베이지안 망에 기초한 불임환자 임상데이터의 분석

정 용 규† · 김 인 철††

요 약

본 논문에서는 베이지안 망을 기초로 불임환자의 임상 데이터에 대한 다양한 분석 실험을 전개하였다. 이 실험을 통해 임신여부에 영향을 주는 요인들간의 상호의존성을 분석해보고, 또 NBN, BAN, GBN 등 제약조건이 다른 다양한 유형의 베이지안 망 분류기들의 분류성능을 서로 비교해보았다. 그리고 우리는 이와 같은 실험을 통해 임신가능여부(Clin)에 직접적인 영향을 미치는 중요한 요인들로 증상(IND), 약물치료법(stimulation), 여성의 나이(FA), 미세조작 난자의 수(ICT), Wallace 사용여부(ETM) 등 5개의 특성들을 가려낼 수 있었고, 이 요인들간의 상호의존성도 찾아낼 수 있었다. 또 서로 다른 유형의 베이지안 망 분류기들 중에서 요인들간의 상호의존관계를 허용하는 좀더 일반적인 BAN과 GBN 등이 그렇지 못한 NBN에 비해 상대적으로 더 높은 분류 성능을 보여준다는 것을 확인하였다. 또 결정트리와 k-최근접 이웃과 같은 다른 분류기들과의 성능 비교를 통해, 임상 데이터의 특성상 확률적 표현과 추론에 기초한 베이지안 망 분류기들이 보다 높은 성능을 보여준다는 사실도 확인할 수 있었다. 또 본 논문에서는 클래스 노드의 Markov blanket에 속한 특성들로 특성집합을 축소하는것을 제안하고, 실험을 통해 이 특성 축소방법이 베이지안 망 분류기들의 성능을 높여 줄 수 있는지 알아보았다.

Bayesian Network-Based Analysis on Clinical Data of Infertility Patients

Yong-Gyu Jung† · In-Cheol Kim††

ABSTRACT

In this paper, we conducted various experiments with Bayesian networks in order to analyze clinical data of infertility patients. With these experiments, we tried to find out inter-dependencies among important factors playing the key role in clinical pregnancy, and to compare 3 different kinds of Bayesian network classifiers (including NBN, BAN, GBN) in terms of classification performance. As a result of experiments, we found the fact that the most important features playing the key role in clinical pregnancy (Clin) are indication (IND), stimulation, age of female partner (FA), number of ova (ICT), and use of Wallace (ETM), and then discovered inter-dependencies among these features. And we made sure that BAN and GBN, which are more general Bayesian network classifiers permitting inter-dependencies among features, show higher performance than NBN. By comparing Bayesian classifiers based on probabilistic representation and reasoning with other classifiers such as decision trees and k-nearest neighbor methods, we found that the former show higher performance than the latter due to inherent characteristics of clinical domain. Finally, we suggested a feature reduction method in which all features except only some ones within Markov blanket of the class node are removed, and investigated by experiments whether such feature reduction can increase the performance of Bayesian classifiers.

키워드 : 베이지안 망(Baysian Networks), 불임환자(Infertility Patients), 특징축소(Features Reduction)

1. 서 론

불임증이란 피임을 하지 않고 부부가 정상적인 성관계를 1년 이상 하였음에도 불구하고 임신이 되지 않는 경우를 말한다. 일반적으로 부부가 모두 정상이면 자연 임신이 1년 이내에 이루어지는 확률이 80~90%에 이른다. 하지만 점점 늘어나는 각종 공해와 이상 기후 및 사회생활에서의 각종 스트레스나 질병 등으로 인해 해마다 불임문제로 병원을 내원하는 환자는 늘고 있다[2]. 최근 연구에 의하면 연령별 불임률은 16~20세에 결혼한 여성의 경우 불임률이 4.5%정도이지만, 35~40세에 결혼한 여성은 32%, 40대에 결혼한 여성은 70%까지 증가한다[1]. 이러한 불임은 33%의 남성 불임요인을 제외하면 원인의 2/3는 여성에게서 발생한다. 더욱이 최근의 의학기술로 인해 남성적 불임원인에 대해서는 거의 95%이상의 임신 성공을 거두고 있어 문제가 되지 않는다. 그러나 여성의 경우, 원인 불명의 경우가 5%로 되어 있고, 자궁 경관, 무 배란, 골반 질환 등에 의한 원인들 대부분이 독립적으로 발생하는 것이 아니라 서로 복합적으로 발생하는 것을 고려할때, 현실적으로 다년간 시술 경험을 가진 전문가가 아니고서는 그 원인을 찾고 원인해결을 위해 적절한 시술을 하여 가임에 이르는 것이 대단히 어렵다. 불임증을 치료하기 위해서는 수개월 단위의 치료기간이 필요하며 치료를 위해서는 여러 가지의 검사가 요구된다. 검사 종

† 정 회 원 : 서울보건대학 전산정보처리과 교수
†† 종신회원 : 경기대학교 전자계산학과 교수
논문접수 : 2002년 7월 15일, 심사완료 : 2002년 8월 22일

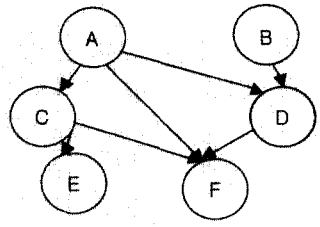
류로는 원인별로 볼 때 배란 여부를 알기 위한 검사, 자궁 이상에 대한 검사, 자궁 경관 이상에 대한 검사, 복강 내 이상에 대한 검사가 있으며 또한 각각에 대해 여러 가지의 구체적인 검사 방법들이 사용된다. 불임문제 해결을 위해서는 검사결과를 가지고 각 원인별로 치료를 해나가는 것이 보통이나 최근 들어서는 인위적으로 배란 촉진을 위한 약물 치료법과 함께, 생성된 난자와 정자를 체외에서 수정시키는 시험관아기시술(IVF-ET)이 가장 많이 사용된다. 시술 절차에 대해 간략히 살펴보면, 시험관아기시술을 위해서는 여러개의 난자와 정자를 따로 취하여 보관하여야 한다. 남성의 경우에는 한번 사정시 보통 ml당 2천만개 이상의 정자를 배출하기 때문에 큰 어려움이 없지만 여성의 경우에는 매달 1개의 난자만이 생성되어 배출되므로 한번에 여러개의 난자를 채취하기 위해서는 배란촉진제를 사용하는데, HCG, FSH, LH, Clomiphene, Parlodel, GnRH등이 많이 사용된다. 이러한 약제들은 대부분 여성의 뇌에서 방출되는 호르몬의 양을 조절하는 일을 담당한다. 약제를 통해 다량 배출된 난자들과 때맞춰 채취된 정자들은 시험관에서 체외수정을 할 수 있는데 수정된 수정란은 다시 여성의 몸에 이식하는 작업을 통해 착상과정을 갖게 된다. 이때 여러 가지 원인으로 인해 정자와 난자가 수정에 이르지 못하면 강제로 난자 속에 정자를 투입하여 수정을 유도하는데 이를 미세 조작술이라 한다. 또한 이식을 위해서 보조도구가 쓰이는데 Wallace를 최근 많이 사용한다. 이렇게 착상된 수정란은 포배기 등을 거쳐 분열하면서 태아로서 성장하게 된다[2]. 하지만 이는 가임을 위한 과정 중의 극히 적은 부분일 뿐이며 서술된 내용 중에도 명확하게 그 원인과 역할을 밝히지 못하는 많은 사실들이 존재한다. 그렇기 때문에 의학계에서도 이러한 문제들을 해결하기 위해 점차적으로 누적된 임상 실험 데이터에 대한 마이닝(mining) 작업을 통해 불임 요인들과 그들 간의 상관관계를 분석해보려는 연구들이 최근 들어 많이 시도되고 있다.

본 연구에서는 2년 동안 수집된 산부인과 불임 환자들에 대한 실제 임상 데이터로부터 불임과 관련된 요인들 간의 의존성을 표현하고 분석하는데 베이지안 망(Bayesian network)을 적용해보고자 한다. 일반적으로 베이지안 망은 잡음과 불확실성이 많은 영역 데이터로부터 분류에 영향을 미치는 속성들간의 상호의존성을 잘 표현하고 이것을 바탕으로 비교적 클래스를 정확하게 예측할 수 있는 견고한 확률적 도구로 알려져 있다. 본 논문에서는 실험을 통해 베이지안 망에 드러난 임신여부에 영향을 주는 요인들간의 상호의존성을 분석해보고, 또 NBN, BAN, GBN등 제약조건이 다른 다양한 유형의 베이지안 망 분류기들의 분류성능을 서로 비교해본다. 그리고 또 결정트리와 k-최근접 이웃과 같은 다른 분류기들과의 분류 성능 비교를 통해 의료 임상 데이터에 대해 베이지안 망 분류기들이 높은 성능을 보여줄수 있는지도 분석해본다. 또 본 논문에서는 하나의 베이지안 망에서 클래스 노드의 Markov blanket에 속한 특성들로 특성집합을 축소하는 것이 베이지안 망 분류기들의 성능을 높여 줄 수 있는지를 실험을 통해 알아본다.

2. 베이지안 망

2.1 기본 개념

베이지안 망(Bayesian network)은 특정 분야의 영역 지식(domain knowledge)을 확률적으로 표현하는 대표적인 수단으로서, (속성)변수들간의 확률적 의존 관계(probabilistic dependency)를 나타내는 그래프와 각 변수별 조건부 확률들로 구성된다[11, 13]. 따라서 하나의 베이지안 망은 각 노드마다 하나의 조건부 확률표(conditional probability table)를 갖는 하나의 비순환 유향 그래프(directed acyclic graph) G =< N, A >로서, B = < N, A, Θ >으로 정의할 수 있다. 이때 각 노드 n ∈N은 하나의 영역변수를, 각 아크 a ∈A는 두 변수간의 확률적 의존성을 나타내며, Θ는 조건부 확률들의 집합을 나타낸다. 일반적으로, 하나의 베이지안 망은 다른 노드들에 배정된 값들을 기초로 특정 노드가 가질 값에 대한 조건부 확률을 계산하는데 이용할 수 있다. 따라서 하나의 베이지안 망은 한 개체의 다른 속성들의 값이 주어졌을 때 분류 클래스 노드(classification node)의 사후 확률 분포(posterior probability distribution)를 구해줌으로써 개체들에 대한 하나의 자동 분류기(classifier)로 이용될 수 있다[15, 19]. 즉 하나의 데이터 집합으로부터 베이지안 망을 학습할때 베이지안 망의 각 노드는 데이터 집합의 각 속성을, 각 아크는 속성들간의 의존성을 표현하게 되며, 이렇게 학습된 베이지안 망을 기초로 분류 클래스를 확률적으로 예측할 수 있다. 예를 들어 변수 A, B, C, D, E, F가 있고, 각 변수들이 yes와 no값을 갖는다고 할때 베이지안 망은 (그림 1)과 같이 각 변수들에 대한 의존성을 그래프로 표현할뿐 아니라 각 변수별로 <표 1>과 같은 조건부 확률도 함께 표현할 수 있다.



(그림 1) 베이지안 망 그래프

〈표 1〉 조건부 확률표 (CPT)

	A(y), B(y)	A(y), ~B(n)	~A(n), B(y)	~A(n), ~B(n)
D(y)	0.4	0.1	**0.8**	0.2
~D(n)	0.6	**0.9**	0.2	0.0

<표 1>은 6개의 변수 중, 변수 D가 각각 yes와 no 값을 가질 하나의 조건부 확률표(conditional probability table)를 나타내고 있다. (그림 1)에서 보듯이 변수 D는 변수 A와 B에 대해서만 종속성을 가지며 따라서 조건부 확률표의 각 열(column)은 이 두 변수 A와 B에 배정 가능한 값들을 나타내며 각 행(row)은 변수 D가 가질 수 있는 값들을 나타낸다. 결국 표상의 각 셀(cell)은 두 변수 A와 B에 배정 가능한 값들에 대해 변수 D가 yes 또는 no를 가질 조건부 확률을 나타낸다. 예컨대 <표 1>에서 변수 A가 no, B가 yes일 때 변수 D가 yes일 확률은 0.8이고, 반면에 D가 no일 확률은 0.2이다. 즉 두 조건부 확률은 각각 P(D = yes | A = no, B = yes) = 0.8, P(D = no | A = no, B = yes) = 0.2이다.

베이지안 망은 조건부 확률 계산에 식 (1)과 같은 베이즈 정리(Bayesian Theorem)를 이용한다. 베이즈 정리는 관측된 데이터 D로부터 가설 h가 옳을 확률 P(h|D)을, 가설 h에 기초한 데이터 D의 조건부 확률 P(D|h)과 가설 h의 사전 확률(prior probability) P(h)을 기초로 계산할 수 있는 방법을 제시한다[14, 23].

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)} \tag{1}$$

2.2 베이지안 망 분류기의 유형

베이지안 망 분류기는 한 개체 j가 클래스 C_i에 속할 확률을 계산함으로써 그 개체를 분류하는 간단한 분류 방법이다. 그러한 확률은 식 (2)와 같이 계산되며, 이때 개체 j는 $A_i = V_i$ 형태의 속성-값 쌍들로 표현되는 것으로 가정한다.

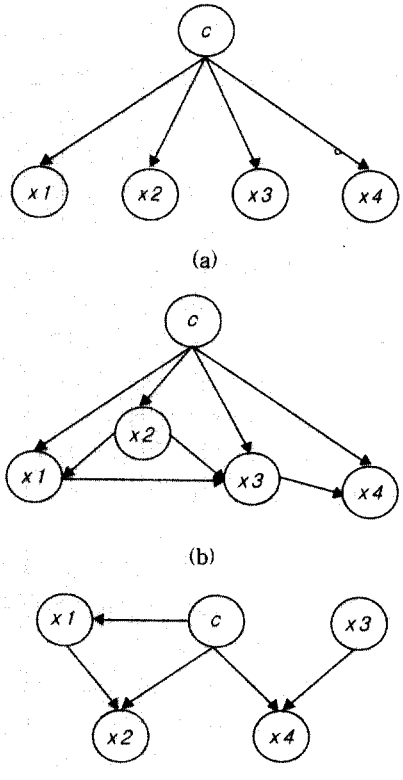
$$P(C_i \mid A_1 = V_{1,} \,\&\, \cdots \,\&\, A_N = V_{N,}) \tag{2}$$

대표적인 3가지 유형의 베이지안 망 분류기에는 (그림 2)에서 보는 바와 같이 나이브 베이지안 망(Nave Bayesian Network, NBN), 베이지안 망으로 확장한 나이브 베이지안 망(Bayesian Network Augmented Nave-Bayes, BAN), 일반 베이지안 망(General Bayesian Network, GBN) 등이 있다[5]. NBN은 (그림 2)의 (a)와 같이 클래스 노드 c를 제외한 다른 모든 속성노드들이 클래스 노드에만 의존적이고, 그들 간에는 서로 독립적이라는 가정을 바탕으로 만들어진 베이지안 망이다. 즉, 한 개체 j가 클래스 C_i에 속할 확률을 계산할 때 식 (3)과 같이 계산한다.

$$\begin{aligned} &P(C_i \mid A_1 = V_{1,} \,\&\, \cdots \,\&\, A_N = V_{N,}) \\ &\approx P(C_i) \prod_{k=1}^{n} P(A_k = V_k \mid C_i) \end{aligned} \tag{3}$$

이러한 NBN은 가정의 단순성에도 불구하고 많은 연구를 통해 비교적 높은 분류 성능을 보여주는 것으로 알려져 있다. 하지만 NBN의 기초가 되는 이 가정은 실세계 문제들에서는 거의 만족되지 않는 가정이므로 최근 들어 속성들간의 독립성 가정을 완화함으로써 NBN의 분류 성능을 높여보려는 많은 시도들이 이루어지고 있다.

BAN은 이와 같은 시도의 하나로서, NBN과는 달리 속성노드들 간에도 상호의존성이 존재한다고 가정하고 이러한 속성간 상호의존성을 하나의 일반 베이지안 망 형태로 표현 가능하도록 NBN을 확장한 것이다. 즉 BAN은 (그림 2)의 (b)와 같이 클래스 노드 c를 제외한 다른 모든 속성들간의 상호의존성을 또 하나의 베이지안 망으로 표현할 수 있다. 베이지안 망 분류기중 가장 일반화된 형태는 (그림 2)의 (c)와 같은 GBN으로서, GBN에서는 기존의 다른 베이지안 망 분류기들과는 달리 클래스 노드조차 일반 속성노드와 차이를 두지 않고 모든 노드들 간의 상호의존성을 하나의 베이지안 망으로 표현한 것이다[3]. 따라서 GBN에서는 클래스 노드도 부모 노드들을 가질 수 있다. GBN은 속성간 상호의존성을 표현할 수 있는 가장 자연스러운 방법이지만 어떠한 제약도 갖지 않은 상태에서 이러한 베이지안 망을 학습하는 데는 매우 높은 학습비용이 소요된다.



(그림 2) (a) Naive Bayesian Network (NBN),
(b) BN Augmented Naive-Bayes (BAN),
(c) General Bayesian Network (GBN)

2.3 베이지안 망의 학습

베이지안 망을 학습하는 과정은 크게 베이지안 망 그래프를 학습하는 과정과 그것을 바탕으로 각 변수의 조건부

확률들을 계산하는 과정으로 나누어 볼 수 있다[4, 11]. 사람이 배경지식을 가지고 베이지안 망 그래프를 직접 수작업으로 그려주거나 편집해주면 이를 바탕으로 훈련 데이터들을 분석하여 조건부 확률들을 자동으로 계산해주는 방식의 많은 베이지안 망 학습 알고리즘과 프로그램들이 존재한다. 하지만 베이지안 망 그래프로부터 각 변수의 조건부 확률을 계산하는 과정은 매우 단순한 과정인데 반해 훈련 데이터로부터 베이지안 망 그래프를 학습하는 과정은 매우 복잡하고 어려운 과정이다. 따라서 기존의 많은 연구들은 바로 이러한 베이지안 망 그래프 학습에 초점이 맞추어져 왔다[5, 8, 11, 20, 21]. 베이지안 망 그래프 학습을 위한 기존의 방법들은 크게 점수 기반 학습 알고리즘(scoring-based learning algorithm)들과 조건부 독립성 기반 학습 알고리즘(CBL, conditional independency-based learning algorithm)들로 나눌 수 있다. 점수 기반의 학습 알고리즘은 영역지식을 바탕으로 임의의 초기 베이지안 망을 만들고 일정한 평가기준을 이용하여 가장 좋은 점수를 받는 양질의 베이지안 망이 만들어질 때까지 계속해서 이 베이지안 망을 고쳐가는 일종의 휴우리스틱 탐색방법이다. 점수를 산출하는데 이용되는 평가기준(scoring criteria)에 따라 엔트로피 기반의 방법(Entropy-based method)[11], 베이지안 점수 방법(Bayesian scoring method), MDL(Minimum Description Length) 방법 등이 제안되었다. 한편 조건부 독립성 기반 학습 알고리즘에서는 노드간 조건부 독립성 테스트(CI test)를 시행하여 임계값 이상의 독립성을 갖는 노드들간에 아크를 삭제하거나 혹은 임계값 이하의 독립성을 갖는 노드들간에 아크를 추가해가는 방법이다[22]. 기존의 많은 연구들을 통해 조건부 독립성 기반의 학습 알고리즘들이 점수 기반 학습 알고리즘들에 비해 비교적 학습시간은 오래 걸리지만 보다 우수한 베이지안 망을 학습할 수 있는 것으로 알려져 있다[8, 12].

본 논문에서 시행하는 의료 데이터마이닝을 위한 베이지안 망 학습에는 대표적인 조건부 독립성 기반 학습 알고리즘의 하나인 Jie Cheng의 CBL 알고리즘[4]을 이용한다. node 수가 N 개일때 요구되는 time complexity가 $O(N^4)$로서 node수에 따라 급격히 증가하게 되므로 본 논문에서는 node수를 줄이기 위하여 feature reduction을 함으로써 이를 효율적으로 적용할 수 있다. 알고리즘의 학습과정은 크게 3단계로 이루어진다 : 초안 작성 단계(drafting phase) 단계, 아크 추가 단계(thickening phase), 아크 삭제 단계(thinning phase). 먼저 초안 작성 단계에서는 식 (4)로 정의되는 두 노드간의 단순 상호정보량(mutual information)을 계산하여 임계값 이상을 갖는 노드들간에 아크들을 추가함으로써 개략적인 베이지안 망 그래프의 초안을 작성한다.

$$I(X_i, X_j) = \sum_{x_i, x_j} P(x_i, x_j) \log \frac{P(x_i, x_j)}{P(x_i)\,P(x_j)} \qquad (4)$$

아크 추가 단계에서는 현재 초안 그래프에 포함되어 있는 아크들 외에 식 (5)와 같이 정의되는 두 노드간의 조건부 상호정보량(conditional mutual information)이 임계값 이상인 경우들을 찾아 해당하는 새로운 아크들을 그래프에 추가한다. 아크 삭제 단계에서는 이미 그래프에 포함되어 있는 기존 아크들에 대해 식 (5)의 조건부 상호정보량이 충분한지 검사하여, 그렇지 못한 아크들을 모두 제거하여 최종 베이지안 망 그래프를 완성한다. 이와 같은 CBL 알고리즘은 노드들의 완전한 순서(total ordering)가 주어진 경우에는 $O(N^2)$, 아무런 순서가 주어지지 않는 경우에는 $O(N^4)$에 비례하는 조건부 독립성 테스트를 요구하는 매우 효율적인 알고리즘이다.

$$I(X_i, X_j \mid C) = \sum_{x_i, x_j, c} P(x_i, x_j, c) \log \frac{P(x_i, x_j \mid c)}{P(x_i \mid c)\,P(x_j \mid c)} \qquad (5)$$

2.4 베이지안 망을 위한 특징 축소

일반적으로 한 개체를 표현하는 중요한 속성(attribute)들을 특징(feature)이라고 한다. 한 개체의 분류 클래스를 판단하는데 큰 영향을 미치지 못하는 특징들은 삭제하고 반대로 중요도가 높은 특징들만을 골라 이들로 분석데이터를 표현하는 처리과정을 특징 축소(feature reduction), 특징 부분집합 선택(feature subset selection), 차원 축소(dimension reduction) 등으로 부른다. 일반적으로 이와 같은 특징 축소를 통해 처리 대상 데이터의 양을 줄임으로써 계산의 효율성을 높일 수 있고, 보다 함축된 분류 지식이나 패턴을 얻을 수 있으며, 때로는 분류기의 성능을 향상시킬 수 있다. 특징 축소를 위한 매우 다양한 방법들이 제안되었는데, 이들은 크게 여과 방법(filtering method)과 포장 방법(wrapper method)으로 나누어 볼 수 있다[16]. 여과 방법은 정보획득량(information gain), 상호정보량(mutual information), χ^2 테스트 등의 척도를 이용하여 각 특징의 중요도를 개별적으로 평가하고 이것이 일정한 수준에 미치지 못하는 특징들을 삭제하는 방식이다. 이에 반해 포장 방법은 가능한 특징집합으로부터 실제로 특정 분류기를 생성하여 이 분류기의 분류성능을 검사해봄으로써 보다 더 나은 분류 성능을 보일 수 있는 특징들의 부분집합을 찾아가는 방식이다. 특징 축소를 위한 알고리즘에 특정 분류기를 생성하고 적용하는 과정을 내포하는 포장 방법이 일반적으로 분류기와는 독립적으로 적용되는 여과 방법에 비해 비용은 많이 소요되나 더 높은 분류 성능을 보이는 특징들을 찾을 수 있다.

Langley와 Sage의 연구[17]에서는 양질의 특징 부분집합을 찾기 위해 전향선택(forward selection)방법을 적용하였고 이렇게 선택된 특징들만을 포함하는 나이브 베이지안 망(NBN)을 생성하여 분류에 적용하였다. Kohavi와 John이 연구[16]에서는 특징 부분집합 선택을 위해 최적 우선 탐색 방법(Best-First Search)을 적용하였다. 이 방법은 결정트리(decision tree)나 나이브 베이지안 망 분류기 등 임의의 분류기

를 포함하는 일종의 포장 방법이다. Pazzani의 연구[18]에서는 나이브 베이지안 망 분류기의 분류 성능을 높이기 위한 방법으로서, 특징 부분집합 선택뿐만 아니라 특징 융합(feature joining)도 적용해 보았다. 본 연구에서는 베이지안 망 분류기의 성능을 높이기 위한 특징 부분집합 선택 방법으로 분류 클래스 노드의 Markov blanket에 속한 특징들을 선택하였다. 베이지안 망에서 한 노드의 Markov blanket은 그 노드의 부모 노드들과 자식 노드들, 그리고 자식 노드들의 또 다른 부모 노드들을 포함하는 노드들의 부분 집합이다. 따라서 베이지안 망에서 클래스 노드의 Markov blanket은 분류에 직접 영향을 주는 특징들만을 포함함으로써 자연스러운 또 하나의 특징 부분집합 선택 방법이 될수 있다. 하지만 이러한 방법으로 선택된 특징들만으로 재구성된 베이지안 망 분류기들이 실제로 어떤 분류 성능의 변화를 가져오는지를 밝힌 실험연구는 많지 않다.

3. 데이터 수집 및 전처리

본 연구에 이용할 실제 의료 임상 데이터의 수집을 위해 서울에 소재한 모 종합병원의 산부인과에 2년 동안 내원한 불임환자의 검사기록과 시술기록, 그리고 시술 결과로 얻어진 임신성공여부가 담긴 데이터를 입수하였다. 수집된 원래 데이터집합에는 약 400여명의 환자에 대하여 총 39개의 검사항목이 기록되어 있었으나 데이터 분석을 위한 전처리 작업(preprocessing)을 통해 검사항목이 일부환자에게만 적용이 되어서 동일한 조건에서 비교할 수 없는 항목들과 임신과 관련성이 거의 없다고 전문가가 판단하는 항목 등을 제거하여, 실제 데이터 분석에 사용할 항목을 임신여부를 나타내는 항목인 Clin을 포함해 <표 2>와 <표 3>에 열거된 것과 같이 총 9개로 줄였다. 일반적으로 데이터 분석을 위한 전처리 작업으로는 이와 같은 차원 축소(dimension reduction) 외에도 필요한 경우 데이터 정제(cleaning), 변환(transformation), 이산화(discretization) 등이 적용될 수 있다. 본 연구를 위해 수집된 데이터의 경우에도 누락 항목 데이터와 잡음(noise)이 포함된 데이터들이 많아 이들을 처리하기 위한 데이터 정제작업이 수행되었다.

그 결과 누락 데이터와 잡음이 많았던 일부 환자의 데이터는 제외하여 총 269개의 정제된 데이터를 얻었다. 또한 이렇게 정제된 의료 데이터 집합에는 정리되지 않은 약어 형태의 데이터 값과 더불어 연속 수치 데이터 값들을 많이 포함하고 있어 적절한 이산화작업이 필요하였다.

대표적인 이산화 방법에는 하나의 속성이 갖는 값의 범위에 대해 동일한 넓이를 갖는 구간을 정한 후 구간마다 대표값을 할당하는 Equal Width Interval Binning방법과, 값의 넓이와 깊이에 구애 받지 않고 영역지식에 의하여 구간을 정하고 각 구간에 속한 값을 대표값을 정하는 Holte's 1R Discretizer방법, 그리고 마지막으로 각 속성값들을 임의의 범위로 구분한 후 엔트로피 계산을 통해 가장 작은 수치를 나타내는 구간을 선택하여 사용하는 Recursive Minimal Entropy Partitioning 방법을 들수 있다[7]. 본 논문에서는 <표 3>에서 열거한 FA, ICT, TO 등 3가지 항목에 대해서 이산화 작업을 하였으며, 적용된 이산화 방법은 Holte's 1R Discretizer방법이다. 이는 각 항목들이 나타내는 값들의 분포가 일정치가 않고 변화가 심한 본 연구에 적합하며, 도메인의 특성에 따라 범위를 도메인 전문가들의 조언을 근거로 정하는 방법으로 domain의 특성으로 인해 영역지식을 가장 잘 반영하는 방법으로 선택한 것이다.

〈표 2〉 정제된 데이터 항목

Attribute	Description	Value	Description	Value
증상 (IND)	Endometriosis Immunological Ovarian O and T O and U Peritoneal	A B C D E F	P and T Tubal T and U Uterine Unexplained	G H I J U
약물치료법 (Stimulatio)n	Long Protocol Short Protocol Ultra Short HMG only FSH	L S U H F	Parlodel Follicular Null Clomiphene FSH-HMG	P RF N C FSH-H
시술방법	IVF-ET	C	ICSI	I
Wallace사용여부 (ETM)	Wallace	W	Default	T
이식일수 (ETD)	Second day Third day	STD TTD	Fifth day Sixth day	FTD SXTD
임상적임신여부 (Clin)	True	1	False	0

〈표 3〉 이산화한 데이터 항목

Attribute	Description	Value	Description	Value
여성의나이 (FA)	20~34 35~40	L M	over 40	H
미세조작난자수 (ICT)	Null 1~5 6~10	N L M	11~15 16~20 over 20	MH H HH
총이식수정란수 (TO)	Null 1~5	N Normal	6~10 over 10	SHIGH HIGH

4. 실 험

4.1 실험 목표

본 연구에서는 앞서 전처리 작업을 통해 정제된 임신관련 의료 임상 데이터집합에 대해, 베이지안 망을 기초로 다양한 분석 실험을 전개하였다. 특별히 의료 데이터집합에 대해 베이지안 망을 적용해보려는 이유는 이 영역 데이터가 갖는 몇가지 특성 때문이다. 먼저 산부인과를 비롯해 대부분의 의료 임상 영역의 경우 원인-결과 관계, 요인간 연관 관계 등 진단과 처방에 필요한 의료 지식이나 이론에 많은 불확

실성과 가변성을 내포하고 있어 명확한 사전 지식을 확보하거나 단언적 추론을 전개하기 어렵다. 따라서 이러한 불확실성을 고려할 수 있는 확률적 학습과 추론 방법이 반드시 필요하다. 다음은 대부분의 임상 데이터들은 병리적 실험이나 시술의 결과를 관측함으로써 얻어지는데 이러한 데이터들은 필연적으로 관측기기의 오차와 관측 환자의 생리상태 및 관측 시점 등에 따라 많은 잡음(noise)과 특이값(outlier)을 포함할 수 밖에 없다. 따라서 이러한 데이터로부터 임신 가능 여부를 자동 예측하기 위해서는 특별히 잡음과 특이값에 견고한 분류 학습법을 적용하여야 한다. 본 연구에서는 앞서 설명한 산부인과 임상 데이터에 대해 다음과 같은 분석 실험 목표를 설정하였다.

- 중요한 임신 요인들간의 상호의존성을 분석

주어진 데이터로부터 베이지안 망을 학습시켜봄으로써, 베이지안 망에서 드러난 임신 요인들간의 상호의존성을 분석해보고 그 의미를 해석해본다.

- 베이지안 망 분류기들 간의 분류성능 비교

주어진 의료 데이터집합으로부터 NBN, BAN, GBN 등 제약조건이 다른 다양한 유형의 베이지안 망 분류기들을 생성하고 이들이 이 의료영역 데이터에서 보여주는 분류성능을 서로 비교해본다.

- 대표적인 다른 분류기들과의 분류성능 비교

의료 영역 데이터의 특성을 고려할때 베이지안 망 분류기가 어느 정도 효과가 있는지를 알아보기 위해 결정트리(decision tree), k-최근접이웃(k-nearest neighbors, k-NN) 등과 같은 대표적인 다른 분류기들과의 분류성능을 비교해본다.

- Markov blanket을 이용한 특징축소 효과 분석

베이지안 망 분류기의 학습 효율과 분류성능을 높이기 위한 특성 축소의 한 방법으로서, 분류 클래스노드의 Markov blanket으로 특성을 축소하는 것이 얼마만큼 효과가 있는지 베이지안 망 분류기별로 분류성능을 비교해본다.

4.2 실험 방법

앞서 설명한 실험목표를 위해 전처리 작업이 끝난 (그림 3)과 같은 실험 데이터로부터 몇 가지 유형의 베이지안 망을 생성한다. 본 연구에서 실험할 베이지안 망 분류기 유형은 NBN(Naive Bayesian Network), BAN(Bayesian Network Augmented Naive-Bayes), GBN(General Bayesian Network) 등으로 각 분류기가 내포하고 있는 가정과 제약이 분류성능에 미치는 효과를 보기 위함이다. 특히 이중에서 특별한 제약 없이 변수들의 의존성을 가장 풍부하게 표현할 수 있는 GBN을 중심으로 클래스 노드인 Clin과 직접 연결 아크를 갖는 임신 요인들을 구해보고 또 이들간의 상호의존성을 분석해본다. 특징 축소 실험을 위해서는 GBN 그래프상에서 분류 클래스 노드의 Markov blanket에 속한 특성들만을 골라낸 뒤, 이 특성들만을 포함하도록 실험 데이터를 축소하고 이 축소된 데이터로부터 다시 NBN, BAN, GBN 등을 학습해내었다. 이와 같이 축소된 특성집합으로부터 얻어진 NBN, BAN, GBN을 각각 NBNSF(NBN with Selected Features), BANSF(BAN with Selected Features), GBNSF(GBN with Selected Features) 등으로 부른다. 특징 축소의 효과를 알아보기 위한 방법으로 특징 축소 이전의 NBN, BAN, GBN 등과 특징 축소 이후의 NBNSF, BANSF, GBNSF 등의 분류성능을 서로 비교해본다. 또 의료 영역 데이터에서의 베이지안 망 분류기의 강점을 확인하기 위하여 동일한 실험 데이터와 실험 조건 하에서 다른 대표적인 분류기인 결정트리와 k-최근접이웃 방법을 적용하고 분류성능을 비교해본다. 이를 위해 결정트리 학습 알고리즘으로는 성능이 비교적 우수한 C4.5를 적용하고, k-최근접이웃을 위해서는 실험을 통해 성능이 가장 좋은 k 값(k = 3)을 설정한다. 분류성능을 알아보기 위한 테스트방법은 3가지를 적용한다. 먼저 269개의 전체 실험 데이터집합을 244개의 훈련 데이터와 25개의 테스트 데이터집합으로 나눈다. 그리고 훈련 데이터집합에 대해 각 베이지안 망 분류기들과 여타 분류기들을 학습한 뒤, 이 분류기들을 각각 훈련 데이터 집합과 테스트 데이터 집합으로 분류성능 테스트를 시행해본다. 끝으로 전체 실험 데이터 집합을 가지고 10회 교차검정방법(10-fold cross validation)으로 각 분류기의 평균 분류성능을 알아본다.

FA	INCL	Stimulation	IVF	ICT	ETML	ETD	TG	CLin
31	T	L	C	0	W	2	4	N
43	T	L	C	0	T	2	4	N
37	T	L	C	0	W	2	2	Y
34	T	L	C	0	T	2	6	Y
34	T	U	C	0	T	2	6	N
27	T	L	C	0	W	2	3	N
38	T	L	C	0	T	2	3	N
35	UN	L	C	0	T	2	2	N
37	E	L	C	0	W	2	3	Y
32	T	L	C	0	T	2	5	Y
34	T	U	C	0	W	3	7	Y
29	T	L	C	0	T	2	4	N
31	E	L	C	0	W	2	5	N
40	UT	L	C	2	W	3	3	N
36	UN	U	I	4	T	3	4	N
28	T	L	C	0	T	3	6	N
27	T	U	C	0	W	3	3	N
36	T	L	C	0	T	3	5	Y
36	T	L	C	0	W	3	6	Y

(그림 3) 실험데이터의 일부

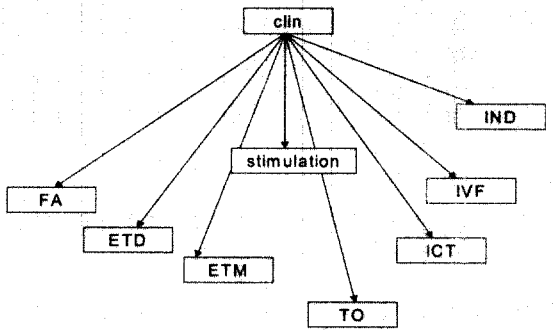
베이지안 망 학습을 위해서는 조건부 독립성 기반의 알고리즘인 Jie Cheng의 CBL 알고리즘과 그것을 구현한 BN PowerConstructor 1.0 프로그램[6]을 이용한다. 다른 분류기와의 비교를 위해서는 Java로 구현된 Weka 패키지를 이용한다. 모든 실험은 Intel Pentium III 700, 256M Memory 사양의 컴퓨터에서 진행하며 실험에 사용된 원시데이터는 Excell파일로 구성된 형태로서 그 일부는 (그림 3)과 같다.

5. 실험결과 및 평가

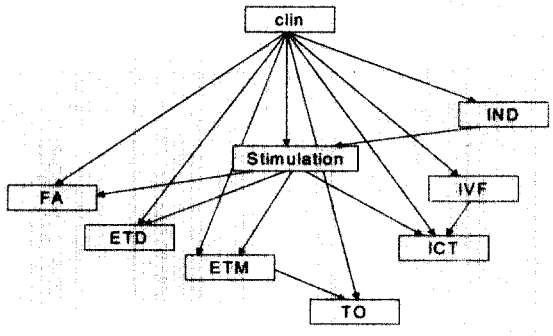
5.1 베이지안 망 분류기의 생성

(그림 4)는 훈련 데이터집합으로부터 학습된 베이지안 망

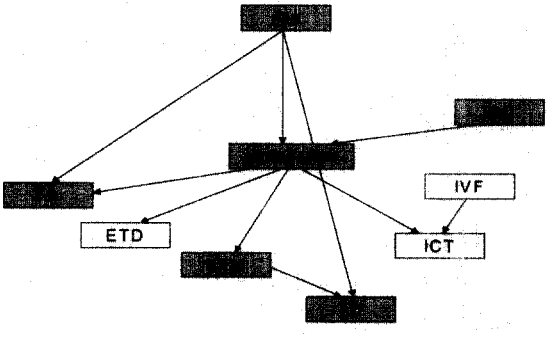
분류기들을 나타내고 있다. (그림 4)의 (a)는 분류 클래스노드를 제외한 모든 항목 즉 임신 요인들간에 독립성을 가정하는 NBN이고, (b)는 임신 요인들간의 상호의존성을 또 하나의 베이지안 망으로 표현한 BAN, (c)는 가장 일반적인 GBN을 나타낸다. 그림에서는 공간문제로 인해 베이지안 망 그래프(graph)만 표현하였으나 각 노드별로 조건부 확률표(CPT)도 함께 구해진다. (그림 4)의 NBN과 BAN은 모두 분류 클래스 노드 Clin을 뿌리(root)로 놓고 베이지안 망을 구한 반면, GBN의 경우 Clin을 일반 노드와 같이 취급하여 베이지안 망을 구하였으나 역시 NBN과 BAN의 경우와 같이 분류 클래스 노드인 Clin은 부모 노드를 하나도 갖지 않는 노드가 되었다.



(a)



(b)

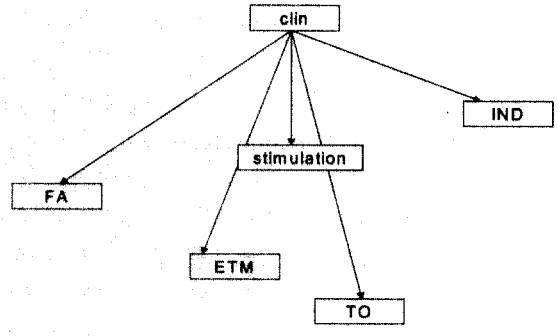


(c)

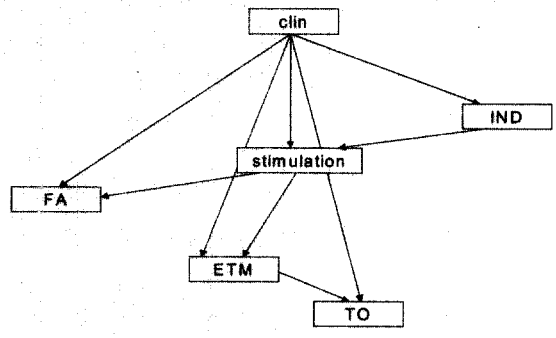
(그림 4) 실험 데이터로부터 얻어진 베이지안 망 분류기들 : (a) NBN, (b) BAN, (c) GBN

(그림 4)의 (c)와 같이 구해진 GBN으로부터 분류 클래스 노드인 Clin의 Markov blanket을 구하면 Clin 자신과 자식 노드들인 FA, stimulation, TO, 그리고 이들의 부모 노드들인 IND, ETM 등이 포함된다.

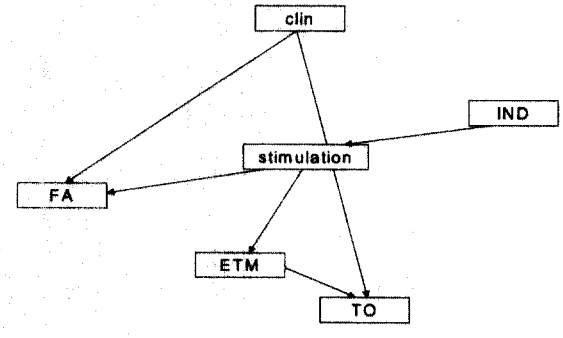
그림에서 검은 박스로 표현된 노드들이 Clin의 Markov blanket에 속한 노드들이다. 이와 같이 선택된 임신 요인들은 각각 Clin(임상적 임신여부), IND(증상), Stimulation(약물치료법), FA(여성의 나이), ICT(미세조작 난자수), ETM (Wallace사용여부), TO(총이식 수정란수) 등이다. 이러한 특징집합만을 포함하도록 실험데이터를 축소한 뒤, 새롭게 구한 베이지안 망 분류기들이 (그림 5)에 나타나 있다. (그림 5)의 (a)는 축소된 특징들로부터 구한 NBN인 NBNSF,



(a)



(b)



(c)

(그림 5) 축소된 특징집합으로부터 얻어진 베이지안 망 분류기들 : (a) NBNSF, (b) BANSF, (c) GBNSF

(b)는 BANSF, (c)는 GBNSF를 각각 나타낸다. 이들은 모두 원래의 NBN, BAN, GBN보다 노드와 아크가 줄어든 부분 그래프들이다. 한 가지 주목할 점은 이들과 같이 축소된 데이터 집합으로부터 다시 학습과정을 통해 구한 베이지안 망 그래프들은 원래 그래프들에서 단순히 선택되지 않은 특징 노드들과 아크들을 제거한 결과와는 다르다는 점이다. 예컨대 (그림 4)의 (c)의 GBN에서는 Clin과 stimulation간의 의존성을 나타내는 아크가 존재하는데 반해, (그림 5)의 (c)의 GBNSF에서는 이러한 아크를 찾아볼 수 없다.

5.2 분류 요인간 의존성 분석

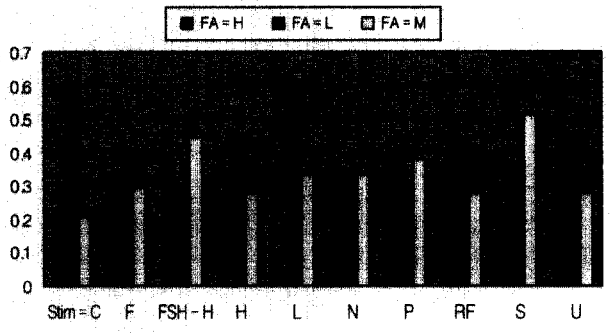
(그림 4)에서 생성된 베이지안 망들은 임신가능여부(Clin)에 직, 간접적으로 영향을 미치는 요인들로 나머지 8개의 특성 모두를 가정하고 그들의 의존관계를 구해낸 것인데 비해, (그림 5)의 베이지안 망들은 임신가능여부(Clin)에 보다 직접적으로 영향을 미치는 요인들을 Markov blanket에 포함되는 증상(IND), 약물치료법(stimulation), 여성의 나이(FA), 미세조작 난자의 수(ICT), Wallace 사용여부(ETM) 등 5개의 특성들로 한정하여 그들 간의 의존성을 구해낸 것이다. (그림 4)의 BAN과 GBN에서 뿌리 노드인 클래스 노드와 그 노드로부터 각 요인 노드들에 연결된 아크들을 제거하고 남는 부분 그래프는 서로 동일하며, 바로 이 그래프는 임신여부를 결정짓는 요인들간의 상호의존성을 나타내는 것이다. 특성이 축소된 (그림 5)의 BANSF와 GBNSF에서도 클래스 노드와 그 노드에서 시작되는 아크들을 제거하고 남는 부분 그래프는 바로 가장 임신여부에 직접적인 영향을 미치는 것으로 판단되는 요인들간의 상호의존성을 나타내는 것이다.

베이지안 망의 각 노드마다 부모 노드들에 대한 의존성 정도를 조건부 확률표(conditional probability table, CPT)로 표현하고 있는데, 이중 하나의 예를 들면 <표 4>와 같다. 이는 여성의 나이(FA)와 약물치료법(stimulation)에 따른 임상적 임신여부(Clin)를 조건부 확률로 나타내고 있다. <표 4>에서 보면 가임에 미치는 항목으로 여성의 나이가 20~34세인 경우 임신이 성공할 확률이 높으며, 그 중 약물치료법(stimulation)이 Long Protocol인 경우와 Clomiphene인 경우 가장 높았으며, 여성의 나이(FA)가 M(34~40)인 경우에는 약물치료법(stimulation)이 Short Protocol일때 가장 높은 확률을 나타냄을 알 수 있다.

또한 <표 4>를 근거로 하여 이것을 그래프로 표현한 것이 (그림 6)이다.

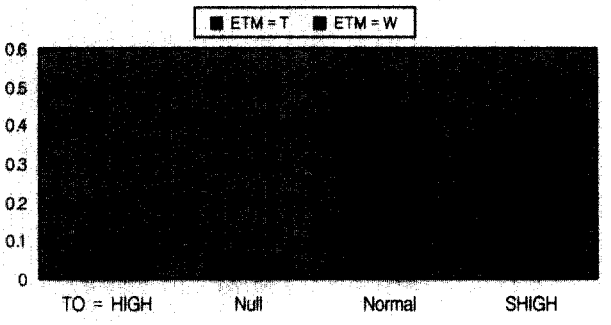
<표 4> 여성의 나이와 약물치료법에 기초한 임신가능성을 나타내는 조건부 확률표 P(Clin = true | FA, Stimulation)

Stimulation	FA = H	FA = L	FA = M
C	.2083333	**.5833333**	.2083334
F	.2962963	.4074074	.2962963
FSH-H	.2777778	.2777778	.4444444
H	.2777778	.4444444	.2777778
L	.0574713	**.6091954**	.3333333
N	.3333333	.3333333	.3333334
P	.3809524	.2380952	.3809524
RF	.2777778	.4444444	.2777778
S	.1282051	.3589744	**.5128205**
U	.2777778	.4444444	.2777778



(그림 6) 여성의 나이와 약물치료법에 기초한 임신가능성 P(Clin = true | FA,Stimulation)

또 다른 예로 Wallace 사용여부(ETM)과 총이식 수정란수(TO)에 따른 임상적 임신여부(Clin)의 조건부 확률을 그래프로 표현하면 (그림 7)과 같다.



(그림 7) Wallace사용여부와 총이식 수정란수에 기초한 임신가능성 P(Clin = truel ETM, TO)

여기서 우리는 Wallace사용여부(ETM)에 관계없이 총이식 수정란수(TO)가 Normal(1~5개) 또는 SHIGH(6~10개)일 경우 임신할 확률이 높았으며 수정란수가 없거나 HIGH(10개 이상)인 경우에는 극히 임신할 확률이 낮음을 알수 있었다.

5.3 분류 성능 비교

<표 5>는 대표적인 일반 분류기인 C4.5 결정트리(decision tree), k-최근접 이웃(k-nearest neighbors, k-NN) 방법들과 본 연구의 관심대상인 3가지 베이지안 망 분류기 NBN, BAN, GBN 등을 각각 동일한 훈련 데이터집합(244개)과 테스트 데이터집합(25개)에 대해 측정한 분류 정확도와, 또한

마지막으로 이 두 데이터집합을 합하여 전체 실험데이터에 대해 10회 교차 검증(10-fold cross validation)[14]을 시행하면서 측정한 평균 분류 정확도를 보여주고 있다. 예상한대로 동일 분류기에 대해 테스트 데이터집합에서의 분류정확도보다 직접 훈련에 사용된 훈련 데이터집합에서의 분류 정확도가 예외 없이 모든 경우 더 높게 나타났고, 10회 교차검증의 분류 정확도는 훈련 데이터집합의 경우보다는 낮으나 테스트 데이터의 경우보다는 약간 높은 성능을 보여주었다. 또 베이지안 망 분류기는 그 유형에 상관없이 모든 경우 결정트리와 k-근접 이웃방법에 비해 높은 분류 정확도를 보여주었다. 이것은 당초 의료 영역 데이터 특성을 고려할 때 베이지안 망 분류기가 다른 분류기보다 우수한 성능을 보이리라고 기대했던 것과 일치한 결과로 해석할 수 있다.

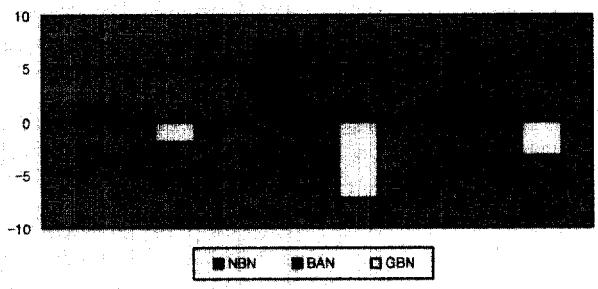
〈표 5〉 분류기별 분류성능 비교

Classifiers	By Train Dataset	By Test Dataset	10-Fold Cross Validation
DT(C4.5)	77.4%	64.3%	73.9%
k-NN (k=3)	78.9%	66.7%	74.8%
NBN	78.4%	67.1%	75.5%
BAN	**81.9%**	70.0%	78.8%
GBN	81.4%	**72.9%**	**79.2%**
NBNSF	79.9%	74.3%	78.4%
BANSF	**82.4%**	**71.4%**	**79.6%**
GBNSF	79.4%	65.7%	75.9%

3가지 서로 다른 유형의 베이지안 망 분류기들간의 분류성능을 비교해보면, NBN, BAN, GBN의 순으로 분류 성능이 증가하였다는 것을 알수 있다. 특히 요인들간의 자유로운 의존관계를 허용하는 BAN과 GBN이 그렇지 못한 NBN에 비해서 상당히 우수한 성능을 나타냄을 알수 있다. 하지만 요인들간의 직접적인 의존관계를 무시하고 모두 서로 독립성을 가정하는 NBN도 다른 일반 분류기에 비해 상당히 높은 성능을 나타낸 것은 주목할만 하다. 이는 확률론적 지식표현과 추론이 비 확률적 분류 예측방법에 비해 불확실성이 높은 의료분야의 데이터에서 효과가 있다는 것을 의미하는 것으로 해석할 수 있다.

5.4 특징축소 효과 분석

<표 5>에서는 특징 축소 이전의 NBN, BAN, GBN의 분류 성능과 더불어 Markov blanket내의 5개 특성들로 특성집합을 축소한 후 얻어진 NBNSF, BANSF, GBNSF 등의 분류성능도 나타나 있다. 특성 축소 이전의 NBN과 BAN에 비해 각각 특성 축소 이후의 NBNSF와 BANSF의 분류 성능이 모두 증가되었음을 알수 있다. 그리고 전체적으로 특성축소를 한 BANSF가 다른 모든 분류기들에 비해 가장 높은 성능을 보였다. 이것은 클래스 노드의 Markov blanket으로 특성집합을 축소하는 것이 이 의료 영역 데이터집합에서는 상당한 효과가 있었음을 보여주는 것이다. 하지만 특이하게 GBN의 경우만 특성이 축소된 GBNSF에서 오히려 분류 성능이 소폭 감소하였다는 사실을 발견할 수 있다. 이러한 현상은 GBN의 경우 분류 클래스를 별도로 두지 않고 있기 때문에 Clin에 대해 간접적 의존성을 갖는 요인들의 배제가 간접적 영향까지 배제한 결과로 이어져 정확도를 떨어뜨렸다고 판단된다.



(그림 8) 특징 축소 효과

6. 결 론

본 논문에서는 베이지안 망을 기초로 산부인과 불임환자의 임상 데이터에 대한 다양한 분석 실험을 전개하였다. 이 실험을 통해 베이지안 망에 드러난 임신여부에 영향을 주는 요인들간의 상호의존성을 분석해보았고, 또 NBN, BAN, GBN 등 제약조건이 다른 다양한 유형의 베이지안 망 분류기들의 분류성능을 서로 비교해보았다. 그리고 우리는 이와 같은 실험을 통해 임신가능여부(Clin)에 보다 직접적으로 영향을 미치는 요인들로 증상(IND), 약물치료법(stimulation), 여성의 나이(FA), 미세조작 난자의 수(ICT), Wallace 사용여부(ETM) 등 5개의 특성들을 가려낼 수 있었고, 이 요인들간의 상호의존성도 찾아낼 수 있었다. 또 본 논문에서는 실험을 통해 서로 다른 유형의 베이지안 망 분류기들 중에서 요인들간의 상관관계를 더 자유롭게 표현할 수 있는 BAN과 GBN들이 그렇지 못한 NBN에 비해 상대적으로 더 높은 분류 성능을 보여준다는 것을 확인하였다. 또 결정트리와 k-최근접 이웃과 같은 다른 분류기들과의 분류 성능 비교를 통해 의료 임상데이터의 특성상 속성들간의 관계에 대한 확률적 표현과 속성값에 대한 확률적 예측이 가능한 베이지안 망 분류기들이 보다 높은 성능을 보여준다는 사실도 확인할 수 있었다. 또 본 논문에서는 하나의 베이지안 망에서 클래스 노드의 Markov blanket에 속한 특성들로 특성집합을 축소하는 것이 베이지안 망 분류기들의 성능을 높여줄 수 있는지를 알아보기 위한 실험을 전개하였고 이를 통해 NBN과 BAN의 경우 그 효과를 입증할 수 있었다. GBN의 경우는 예상외로 성능이 저하되는 결과를 보여주었으나 그 원인은 간접적 의존성을 갖는 요인들이라도 GBN에서는 분류 클래스를 별도로 두지 않기 때문에 특징축소가 전체

의 정확도를 떨어뜨림을 확인할 수 있었다.

참 고 문 헌

[1] 대한산부인과학회, 부인과학(개정판), 도서출판 칼빈서적, 1991.

[2] 정혁, "불임, 무엇이 문제인가 -그 원인과 치료", 우리출판사, 1997.

[3] Bouckaert, R., "Bayesian Belief Networks : From Construction to Inference," Doctoral Dissertation, University of Utrecht, The Netherlands, 1995.

[4] Cheng, J., Bell, D. A. and Liu, W., "An Algorithm for Bayesian Belief Network Construction from Data," Proceedings of AI & STAT-97, Florida, pp.83-90, 1997.

[5] Cheng, J. and Greiner, R., "Learning Bayesian Belief Network Classifiers : Algorithms and System," Proceedings of the fourteenth Canadian conference on artificial intelligence, 2001.

[6] Cheng, J., "BN PowerConstructor," http://www.cs. ualberta.ca/~jcheng/bnsoft.htm.

[7] Dougherty, J., Kohavi, R., and Sahami, M., "Supervised and Unsupervised Discretization of Continuous Features," Proceedings of ICML-95, pp.194-202, 1995.

[8] Friedman, N., "Learning Bayesian Networks in the Presence of Missing Values and Hidden Variables," Proceedings of ICML-97, pp.125-133, 1997.

[9] Friedman, N., Linial, M., Nachman, I., Peter, D., "Using Bayesian networks to Analyze Expression Data," Journal of Computational Biology, 2000.

[10] Gorrill, Marsha-J. ; Kaplan, Paul-F. ; Patton, Phillip-E. ; Burry, Kenneth-A., "Initial Experience with Extended Culture and Blastocyst Transfer of Aryopreserved Embryos," American Journal of Obstetric & Gynecology, Vol.180, No.6, 1999.

[11] Heckerman, D., "A Tutorial on Learning Bayesian Networks," Technical Report MSR-TR-95-06, Microsoft Research, 1995.

[12] Heckerman, D., Meek, C. and Cooper, G., "A Bayesian Approach to Causal Discovery," Technical Report MSR-TR-97-05, Microsoft Research, 1997.

[13] Jensen, F. V., An Introduction to Bayesian Networks, New York : Springer-Verlag, 1996.

[14] Jiawei Han, Micheline Kamber, Data Mining : Concepts and Techniques, Morgan Kaufmanm. 2001.

[15] Kevin Patrick Merphy, "A Brief Introduction to Graphical Models and Bayesian Networks," Technical Report, Department of Computer Science, UC Berkley, 2001.

[16] Kohavi, R. and John G., "Wrappers for Feature Subset Selection," Artificial Intelligence, Special Issue on Relevance, Vol.97, No.1-2, pp.273-324, 1997.

[17] Langley, P. and Sage, S., "Induction of Selective Bayesian Classifiers," Proceedings of UAI-94, 1994.

[18] Pazzani, M. J., "Searching for Dependencies in Bayesian Classifiers," Proceedings of AI & STAT-95, 1995.

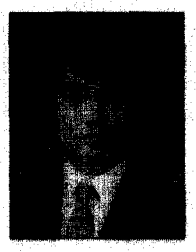
[19] Pearl, J., Probabilistic Reasoning in Intelligent Systems, Morgan Kaufmanm, 1988.

[20] Provan, G. M. and Singh, M., "Learning Bayesian Networks Using Feature Selection," Learning from Data, Lecture Notes in Statistics, Berlin : Springer-Verlag, Vol.112, pp. 291-300, 1996.

[21] Singh, M., "Learning Bayesian Networks from Incomplete Data," Proceedings of AAAI-97, The MIT Press, pp.534-539, 1997.

[22] Sprites, P., Gleymour, C., and Sceines, R., Causation, Prediction, and Search, New York : Springer-Verlag, 1993.

[23] Tom M. Mitchael, Machine Learning, McGrow-Hill, 1997.



정 용 규

e-mail : ygjung@shjc.ac.kr

1981년 서울대학교 물리교육과(이학사)

1994년 연세대학교 전자계산전공 (공학석사)

2002년 경기대학교 전자계산학과 (박사수료)

1999년~현재 서울보건대학 전산정보처리과 교수

관심분야 : 데이터마이닝, 의료정보시스템

김 인 철

e-mail : kic@kyonggi.ac.kr

1987년 서울대학교 대학원 전산과학과 (이학석사)

1995년 서울대학교 대학원 전산과학과 (이학박사)

1989년~1995년 경남대학교 전산통계학과 조교수

1996년~현재 경기대학교 정보과학부 전자계산학전공 부교수

관심분야 : 지능형 에이전트, 분산인공지능, 데이터마이닝