베이지안망을 이용한 불임요인 분석 및 가임예측

정용규[0] 진 훈 김인철
경기대학교 전자계산학과
ygjung@shjc.ac.kr, {jinun, kic}@kyonggi.ac.kr

Analysis of Infertility Factors and Prediction of Pregnancy Using Bayesian Networks

Yong-Gyu Jung[0] Hoon Jin In-Cheol Kim
Dept. of Computer Science, Kyonggi University

요 약

의료 분야에서의 데이터는 특성상 여러 측면을 복합적으로 고려해야 할 뿐만 아니라, 다른 분야에서의 데이터 성격과는 다르게 원인과 그 원인에 대한 해결책을 바로 찾아내기가 쉽지 않다. 본 연구에서는 불임환자들에 대한 검사기록 및 임신결과가 기록된 데이터를 이용하여 베이지안망 분류기를 생성하고 이를 통해 가임여부를 결정짓는 중요 항목 들간의 의존성을 조건확률로 나타내고 비교하였다. 또한 휴리스틱망, 나이브베이지안망 분류기를 생성하여 성능을 비교하였다. 결과적으로 총수정란수는 최상급수정란이식수에 강한 영향을 갖는다는 사전지식의 타당함을 입증할 수가 있었으며, 또한 성숙난자수가 총수정란수에 강한 영향을 미치고 화학적임신과 임상적임신과 학습은 서로 독립이라는 가설에 대하여 전자의 경우는 간접적인 의존성을 갖고, 후자의 경우는 화학적 결과가 임상적 결과에 강한 의존성이 존재함을 밝혀낼 수 있었다. 분류기 간의 성능에서는 자동생성된 베이지안망이 가장 우수한 정확도를 가짐을 측정할 수 있었다.

1. 서론

최근 각종 공해와 스트레스 등에 의해 불임증 환자의 수는 점점 증가하는 추세이며 이는 전체 가임 연령에 있는 성인의 10~15%가 불임환자로 알려져 있다. 불임이란 부부가 피임을 하지 않고 정상적인 부부생활을 함에도 불구하고 임신이 되지 않는 경우를 뜻한다. 불임증은 보통 복합적 원인이 많으며 통계상 불임환자의 약 35%가 복합적 원인에 의한다[1]. 그러므로 불임환자의 경우, 현 의료체계에서 문진(問診)상태만을 가지고 가임의 가능성을 확보하기는 어렵다. 또한 불임환자의 각종 검사와 시술과정에서 얻은 정보 들간의 상관 관계를 규명하지 않은 상태에서 임신의 여부를 예측하기란 더욱 어려운 일이다. 그러므로 지금까지의 시술은 담당하는 의료전문가의 오랜 경험과 방법들을 통한 직관에 의존할 수 밖에 없었고, 이렇게 쌓이는 지식들은 공개적인 형태로 자료화되지도 않는 형편이다. 이에 본 연구는 불임환자의 검사항목과 시술과정에서 얻은 정보에 대하여 종합적으로 상관관계를 규명함으로써 불임요인을 분석하고 가임여부에 대하여 예측하고자 하였다. 실험을 위해 서울 모 종합병원에 2년 동안 래원(來院)한 400여명의 불임환자들을 대상으로 조사한 검사항목, 시술방법, 시술결과가 기록된 데이터와 충분하지 못한 사전 지식을 입력 자료로 사용하였다. 그리고 이 자료에 베이지안망 알고리즘을 적용하여 상관관계를 분석하였는데, 이는 베이지안망 알고리즘이 불확실한 지식체계 속에서 항목들 간의 상관관계를 분석하여 그래프 구조로 보여줄 수 있다는 장점을 가지기 때문이다. 또한 항목들 간의 상관관계를 독립이라 가정하는 나이브베이지안 알고리즘을 적용한 결과와도 비교분석하여 베이지안망의 우수성을 입증할 수 있었다. 다음 장에는 베이지안망에 대해 간략히 소개하였고 3장에서 데이터 사전처리, 분류기학습, 분류실험결과를 제시하였다. 그리고 4장에서 결론을 기술하였다.

2. 베이지안망

베이지안망이란 관측하려는 대상들을 노드로 표현하고 대상들 간의 순서적, 관계적 의미를 링크로 표현하여 노드를 연결한 그래프이며 다음과 같은 특징을 갖는다.

■ 베이지안 망 알고리즘의 특징
1. 지식 표현이 용이하다.
2. 불확실한 지식조건 하에서 추론이 가능하다.
3. 항목 들간의 관련성 또는 (원인-영향) 관계를 밝히는데 적합하다.
4. 지식체계를 그래프적 구조로 표현함으로써 이해하기 쉽고 수정이 간편하다.

베이지안망에 관한 연구는 데이터로부터 그래프적 모델(Bayesian Belief Network)을 학습하기 위해 두 가지 방향에서 다르게 시도되었다. 하나는 Chow와 Liu의 검색점수측정방법(search & scoring method)이며 다른 하나는 Wermuth와 Lauritzen의 의존성분석방법(dependency analysis method)이다. 후자의 경우 전자가 여러 요소(Bayesian, MDL 또는 KL)의 엔트로피(entropy)를 측정하여 이를 최대화할 수 있는 구조를 학습하는데 반하여, 노드들 간의 조건독립적(Conditional

Independence) 관계를 'd-separation'의 개념에 따라 측정하여 학습하는 방법이다. 각기 장단점이 있지만 본 논문에서는 후자의 방법을 이용하였다. 또한 학습을 통해 생성되는 베이지안망 분류기에도 그 형태에 따라 Naïve-Bayes, Tree-Augmented Naïve-Bayes(TAN), BN Augmented Naïve-Bayes(BAN), Bayesian Multi-Net, General Bayesian Network(GBN)로 나눌 수 있다[2].

3. 데이터 사전처리

불임요인은 크게 남성요인과 여성요인, 면역학적요인, 원인불명으로 나뉠 수 있다. 남성요인은 주로 정자의 수와 운동성에 관련이 있으며 여성요인은 난소요인, 난관요인, 자궁경부요인, 자궁요인, 복막요인으로 분류할 수 있다[1]. 본 논문에서는 40여 가지의 검사항목 중에서 중요도가 낮은 항목, 특이값 발생 항목, 그리고 의존성을 전혀 예측할 수 없는 항목에 대하여 관련 전문가의 조언과 휴리스틱한 방법을 이용하여 데이터를 정제하였다. 이렇게 선택된 특징(feature)들은 난소 및 자궁요인과 이에 대한 시술방법, 그리고 최종 임신여부 항목이다. 결과적으로 총 384개의 데이터집합에 대한 12가지 특징들이 선택되었고 항목 간의 의존성(시간적 발생순서, 간접적 원인관계, 직접적 원인관계)을 조사하여 [표 1]과 같이 작성하였다.

표 1 실험데이터 집합

기호	특징명	종속성		
		시간	직접원인	간접원인
Eten	자궁내막두께	0		>Etn
Zth	투명대두께	4	>Ah	
Mature	성숙난자수	1	>Ter, Tpl	>hCG, Clin
MI	성숙직전난자수	1		
Ter	총수정란수	2	>G1, G2	>hCG, Clin
Tpl	총다전핵수정란수	2		
G1	최상급수정란이식수	4	>hCG, Clin	
G2	상급수정란이식수	4	>hCG, Clin	
Etn	이식방법	3	>hCG, Clin	
Ah	보조부화술	5		>Etn
HCG	화학적임신유무	6		
Clin	임상적임신유무	6		

[표 2]는 [표 1]에 나타나는 각 항목들에 대한 설명이다.

표 2 항목 설명

자궁내막	자궁의 점막을 의미한다.
투명대	배란 시에 배출되는 난자를 둘러싸고 있는 물질이다.
이식방법	수정란을 어떤 방법으로 이식할지를 결정한다.
보조부화술	투명대의 두께를 얇게 하여 수정을 돕는 시술이다.
화학적임신	화학적 물질(hCG)의 분비량을 측정하여 임신여부를 판단한다.
임상적임신	물리적 촬영 등을 통하여 임신여부를 판단한다.

또한 [표 1]에서 '시간'이란 시간적 종속성을 의미하는데, 즉 선행되는 특징과 이를 따라 이루어지는 후행 특징으로 구분이 가능하다. '직접 원인'이란 선행 특징과 후행하는 특징 간의 (원인-결과) 관계가 존재함을 나타내며 '간접 원인'이란 선-후행 특징 간에 비 직접적이지만 임시적 순서의 의미, 또는 간접적 영향을 미치는 관계가 성립함을 나타낸다.

본 실험데이터의 경우 대부분의 항목값들이 수치표현으로 이루어져 있고 한 항목에 대하여 다수의 값들이 존재한다. 그러므로 데이터를 이산화시키는 과정이 필요하다. 이산화과정에는 Equal Width Interval Bining, Holte's 1R Discretizer, Recursive Minimal Entropy Partitioning 방법이 있다[3]. 우리는 특징을 고려하여 두번째 방법을 사용하였고 다음과 같이 데이터를 변환하였다.

{hCG, Clin : 임신(1), 불임(0)}

ETen	Mature	MI	Etn	TER	TPL	AH	Zth	G1	G2	hCG	Clin
M	L	L	W	M	ML	A	R	L	M	0	0
M	L	L	T	M	ML	A	R	L	M	0	0
M	L	L	W	L	L	A	R	L	L	1	1
ML	L	LL	T	L	M	A	R	H	L	1	1
MI	L	H	T	L	H	A	R	M	M	0	0

그림 1 정제된 실험 데이터

[그림 1]에서 LL/L/ML/M/MH/H/HH/VHH/R/NR은 각각 Lowest/Low/Low/Middle Low/Middle/Middle High/High/Highest/Very Highest/Right/Non-Right를 나타낸다.

4. 베이지안망 분류기 학습

실험을 위해 총 384개의 데이터집합을 285개의 훈련데이터집합(training set)과 99개의 검증데이터집합(test set)으로 구분하였고 분류기 생성을 위해서 Cheng, J의 BNPowerPredictor를 이용하였다[2].

먼저 [표 1]을 이용해 휴리스틱으로 생성한 베이지안망 그래프는 훈련데이터집합을 이용해 생성된 분류기의 구조는 [그림 2]이고 훈련데이터집합을 이용해 자동으로 생성한 그래프는 [그림 3]이다.

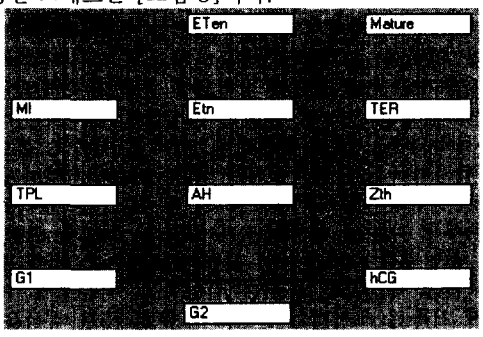

그림 2 휴리스틱을 이용한 베이지안망 그래프

[그림 3]은 [그림 2]와 비슷하면서도 다른 형태로 나타나는 것을 알 수 있다. 이는 불충분한 사전지식과 실제 관측 데이터의 차이로 이해할 수 있다.

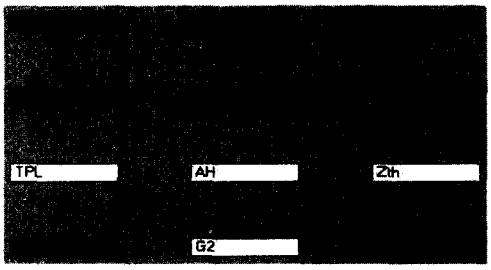

그림 3 자동생성된 베이지안망 그래프

[그림 3]에서 전체적인 불임 확률은 0.729, 가임 확률은 0.270이 됨을 알 수 있고 각 특징들에 대해 조건부 확률을 구해보면 MI의 경우는 10개, TER은 120개, G1은 12, hCG는 2개의 엔트리(entries)를 가짐을 있다.

표 3 G1의 조건부 확률

Clin	TER	H	HH	L	M
N	H	.3863636	.1136364	.2954545	.2045455
	HH	.1785714	.1785714	.1785714	.4642858
	L	.0709459	.0168919	.5709459	.3412163
	LL	.0320513	.0320513	.8269231	.1089743
	M	.125	.027439	.4420732	.4054878
	MH	.2788461	.125	.2788461	.3173078
Y	H	.3541667	.1875	.1875	.2708333
	HH	.25	.25	.25	.25
	L	.1413043	.0978261	.4456522	.3152174
	LL	.1785714	.1785714	.3214286	.3214286
	M	.1838235	.125	.0955882	.5955883
	MH	.29	.13	.21	.37

사전지식을 고려할 때 가임여부에 중요한 영향을 미치는 특징은 G1이다. 이러한 G1은 사전지식에서도 TER에 직접적인 의존성이 있었는데 실제 조건부 확률을 계산한 결과 동일하게 강한 의존적인 관계를 가지는 것을 알 수 있었다. 그리고 [표 1]에서 Mature가 TER에 강한 영향을 미치고, hCG와 Clin은 서로 독립으로 가정할 수 있다. 그렇지만 [그림 3]을 통해서 Mature는 TER에 대하여 MI를 통한 간접적인 관련성이 있을 뿐이며, hCG가 Clin에 강하게 의존적임을 알 수 있다. 반면에 실제 현실에서는 Zth를 판정하여 AH를 결정짓는데 대해 학습된 분류기에서는 서로 독립적인 관계로 나타남을 알 수 있다. 이는 분류기 생성에 사용되는 훈련데이터집합이 이와 같은 속성을 구별할 수 있도록 반영하고 있지 않는 경우이기 때문이다.

5. 분류실험

조건부 확률을 이용하여 99개의 테스트데이터집합에 대하여 실험한 결과 95%의 신뢰도 구간에서 95.96%의 높은 확률을 나타냄을 알 수 있었다(오차범위:3.88%). 이는 [그림 2]를 가지고 실험한 정확도인 77.20%(오차범위: 8.81%)에 비해 대단히 높은 수치임을 알 수 있다.

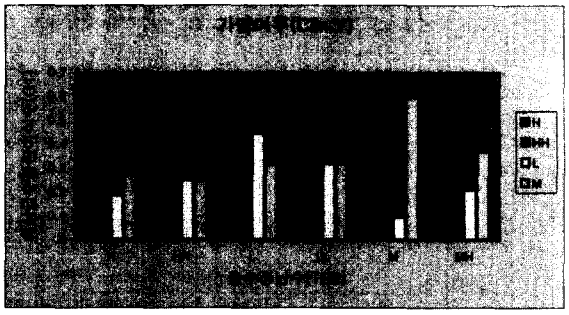

그림 4 TER과 G1의 상관관계표(가임상태)

정확도 차이를 나타내는 주된 원인은 최근 많이 사용되는 베이지안망의 구조가 BAN을 확장한 형태인 BMN인 것을 볼 때, 사전지식을 주로 (root-leaf)노드순, 완전순서(complete order)와 링크없음(forbidden)노드를 주된 입력으로 처리하기 때문인 것을 추측할 수 있다.
또한 특징들 간의 상관관계를 조건독립이라 가정하는 나이브베이지안망(Naïve-Bayesian Network)을 생성하여 정확도를 측정한 결과 91.49%(오차범위:5.64%)의 높은 성능을 나타냄을 알 수 있었다.

6. 결론

본 연구에서는 실제 불임환자들을 대상으로 한 검사결과 데이터를 가지고 베이지안망 분류기를 생성한 후 이를 사전지식과 조건부 확률을 기반으로 비교분석하였고 또한 휴리스틱망 및 나이브베이지안망과도 분류성능을 비교하였다. 결과적으로 보면 클래스를 정의한 상태에서 자동생성된 베이지안망이 휴리스틱망보다는 훨씬 우수한 성능을 나타내고, 나이브베이지안망보다는 근소하게 우위에 있음을 알 수 있었다. 본 논문은 의학연구에서 이루어지는 데이터마이닝 관련 연구들이 대부분 두 가지의 항목들을 서로 비교하여 상관관계를 찾고자 하는 데에 반하여, 첫째로 종합적인 항목들을 고려하여 그들 간의 의존성을 찾고 이를 도식적으로 표현함으로써 관련 전문가로 하여금 쉽게 망 구조를 이해하고 수정할 수 있도록 하였다는 것과 두번째로 임신여부에 강한 영향을 주는 항목간의 의존성을 실제 데이터를 토대로 조건확률에 의거하여 계산하고 나타냄으로써 사전적 지식과 일치성, 이질성을 제시하였다는데에 의의를 갖는다.

참고 문헌

[1] 대한산부인과학회, *부인과학(개정판)*, 도서출판 칼빈서적, pp.389-436, 1991.

[2] Cheng, J. and Greiner, R., "Learning Bayesian Belief Network Classifiers: Algorithms and System", *Proceedings of the 14th Canadian conference on artificial intelligence*, 2001.

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