 TAN의 네트워크보다 단순하지만 실제 인과관계는 더 잘 표현하면서 분류성능은 비슷하므로 NTAN이 TAN보다 효율적인 모델임을 의미한다.

시뮬레이션 II에서는 표본의 수가 충분히 확보된 경우 변수의 개수에 따른 분류성능과 네트워크 유사도를 평가하였다. 이를 통해 변수의 개수가 많고 관계가 많을수록 TAN과 NTAN의 성능이 향상되는 것을 보았다. 변수가 10개인 경우와 70개인 경우를 비교하여 TAN과 NTAN의 정확도는 각각 0.082, 0.084 상승하고 AUC는 0.156, 0.157 상승하였다. NTAN의 분류성능은 TAN과 비교하여 같거나

우수한 성능을 보여주며, 네트워크 형성 시 NTAN이 더 자연스러운 네트워크를 형성하였다. 이는 NTAN의 네트워크 유사도가 TAN과 TAN-II의 네트워크 유사도보다 높고 NTAN의 네트워크에 변수집합 E가 제외된 네트워크가 생성됨을 확인함으로써 입증하였다. 그리고 NNB의 경우 NB보다 성능이 향상되는 것을 보았다. NNB가 NB에 비해 정확도 면에서 변수가 10개인 경우에는 0.014, 25개인 경우에는 0.019, 40개인 경우에는 0.046, 70개인 경우에는 0.077 향상되었고 전체적인 평가지표가 상승하여 더 좋은 분류성능을 나타낸다. NB는 변수들의 독립성 가정으로 인해 변수간 관계를 사용하지 않았으나, NNB의 경우 제외시킬 변수집합 E를 클래스 변수에 포함시킴으로써 변수간 관계를 사용하여 향상된 성능을 얻어낼 수 있었다.

그리고 국민건강보험공단에서 제공하는 건강검진 데이터인 시력검사 데이터, 혈액검사 데이터와 UCI 데이터인 유방암 재발 여부 데이터를 통해 TAN과 NTAN을 비교하여 NTAN의 네트워크가 자연스러운 결과 해석을 도출해낼 뿐 아니라 분류성능이 향상됨을 확인하였다. 먼저 시력검사 데이터에서 기존의 TAN 네트워크는 나이와 당뇨병간, 고혈압, 좌측 시력, 녹내장과 나이간, 성별과 우측 시력간 관계가 형성되었다. 이는 나이가 당뇨병에 영향을 받고, 성별이 우측 시력에 영향을 받는 등 부적절한 관계 해석을 야기한다. 반면 NTAN의 경우 나이와 성별이 네트워크에서 제외되어 좌측 시력, 고혈압, 녹내장과 당뇨병간, 그리고 우측 시력과 좌측 시력간 관계가 형성되는 자연스러운 네트워크가 형성되었다. 이는 우측 시력이 좌측 시력에 영향을 받고 고혈압이 당뇨병에 영향을 받는다 등의 해석을 할 수 있다. NTAN의 정확도는 TAN에 비해 0.021 상승하였으며, F1의 경우 0.013, AUC는 0.004 상승하였다. K-NN, 의사결정나무, 랜덤포레스트에

비해서는 각각 0.031, 0.024, 0.028 높은 정확도를 보이고 있다. 그리고 NNB의 경우 NB보다 정확도가 0.066 상승하였으며, F1은 0.042 상승하였다. 또한 NNB는 모델 중에서 가장 높은 정확도, AUC, 재현율, F1을 보인다. 이를 통해 NTAN과 NNB가 기존의 TAN과 NB에 비해서도, 다른 머신러닝 기법들에 비해서도 좋은 분류성능을 보임을 확인하였다.

혈액검사 데이터의 경우 기존의 TAN 네트워크에서 허혈성 심장질환, 헤모글로빈 수치와 나이간, 빈혈, 성별과 헤모글로빈 수치간, 그리고 성별과 HDL 콜레스테롤간, HDL 콜레스테롤과 중성지방간, 중성지방과 총 콜레스테롤간 관계가 형성되었다. 하지만 이는 허혈성 심장질환과 헤모글로빈 수치가 나이에게 영향을 받고, 성별이 헤모글로빈 수치에게 영향을 받는 등 부적절한 관계 해석을 할 수 있다. NTAN 네트워크의 경우 나이와 성별이 네트워크에서 사라지며 허혈성 심장질환 → 빈혈 → 헤모글로빈 수치 → 중성지방 형태로 관계가 형성되고 총 콜레스테롤, HDL 콜레스테롤과 중성지방간 관계가 형성되었다. 이러한 네트워크를 통해 헤모글로빈 수치가 빈혈에게, HDL 콜레스테롤이 중성지방에 영향을 받는다 등의 해석을 할 수 있게 되었다. 분류성능 면에서 NTAN의 정확도는 TAN의 정확도에 비해 0.073 상승하였으며 F1은 0.05 상승하였다. 또한 K-NN, 의사결정나무, 랜덤포레스트에 비해 정확도는 0.091, 0.063, 0.072 높고 F1은 0.062, 0.043, 0.049 높다. 하지만 TAN에 비해 AUC는 0.141, 특이도는 0.28 떨어지는 성능을 보이고 있다.

UCI에서 제공하는 유방암 재발 여부 데이터의 경우 기존의 TAN 네트워크에서 최대 절제 암 크기 직경 → 나이 → 폐경시기로 관계가 형성되었다. 이는 나이가 최대 절제 암 크기 직경에 영향을 받고, 폐경 시기가 나이에 영향을 받는

부자연스러운 해석을 초래한다. 그에 비해 NTAN은 네트워크에서 나이 변수가 사라져 최대 절제암 크기 직경 → 폐경시기로 관계가 형성되었다. NTAN의 네트워크를 통해서 림프 노드 캡슐 전이성 암 여부가 림프 노드 개수에 영향을 받고, 암의 조직학적 등급이 최대 절제 암 크기 직경에 영향을 받는다는 등의 해석이 가능하다. 분류성능 면에서 NTAN은 TAN에 비해 정확도가 0.01, AUC가 0.023, 재현율이 0.012, 특이도가 0.006, 정밀도가 0.005, F1이 0.01 상승하였다. 랜덤포레스트에 비해서 0.032 높은 정확도를 보여주지만, K-NN에 비해 0.031 떨어지는 정확도를 보이고 있다. 그리고 NNB는 NB에 비해 정확도가 0.004, AUC는 0.012, 재현율은 0.021, F1은 0.008 높으나, 특이도는 0.037, 정밀도는 0.01 낮은 수치를 보이고 있다. UCI 데이터의 경우 NTAN과 TAN의 성능이 NNB와 NB에 비해 떨어지는 성능을 보이고 있다. NTAN은 NNB에 비해 정확도는 0.069, AUC는 0.072, F1은 0.063만큼 떨어지는 성능을 보이고 있다. 이는 시뮬레이션 I에서 확인하였듯이 표본의 수가 학습할 모수에 비해 적어 나타나는 NTAN과 TAN의 단점이다.

건강검진 데이터와 UCI 데이터를 통해 NTAN이 TAN, TAN-II보다 높은 정확도, 재현율, 정밀도, F1을 나타내고 TAN에 비해 자연스러운 네트워크를 형성하면서 적절한 관계해석이 가능하다는 장점이 있음을 확인하였다. NNB의 경우에도 NB에 비해 높은 정확도를 보여주고 있다. 그러나 혈액검사 데이터에 적용하였을 때 NTAN과 NNB의 특이도가 기존의 TAN과 NB에 비해 매우 낮아지는 현상이 발생하였다. 그리고 NTAN과 TAN의 경우 표본의 수가 적은 경우에 NB에 비해 분류성능이 떨어지는 단점이 존재하였는데, 이를 해결하기 위한 후속 연구가 필요하다고 보인다.

참 고 문 헌

Breiman, L. (2001). Random forests. *Machine learning*, 45, 5-32.

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

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321-357.

Chow, C. K., & Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory*, 14, 462-467.

Dal Pozzolo, A., Caelen, O., Bontempi, G. (2015). unbalanced: Racing for Unbalanced Methods Selection. R package version 2.0.

Dana, A. L.-D., & Alashqur, A. (2014). Using decision tree classification to assist in the prediction of Alzheimer's disease. 2014 6th *international conference on computer science and information technology (CSIT)*, (pp. 122-126).

Duda, R. O., Hart, P. E., & Stork, D. G. (1973). *Pattern classification and scene analysis* (Vol. 3). Wiley New York.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine learning*, 29, 131-163.

Gakii, C., & Jepkoech, J. (2019). A classification model for water quality analysis using decision tree. *European Journal of Computer Science and Information Technology, 7(3)*, pp.1-8

Gu, Q., Cai, Z., Zhu, L., & Huang, B. (2008). Data mining on imbalanced data sets. *2008 International Conference on advanced computer theory and engineering*, (pp. 1020-1024).

Hansen K.D., Gentry J, Long L, Gentleman R, Falcon S, Hahne F, Sarkar D (2022). Rgraphviz: Provides plotting capabilities for R graph objects. R package version 2.40.0.

Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5, 1.

Islam, M. R., Kamal, A. R. M., Sultana, N., Islam, R., & Moni, M. A. (2018). Detecting depression using k-nearest neighbors (knn) classification technique. *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*, (pp. 1-4).

Jiang, L., Zhang, H., Cai, Z., & Su, J. (2005). Learning tree augmented naive bayes for ranking. *International Conference on Database Systems for Advanced Applications,* (pp. 688-698).

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 29*, 119-127.

Kumar, M., Hanumanthappa, M., & Kumar, T. S. (2012). Intrusion detection system using decision tree algorithm. *2012 IEEE 14th international conference on communication technology*, (pp. 629-634).

Lam, W., & Bacchus, F. (1994). Using new data to refine a Bayesian network. In *Uncertainty Proceedings 1994* (pp. 383-390). Elsevier.

Langley, P., Iba, W., & Thompson, K. (1992). An analysis of Bayesian classifiers. *American Association for Artificial Intelligence, 90*, pp. 223-228.

Liaw, A., & Wiener, M (2002). Classification and regression by randomForest. *R news, 2*, 18-22.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1), 21-27.

Mihaljevic, B., Bielza Lozoya, M. C., & Larrañaga Múgica, P. M. (2018). Bnclassify: learning bayesian network classifiers. *R JOURNAL, 10,* 455-468.

Moldagulova, A., & Sulaiman, R. B. (2017). Using KNN algorithm for classification of textual documents. *2017 8th International Conference on Information Technology (ICIT)*, (pp. 665-671).

Najafi, R., & Afsharchi, M. (2012). Network intrusion detection using tree augmented naive-bayes. *The Third International Conference on Contemporary Issues in Computer and Information Sciences (CICI)*, (pp. 396-402).

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1, 81-106.

Quinlan, J. R. (2014). *C4. 5: programs for machine learning*. Elsevier.

Vapnik, V. (1999). *The nature of statistical learning theory*. Springer science & business media.

Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

Ripley,B., Venables, W., & Ripley, M. B. (2015). *Modern Applied Statistics with S*, Fourth edition. Springer, New York. ISBN 0-387-95457-0,

Sharma, M., Singh, S. K., Agrawal, P., & Madaan, V. (2016). Classification of clinical dataset of cervical cancer using KNN. *Indian Journal of Science and Technology, 9,* 1-5.

Stephenson, T. A. (2000). *An introduction to Bayesian network theory and usage*. Research Report 00-01, IDIAP.

Sun, A., Lim, E.-P., & Ng, W.-K. (2002). Web classification using support vector machine. *Proceedings of the 4th international workshop on Web information and data management,* (pp. 96-99).

Suzuki, J. (1993). A construction of Bayesian networks from databases based on an MDL principle. *Uncertainty in Artificial Intelligence*, (pp. 266-273).

Therneau, T., Atkinson, B., and Ripley, B. (2015). rpart: Recursive Partitioning and Regression Trees. URL http://CRAN.R-project.org/package=rpart. R package version 4.1-10.

Xuan, S., Liu, G., Li, Z., Zheng, L., Wang, S., & Jiang, C. (2018). Random forest for credit card fraud detection. *2018 IEEE 15th international conference on networking, sensing and control (ICNSC)*, (pp. 1-6).

김현미, & 정성환. (2013). 망막 질환 진단을 위한 베이지안 네트워크에 기초한 데이터 분석. “Journal of Korea Multimedia Society”, 16, 269-280.

정용규, & 김인철. (2002). 베이지안 망에 기초한 불임환자 임상데이터의 분석. “정보처리학회논문지 B, 제 9-B 권”, 625-633.

홍종선, 오세현, & 최예원. (2022). 혼동행렬의 상관계수를 이용한 최적분류점. 응용통계연구, 35(1), 77-91.

부 록

<부록 1> p=40 인 경우

$Y \sim B(P = 0.5)$, $\quad \text{sex} \sim B(P = 0.5)$,

$x_1 \left| Y \sim \text{B}\left(\text{P} = \frac{\exp(-1.5 + 3 \times Y)}{1 + \exp(-1.5 + 3 \times Y)}\right)\right.$, $\quad x_2 \left| x_1 \sim B\big(P = \Phi(-1 + 2 \times x_1)\big)\right.$,

$x_3 | x_1 \sim N(\mu = 10 + 5 \times x_1,\ \text{sd} = 2)$, $\quad \text{age} | x_2 \sim B\left(P = \frac{\exp(-1 + 2 \times x_2)}{1 + \exp(-1 + 2 \times x_2)}\right)$,

$x_4 | x_2 \sim N(\mu = 50 + 15 \times x_2,\ \text{sd} = 3)$,

$x_5 | \text{age}, x_2 \sim B\left(P = \frac{\exp(1.5 - 3 \times \text{age} - x_2)}{1 + \exp(1.5 - 3 \times \text{age} - x_2)}\right)$,

$x_6 | x_3 \sim N(\mu = 100 + 1.5 \times x_3,\ \text{sd} = 5), x_7 | \text{age}, x_2 \sim B\big(P = \Phi(1 - 2 \times \text{age} - 0.5 \times x_2)\big)$,

$x_8 \left| Y \sim B\left(P = \frac{\exp(-1 + 2 \times Y)}{1 + \exp(-1 + 2 \times Y)}\right)\right.$, $\quad x_9 \left| \text{age}, x_2 \sim B\big(P = \Phi(-1.5 + 3 \times \text{age} - x_2)\big)\right.$,

$x_{10} | x_8 \sim N(\mu = 30 + 3 \times x_8,\ \text{sd} = 1.5)$, $\quad x_{11} | x_9 \sim N(\mu = 15 + 2 \times x_9\ ,\text{sd} = 1)$,

$x_{12} | x_{10} \sim N(\mu = 50 + 1.5 \times x_{10},\ \text{sd} = 2)$, $\quad x_{13} | x_9 \sim B\big(P = \Phi(-1 + 2 \times x_9)\big)$,

$x_{14} \left| x_{13} \sim B\left(P = \frac{\exp(-1 + 2 \times x_{13})}{1 + \exp(-1 + 2 \times x_{13})}\right)\right.$, $\quad x_{15} \left| Y \sim B\left(P = \frac{\exp(-0.5 + Y)}{1 + \exp(-0.5 + Y)}\right)\right.$,

$x_{16} | \text{age}, x_2 \sim B\big(P = \Phi(2 - 4 \times \text{age} - x_2)\big)$, $\quad x_{17} | x_{15} \sim N(\mu = 50 - 3 \times x_{15},\ \text{sd} = 2)$,

$x_{18} | x_{16} \sim B\big(P = \Phi(1 - 2 \times x_{16})\big)$, $\quad x_{19} | x_{16} \sim B\left(P = \frac{\exp(-1 + 2 \times x_{16})}{1 + \exp(-1 + 2 \times x_{16})}\right)$,

$x_{20} \left| x_{19} \sim B\left(P = \frac{\exp(-1.5 + 3 \times x_{19})}{1 + \exp(-1.5 + 3 \times x_{19})}\right)\right.$, $\quad x_{21} \left| x_{17} \sim N(\mu = 10 + x_{17},\ \text{sd} = 1)\right.$,

$x_{22} \left| x_{18} \sim B\left(P = \frac{\exp(-0.5 + x_{18})}{1 + \exp(-0.5 + x_{18})}\right)\right.$, $\quad x_{23} \left| Y \sim B\left(P = \frac{\exp(-0.5 + Y)}{1 + \exp(-0.5 + Y)}\right)\right.$,

$x_{24} | \text{age}, x_2 \sim B\big(P = \Phi(2 - 4 \times \text{age} - x_2)\big)$, $\quad x_{25} | x_{23} \sim N(\mu = 15 - 5 \times x_{23},\ \text{sd} = 1)$,

$x_{26}|x_{24} \sim N(\mu = 25 + 3 \times x_{24},\ \mathrm{sd} = 1)$, $x_{27}|x_{25} \sim N(\mu = 30 + 1.25 \times \mathrm{x}_{25},\ \mathrm{sd} = 1.5)$,

$x_{28}|x_{24} \sim B\big(P = \Phi(1 - 2 \times x_{24})\big)$, $x_{29}|x_{28} \sim B\big(P = \Phi(-1 + 2 \times x_{28})\big)$,

$$x_{30}\left|Y \sim B\left(P = \frac{\exp(1 - 2 \times Y)}{1 + \exp(1 - 2 \times Y)}\right), x_{31}\right|\mathrm{age}, x_2 \sim B\left(P = \frac{\exp(2 - 4 \times \mathrm{age} - x_2)}{1 + \exp(2 - 4 \times \mathrm{age} - x_2)}\right),$$

$x_{32}|x_{30} \sim N(\mu = 5 + 4 \times x_{30},\ \mathrm{sd} = 1)$, $x_{33}|x_{31} \sim B\big(P = \Phi(-1.5 + 3 \times x_{31})\big)$,

$$x_{34}|x_{31} \sim B\left(P = \frac{\exp(-1 + 2 \times x_{31})}{1 + \exp(-1 + 2 \times x_{31})}\right),$$

$$x_{35}|x_{34} \sim B\left(P = \frac{\exp(-1.5 + 3 \times x_{34})}{1 + \exp(-1.5 + 3 \times x_{34})}\right),$$

$$x_{36}|x_{32} \sim N(\mu = 11 - x_{32},\ \mathrm{sd} = 1), \quad x_{37}|x_{33} \sim B\left(P = \frac{\exp(0.5 - x_{33})}{1 + \exp(0.5 - x_{33})}\right).$$

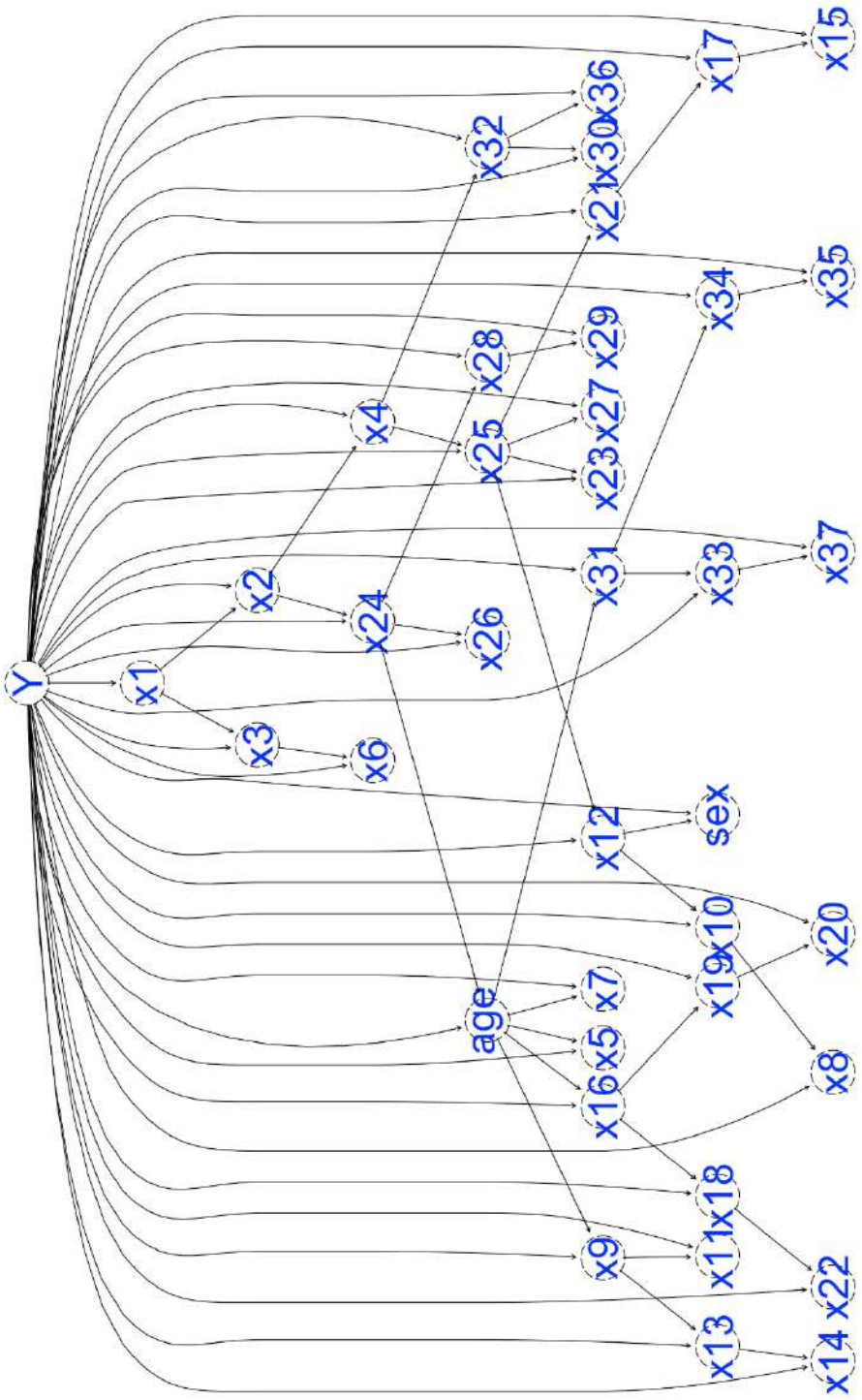


그림 부록1 – 1: p=40인 경우의 TAN

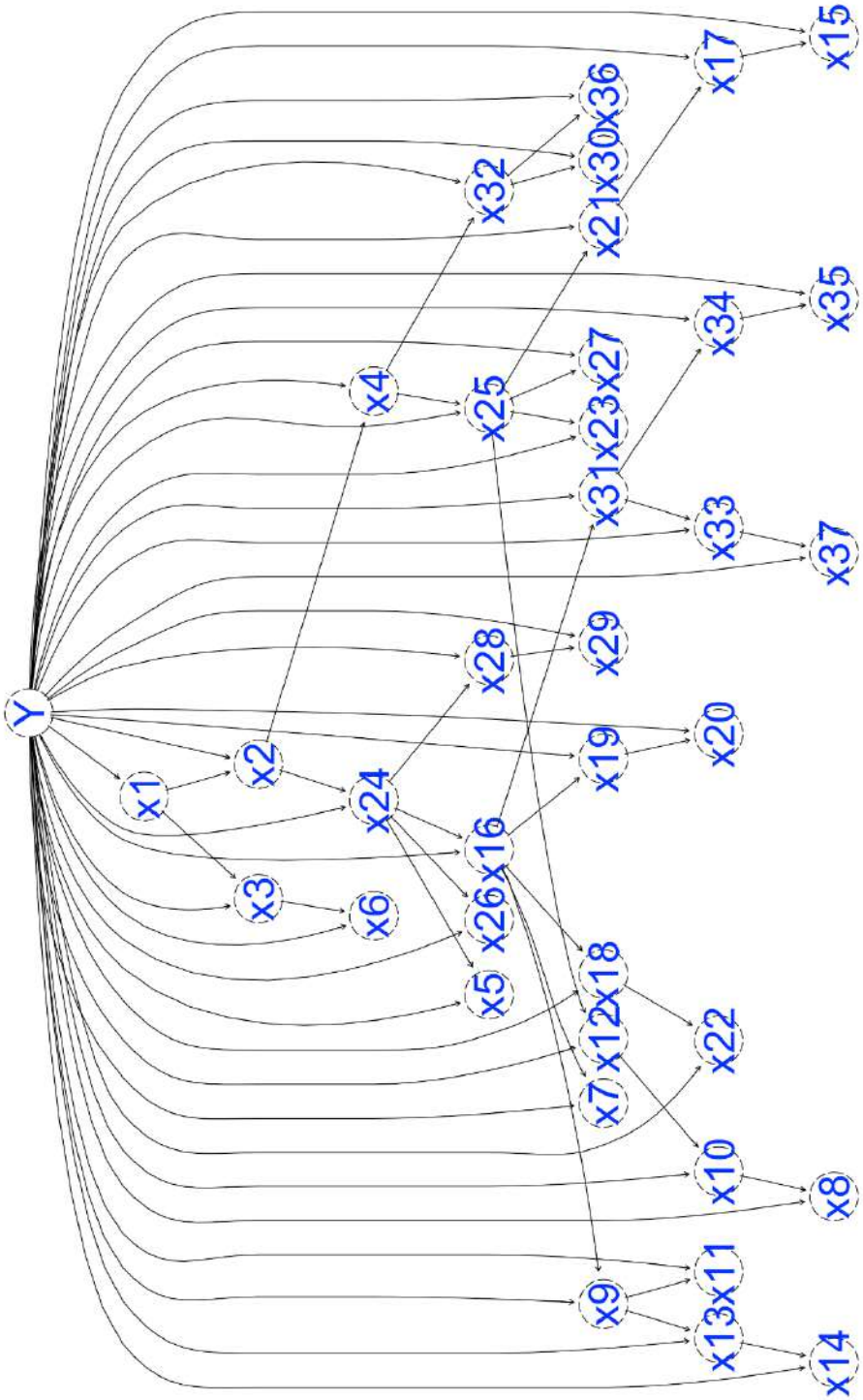


그림 부록1 – 2: p=40인 경우의 TAN-II

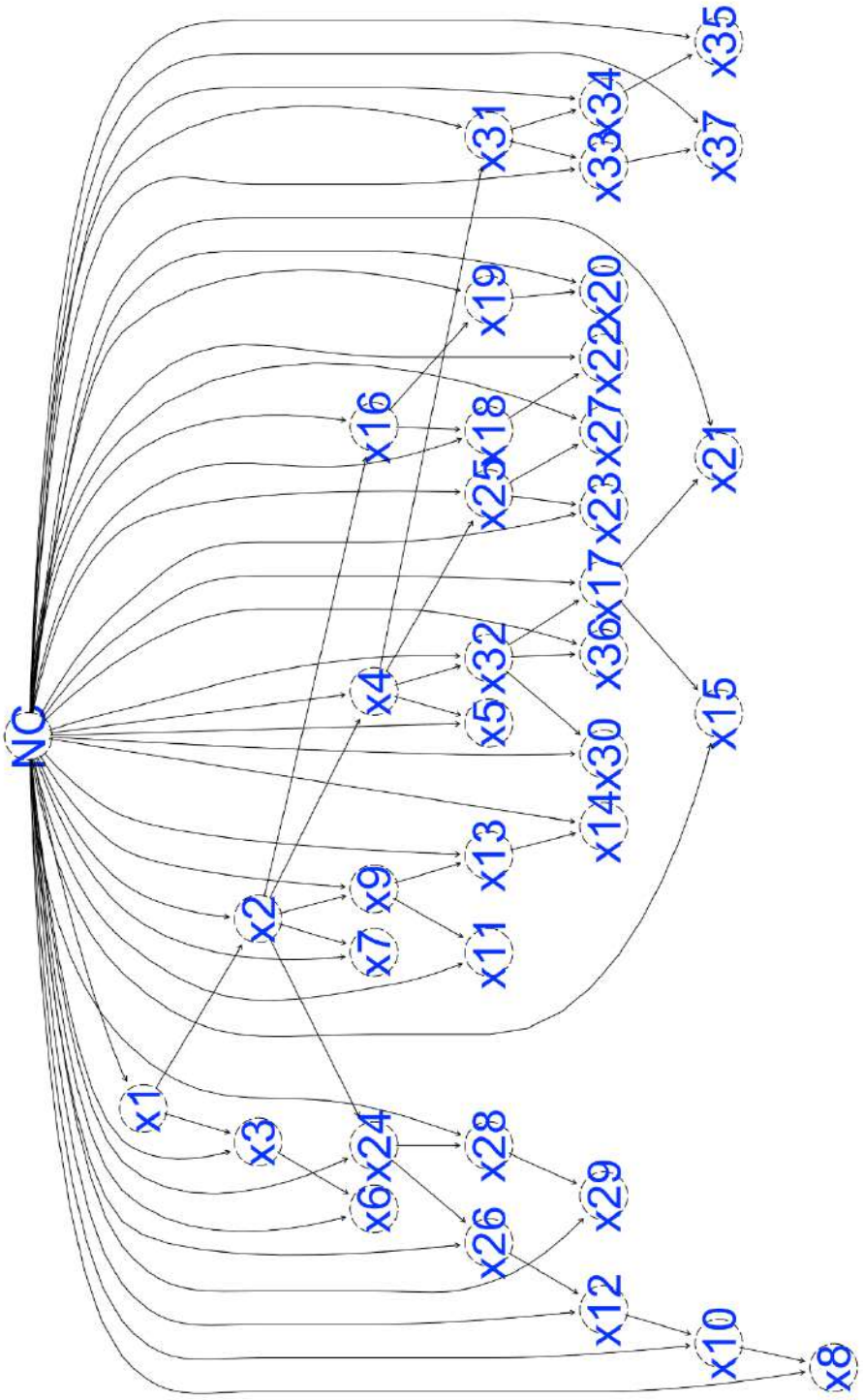


그림 부록1 – 3: p=40인 경우의 NTAN

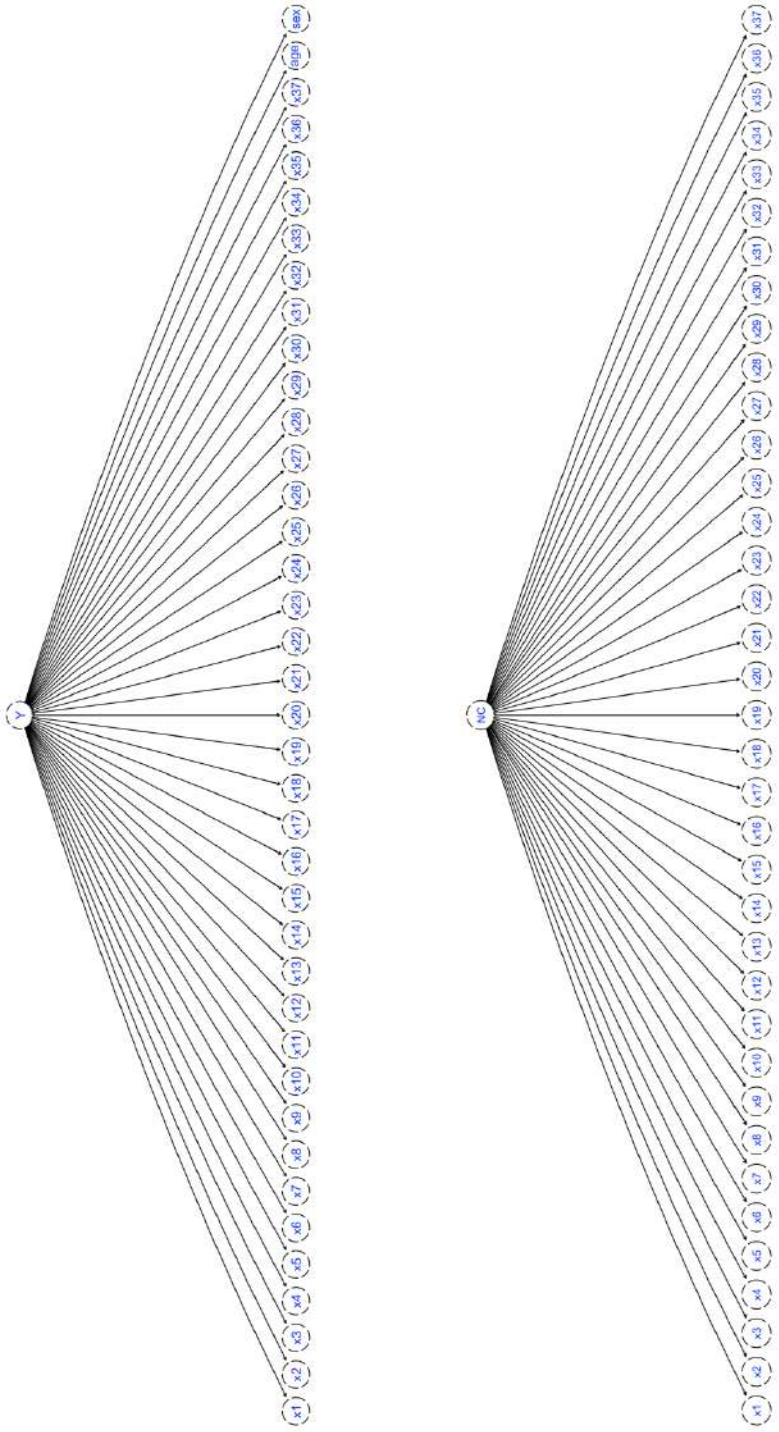


그림 부록1 – 4: p=40인 경우의 NB(위), NNB(아래)

<부록 2> p=70 인 경우

$Y \sim B(P = 0.5),\quad \text{sex} \sim B(P = 0.5),$

$x_1 \left| Y \sim \text{B}\left(\text{P} = \frac{\exp(-1.5 + 3 \times Y)}{1 + \exp(-1.5 + 3 \times Y)}\right),\quad x_2 \right| x_1 \sim B\big(P = \Phi(-1 + 2 \times x_1)\big),$

$x_3 | x_1 \sim N\big(\mu = 10 + 5 \times x_1,\ \text{sd} = 2\big),\quad \text{age} | x_2 \sim B\left(P = \frac{\exp(-1 + 2 \times x_2)}{1 + \exp(-1 + 2 \times x_2)}\right),$

$x_4 | x_2 \sim N\big(\mu = 50 + 15 \times x_2,\ \text{sd} = 3\big),$

$x_5 | \text{age}, x_2 \sim B\left(P = \frac{\exp(1.5 - 3 \times \text{age} - x_2)}{1 + \exp(1.5 - 3 \times \text{age} - x_2)}\right),$

$x_6 | x_3 \sim N\big(\mu = 100 + 1.5 \times x_3,\ \text{sd} = 5\big), x_7 | \text{age}, x_2 \sim B\big(P = \Phi(1 - 2 \times \text{age} - 0.5 \times x_2)\big),$

$x_8 \left| Y \sim B\left(P = \frac{\exp(-1 + 2 \times Y)}{1 + \exp(-1 + 2 \times Y)}\right),\quad x_9 \right| \text{age}, x_2 \sim B\big(P = \Phi(-1.5 + 3 \times \text{age} - x_2)\big),$

$x_{10} | x_8 \sim N\big(\mu = 30 + 3 \times x_8,\ \text{sd} = 1.5\big),\quad x_{11} | x_9 \sim N\big(\mu = 15 + 2 \times x_9,\ \text{sd} = 1\big),$

$x_{12} | x_{10} \sim N\big(\mu = 50 + 1.5 \times x_{10},\ \text{sd} = 2\big),\quad x_{13} | x_9 \sim B\big(P = \Phi(-1 + 2 \times x_9)\big),$

$x_{14} \left| x_{13} \sim B\left(P = \frac{\exp(-1 + 2 \times x_{13})}{1 + \exp(-1 + 2 \times x_{13})}\right),\quad x_{15} \right| Y \sim B\left(P = \frac{\exp(-0.5 + Y)}{1 + \exp(-0.5 + Y)}\right),$

$x_{16} | \text{age}, x_2 \sim B\big(P = \Phi(2 - 4 \times \text{age} - x_2)\big),\quad x_{17} | x_{15} \sim N\big(\mu = 50 - 3 \times x_{15},\ \text{sd} = 2\big),$

$x_{18} | x_{16} \sim B\big(P = \Phi(1 - 2 \times x_{16})\big),\quad x_{19} | x_{16} \sim B\left(P = \frac{\exp(-1 + 2 \times x_{16})}{1 + \exp(-1 + 2 \times x_{16})}\right),$

$x_{20} \left| x_{19} \sim B\left(P = \frac{\exp(-1.5 + 3 \times x_{19})}{1 + \exp(-1.5 + 3 \times x_{19})}\right),\quad x_{21} \right| x_{17} \sim N(\mu = 10 + x_{17},\ \text{sd} = 1),$

$x_{22} \left| x_{18} \sim B\left(P = \frac{\exp(-0.5 + x_{18})}{1 + \exp(-0.5 + x_{18})}\right),\quad x_{23} \right| Y \sim B\left(P = \frac{\exp(-0.5 + Y)}{1 + \exp(-0.5 + Y)}\right),$

$$x_{24}|\text{age},x_2 \sim B\big(P = \Phi(2 - 4 \times \text{age} - x_2)\big), \quad x_{25}|x_{23} \sim N\big(\mu = 15 - 5 \times x_{23},\ \text{sd} = 1\big),$$

$$x_{26}|x_{24} \sim N\big(\mu = 25 + 3 \times x_{24},\ \text{sd} = 1\big), \quad x_{27}|x_{25} \sim N\big(\mu = 30 + 1.25 \times \text{x}_{25},\ \text{sd} = 1.5\big),$$

$$x_{28}|x_{24} \sim B\big(P = \Phi(1 - 2 \times x_{24})\big), \quad x_{29}|x_{28} \sim B\big(P = \Phi(-1 + 2 \times x_{28})\big),$$

$$x_{30}\Big|Y \sim B\left(P = \frac{\exp(1 - 2 \times Y)}{1 + \exp(1 - 2 \times Y)}\right), x_{31}\Big|\text{age},x_2 \sim B\left(P = \frac{\exp(2 - 4 \times \text{age} - x_2)}{1 + \exp(2 - 4 \times \text{age} - x_2)}\right),$$

$$x_{32}|x_{30} \sim N\big(\mu = 5 + 4 \times x_{30},\ \text{sd} = 1\big), \quad x_{33}|x_{31} \sim B\big(P = \Phi(-1.5 + 3 \times x_{31})\big),$$

$$\text{x}_{34}\Big|x_{31} \sim B\left(P = \frac{\exp(-1 + 2 \times x_{31})}{1 + \exp(-1 + 2 \times x_{31})}\right), x_{35}\Big|x_{34} \sim B\left(= \frac{\exp(-1.5 + 3 \times x_{34})}{1 + \exp(-1.5 + 3 \times x_{34})}\right),$$

$$x_{36}|x_{32} \sim N\big(\mu = 11 - x_{32},\ \text{sd} = 1\big), \quad x_{37}|x_{33} \sim B\left(P = \frac{\exp(0.5 - x_{33})}{1 + \exp(0.5 - x_{33})}\right),$$

$$x_{38}\Big|Y \sim B\left(P = \frac{\exp(-0.5 + Y)}{1 + \exp(-0.5 + Y)}\right), x_{39}\Big|\text{age},x_2 \sim B\left(P = \frac{\exp(2 - 4 \times \text{age} - x_2)}{1 + \exp(2 - 4 \times \text{age} - x_2)}\right),$$

$$x_{40}|x_{38} \sim N\big(\mu = 20 - 5 \times x_{38},\ \text{sd} = 1\big), \quad x_{41}|x_{39} \sim N\big(\mu = 25 + 2 \times x_{39},\ \text{sd} = 1\big),$$

$$x_{42}|x_{40} \sim N\big(\mu = 10 + 1.25 \times x_{40},\ \text{sd} = 1.5\big), \quad x_{43}|x_{39} \sim B\big(P = \Phi(1 - 2 \times x_{39})\big),$$

$$\text{x}_{44}|x_{43} \sim B\big(P = \Phi(-1 + 2 \times x_{43})\big),\ x_{45}|Y \sim B\left(P = \frac{\exp(0.5 - Y)}{1 + \exp(0.5 - Y)}\right),$$

$$x_{46}|\text{age},x_2 \sim B\big(P = \Phi(2 - 4 \times \text{age} - x_2)\big), \quad x_{47}|x_{45} \sim N\big(\mu = 7 + 2 \times x_{45},\ \text{sd} = 1\big),$$

$$x_{48}|x_{46} \sim B\big(P = \Phi(-1.5 + 3 \times x_{46})\big), \quad x_{49}|x_{46} \sim B\left(P = \frac{\exp(-1 + 2 \times x_{46})}{1 + \exp(-1 + 2 \times x_{46})}\right),$$

$$x_{50}\Big|x_{49} \sim B\left(P = \frac{\exp(-1.5 + 3 \times x_{49})}{1 + \exp(-1.5 + 3 \times x_{49})}\right), x_{51}\Big|x_{47} \sim N\big(\mu = 25 - 1.5 \times x_{47},\ \text{sd} = 1\big),$$

$$x_{52}\Big|x_{48} \sim B\left(P = \frac{\exp(0.5 - x_{48})}{1 + \exp(0.5 - x_{48})}\right), \quad x_{53}\Big|Y \sim B\left(P = \frac{\exp(1 - 2 \times Y)}{1 + \exp(1 - 2 \times Y)}\right),$$

$$x_{54}|\text{age},x_2 \sim B\left(P = \frac{\exp(2 - 4 \times \text{age} - x_2)}{1 + \exp(2 - 4 \times \text{age} - x_2)}\right),$$

$$x_{55}|x_{53} \sim N\big(\mu = 20 + 5 \times x_{53},\ \text{sd} = 1\big),$$

$x_{56}|x_{54} \sim N(\mu = 25 - 2 \times x_{54},\ \mathrm{sd} = 1)$, $x_{57}|x_{55} \sim N(\mu = 15 + 1.3 \times x_{55},\ \mathrm{sd} = 1.5)$,

$x_{58}|x_{54} \sim B(P = \Phi(1 - 2 \times x_{54}))$, $x_{59}|x_{58} \sim B(P = \Phi(-1 + 2 \times x_{58}))$,

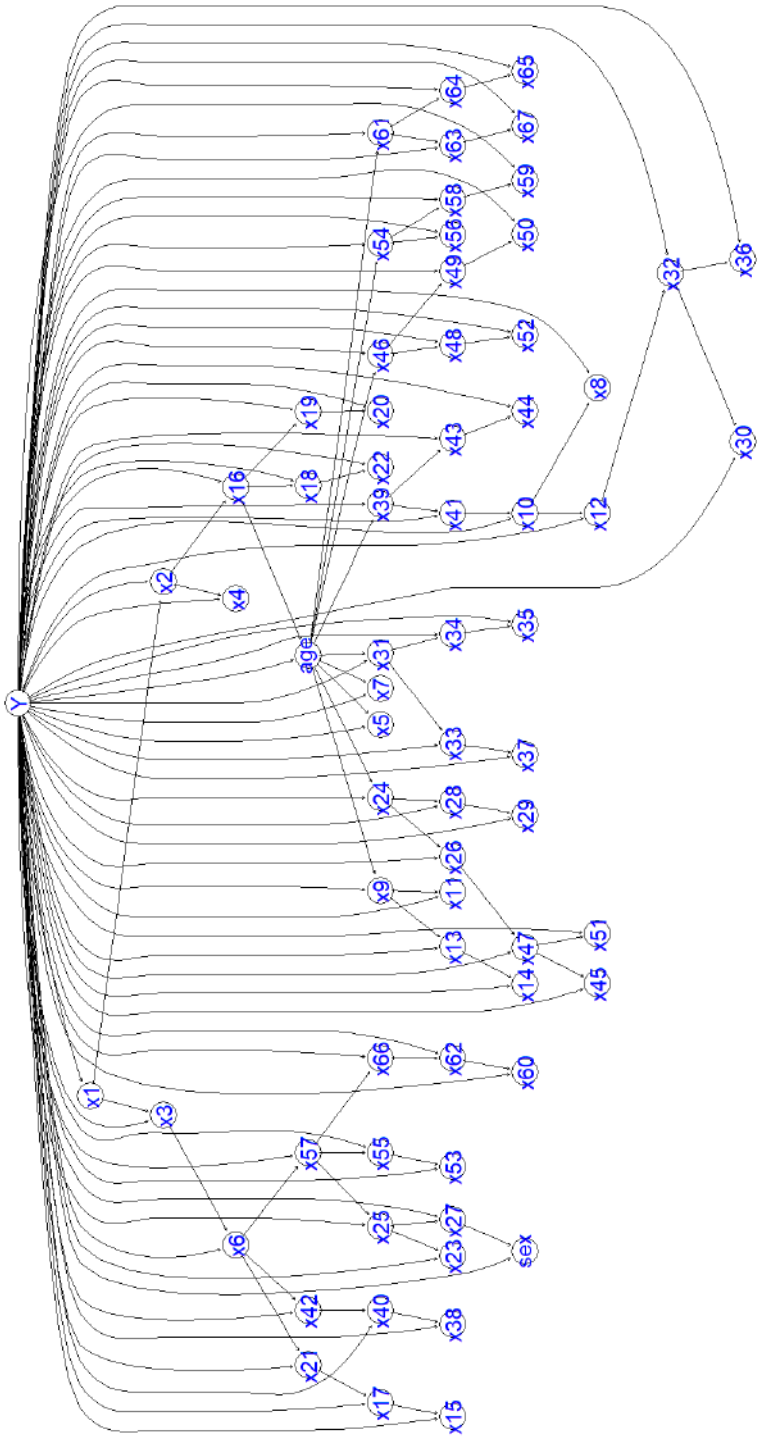
$x_{60}|Y \sim B(P = \Phi(-0.5 + Y))$, $x_{61}|\mathrm{age}, x_2 \sim B\left(P = \frac{\exp(2 - 4 \times \mathrm{age} - x_2)}{1 + \exp(2 - 4 \times \mathrm{age} - x_2)}\right)$,

$\mathrm{x}_{64}|x_{61} \sim B\left(P = \frac{\exp(1 - 2 \times x_{61})}{1 + \exp(1 - 3 \times x_{61})}\right)$,

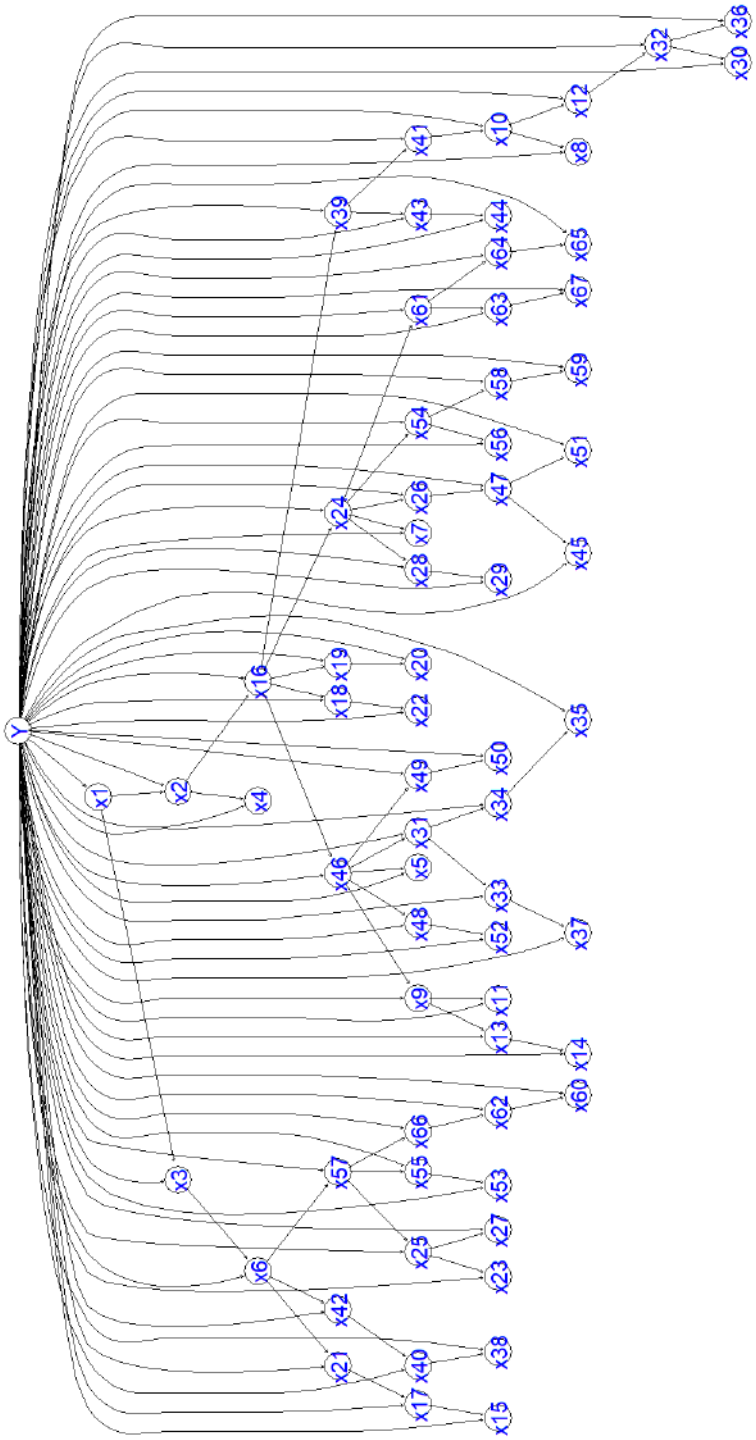
$\mathrm{x}_{65}|x_{64} \sim B\left(P = \frac{\exp(-1.5 + 3 \times x_{64})}{1 + \exp(-1.5 + 3 \times x_{64})}\right)$,

$x_{62}|x_{60} \sim N(\mu = 10 + 1.5 \times x_{60},\ \mathrm{sd} = 1)$, $x_{63}|x_{61} \sim B(P = \Phi(-1.5 + 3 \times x_{61}))$,

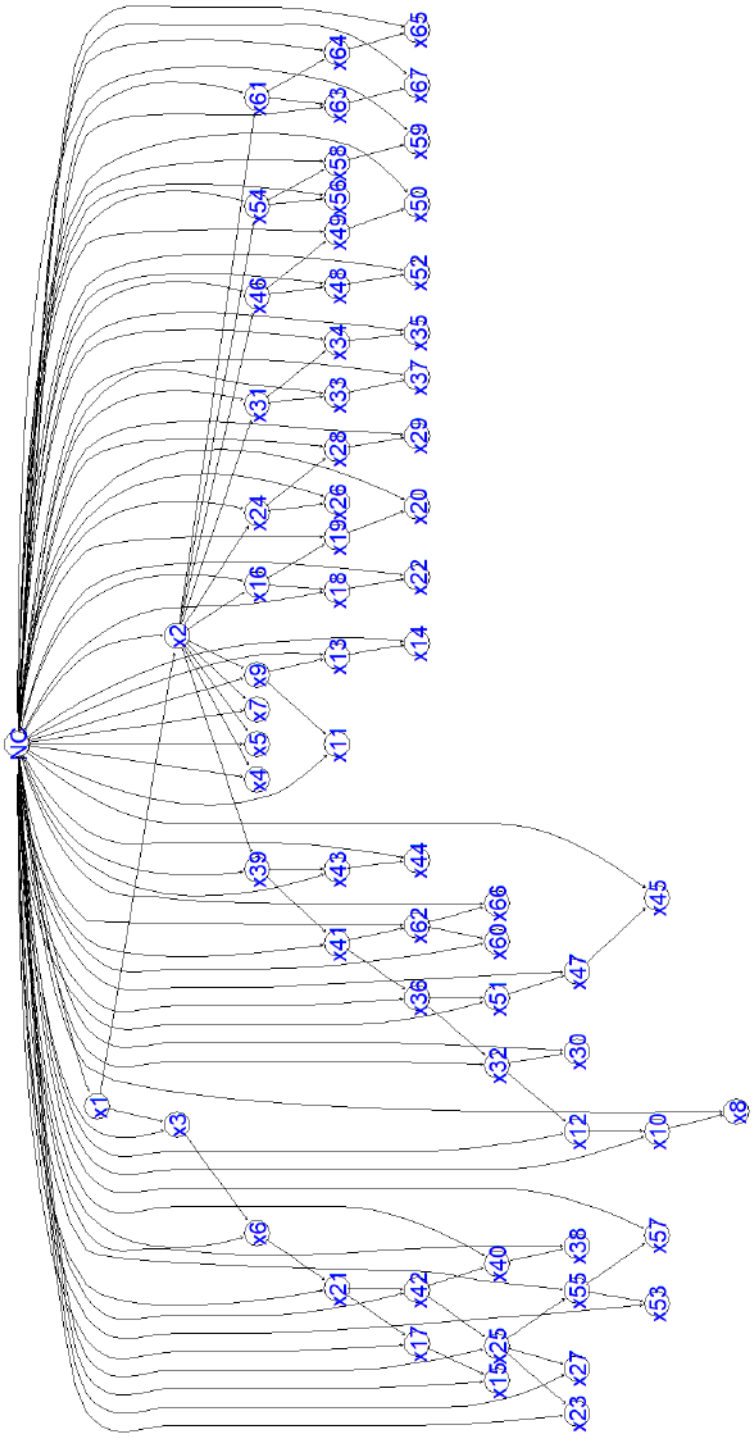
$x_{66}|x_{62} \sim N(\mu = 20 - 2 \times x_{62},\ \mathrm{sd} = 1)$, $x_{67}|x_{63} \sim B\left(P = \frac{\exp(0.5 - x_{63})}{1 + \exp(0.5 - x_{63})}\right)$.



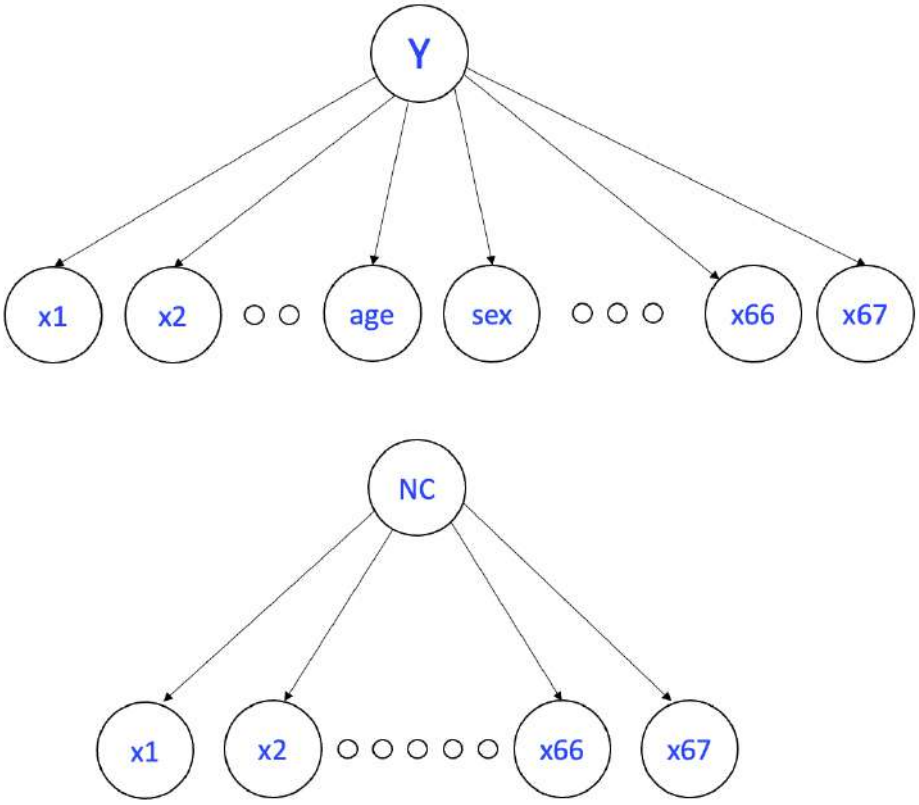
부록 그림2 – 1: p=70인 경우의 TAN



부록 그림2 – 2: p=70인 경우의 TAN-II



부록 그림2 – 3: p=70인 경우의 NTAN



부록 그림2 – 4: p=70인 경우의 NB(위), NNB(아래)*

* 나이브 베이즈의 경우 변수가 모두 조건부 독립이므로, 클래스 변수로부터 모든 변수에게로 화살표 방향이 향하는 형태로 표현된다.

ABSTRACT

Proposal of the improved Tree Augmented Naïve Bayes model

Hyun Woo Lim

Department of Statistics

Sungkyunkwan University

As the application of classification techniques to big data grows in many fields, so does the use of various machine learning techniques. Bayesian networks, also known as directed acyclic graphs, are widely used in many fields due to their ability to visualize causal relationships between variables using the information contained in big data. As examples of Bayesian network classifiers, Naive Bayes and TAN are used. TAN is particularly useful for alleviating Naive Bayes' "strong assumption of conditional independence," which is easily violated in practice, and for visualizing causal relationships between variables. However, because the existing TAN network employs all variables in the construction process, an unnatural network was formed as a result of variables that should not be included in the causal relationship being included in the network, making interpretation of the network impossible.

In this paper, we propose a New Tree Augmented Naive Bayes (NTAN) method for solving this problem, which includes the set of variables to be excluded from the network construction process in the class variable and excludes them from the causal relationship. Rather than simply excluding the variable, the information possessed by the variable was used as prior information to prevent information loss and to build a natural network by including it in the class variable.

The simulation confirmed that NTAN performed similarly to TAN in classification, but it formed a more natural network, making it easier to interpret the relationship between variables. And, using real data, health checkup data and breast cancer recurrence data, a natural network was constructed with NTAN, and the relationship between variables was examined. Furthermore, it was confirmed that NTAN's classification performance was superior to that of TAN. As a result, we show that the method proposed in this dissertation not only forms a natural network but also improves classification performance when compared to TAN.

Keywords: Bayesian Network, directed acyclic graph, Tree Augmented Naïve Bayes, Machine Learning.