의료데이터마이닝을 위한 특징축소와 베이지안망 학습

정용규
서울보건대학 전산정보처리과
ygjung@shjc.ac.kr

Features Reduction and Baysian Networks Learning for Medical Datamining

Yong-Gyu JUNG
Dept. of Computer Inform., Seoul Health College

요 약

본 연구에서는 베이지안망을 기초로 불임환자의 임상 데이터에 대한 다양한 실험을 전개한다. 실험을 통해 임신여부에 영향을 주는 요인들간의 상호 의존성을 분석하고, 또 제약조건이 다른 다양한 베이지안망의 대표적 유형으로 나이브 베이지안망(NBN), 베이지안망으로 확장한 나이브 베이지안망(BAN), 일반 베이지안망(GBN) 분류기들의 분류성능을 서로 비교 분석한다. 베이지안망을 적용할 때 변수의 수가 많아짐에 따라 베이지안망의 구조를 학습하는데 탐색공간이 넓어져 시간의 요구량이 급격히 많아진다. 따라서 이런 탐색공간을 효율적으로 줄이기 위하여 클래스 노드의 Markov blanket에 속한 특징들로 집합을 축소하는 것을 제안하고, 실험을 통해 이 특징 축소 방법이 베이지안망 분류기들의 성능을 높여 줄 수 있는지 알아본다.

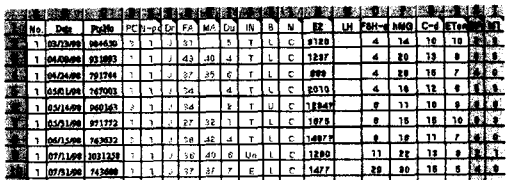
1. 서론

의료데이터의 실험적 분석은 수많은 원인과 결과들의 인과관계를 설명하며 이는 과거로부터 가장 많이 사용되어온 방법이다. 결국 원인들 간의 상관관계는 확률적인 분포를 갖고 있으며 우리가 결국 밝혀내려는 분류 클래스들도 여러 특징에 의해 확률적 분포를 갖는다. 의사들의 진료행위와 치료결과에 근거하여 이런 방법은 가장 효율적인 치료패턴을 찾아주므로, 이에 근거하여 진료의 방법을 결정할 수 있고, 또한 의료의 질 향상을 도모할 수 있게 한다.

본 연구에서는 병원에 래원한 환자들에 대한 실재의 임상 데이터로부터 불임과 관련된 특징들 간의 의존성을 표현하고 분석하는데 베이지안망(Bayesian network)을 적용해 보고자 한다. 실험을 통해 베이지안망의 여러 유형들을 학습하고 이를 바탕으로 영역지식을 표현한다. 이런 영역지식이 기존의 전문가들이 갖고 있는 지식과 일치하는지도 살펴본다. 또한 베이지안망 학습은 대상이 되는 특징들의 수에 많은 시간적 부하를 주게 되므로 특징들의 수를 줄이는 방법에 대해서도 연구해 본다. 이러한 연구를 통하여 베이지안망이 변수들 간의 확률관계를 축약된 형태로 표현하는데 최적의 모델로서, 확률적 추론, 예측, 의사결정을 요하는 분야에 잘 적용되는 분류기임을 의료영역에서 확인하기 위하여 임상 데이터를 분석한다.

2. 실험데이터

본 연구에서 이용하게 될 실제의 의료 임상 데이터는 서울에 소재한 모 종합병원의 산부인과에 2년 동안 래원한 불임환자들의 검사기록과 시술과정, 그리고 시술 결과로 얻어진 임신의 성공여부가 담긴 데이터이다. 불임환자는 불임원인을 먼저 파악하기 위해 각종 검사를 하게 되고 원인별로 불임시술의 방법을 선택하게 된다. 이런 환자의 정보들은 차트에 의해 자세히 기록되어 있으나, 우리는 실험을 위하여 관련된 정보를 요약하여 정리하게 되었는데, 그 결과는 [그림1]과 같다.



[그림1] 수집된 실험데이터 일부

불임요인은 크게 남성요인과 여성요인, 면역학적 요인, 원인불명으로 나뉠 수 있다. 남성요인은 주로 정자의

수와 운동성에 관련이 있으며 여성요인은 난소요인, 난관요인, 자궁경부요인, 자궁요인, 복막요인으로 분류할 수 있다. 본 연구에서는 40여 가지의 검사항목 중에서 중요도가 낮은 항목, 특이값 발생 항목, 그리고 의존성을 전혀 예측할 수 없는 항목에 대하여 휴리스틱 방법을 이용하여 데이터를 정제하였다. 이렇게 선택된 특징들은 <표1>과 같이 9개가 선택되었다.

<표1> 실험 데이터집합

코드값	특징이름	설명
Clin	임상적임신여부	초음파등을 통한 임신의 성공여부
FA	여성의 나이	여성의 실제나이
ETD	이식일수	수정후 자궁내 착상까지 경과일수
ETM	Wallace사용여부	보조부화술의 사용여부
Stim	약물치료법	배란을 촉진하기 위한 약물투여
TO	총이식 수정란수	이식된 수정란의 수
ICT	미세조작난자수	미세조작을 통한 수정된 난자수
IVF	시술방법	시험관아기 시술법
IND	증상	불임의 원인

3. 전처리

일반적으로 데이터 분석을 위한 전처리 작업으로는 특징을 축소(dimension reduction)하는 방법 이외에도 필요한 경우 데이터 정제(cleaning), 변환(transformation), 이산화(discretization) 등이 적용될 수 있다. 정제작업은 누락항목 데이터와 잡음(noise) 등이 포함된 데이터들을 채우거나 삭제하는 방법으로 본 연구를 위해 수집된 데이터의 경우에도 이런 경우들이 많아 이들을 처리하기 위한 데이터 정제작업이 수행되었다.

누락 항목이 많은 미세 조작술, 보조도구 사용유무, 총이식 수정란수 등은 정상인 경우 기입하지 않은 경우들이 대부분을 차지하고 있어, 정상값 또는 평균값으로 적용하였다. 이식일수, 총이식 수정란수의 데이터 항목은 빈도수가 현저히 떨어지는 값 또는 특이값이 있는 경우가 많은데, 이들은 해당 레코드를 지우는 방법으로 정제작업을 하였다. 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 연속수치 데이터 값들을 많이 포함하고 있어 적절한 이산화 작업이 필요하였다.

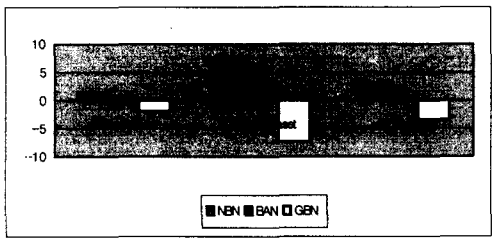
4. 특징축소를 통한 성능개선

본 연구의 관심대상인 3가지 베이지안망의 기본분류기는 나이브 베이지안망(NBN), 베이지안망으로 확장한 나이브 베이지안망(BAN), 일반 베이지안망(GBN)이다. 본 연구에서는 NBN, BAN, GBN를 각각 동일한 훈련 데이터 집합(244개)과 테스트 데이터 집합(25개)에 대해 측정한 분류 정확도와 비교하였다. 또한 마지막으로 이 두 데이터 집합을 합하여 전체 실험 데이터에 대해 10회 교차 검증(10-fold cross validation)을 시행하면서 측정한 평균 분류 정확도를 <표2>에서 보여주고 있다. <표2>에 의하면 동일 분류기에 대해 테스트 데이터 집합에서의 분류 정확도보다 직접 훈련에 사용된 훈련 데이터 집합에서의 분류 정확도가 예외 없이 모든 경우 더 높게 나타났고, 10회 교차검증의 분류 정확도는 훈련 데이터 집합의 경우보다는 낮으나 테스트 데이터의 경우보다는 약간 높은 성능을 보여 주었다. 3가지 서로 다른 유형의 베이지안망 분류기들간의 분류성능을 비교해보면, NBN, BAN, GBN의 순으로 분류 성능이 증가하였다는 것을 알 수 있다. 특히 특징들간의 자유로운 의존관계를 허용하는 BAN과 GBN이 그렇지 못한 NBN에 비해서 상당히 우수한 성능을 나타냄을 알 수 있다. 하지만 특징들간의 직접적인 의존관계를 무시하고 모두 서로 독립성을 가정하는 NBN도 다른 일반 분류기에 비해 상당히 높은 성능을 나타냈다.

<표2> 분류기별 분류성능 비교

Classifiers	By Train Dataset	By Test Dataset	10-Fold Cross Validation
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%

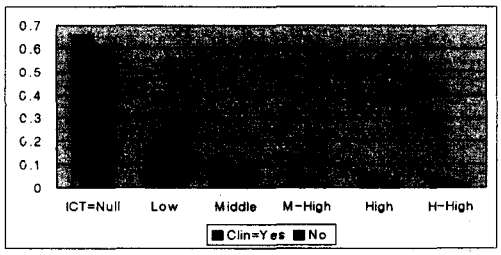
이제까지 우리는 특징 축소 이전의 NBN, BAN, GBN의 분류 성능과 더불어 Markov blanket내의 5개 특징들로 특징집합을 축소한 후 얻어진 NBNSF, BANSF, GBNSF 등의 분류성능에 대해서도 살펴본다. 특징 축소 이전의 NBN과 BAN에 비해 각각 특징 축소 이후의 NBNSF와 BANSF의 분류 성능이 모두 증가되었음을 알 수 있다. 특징을 축소한 BANSF가 다른 모든 분류기들에 비해 가장 높은 성능개선을 보였다. 이것은 클래스 노드의 Markov blanket으로 특징 집합을 축소하는 것이 의료분야 데이터 집합에서는 상당한 효과가 있었음을 보여주는 것이다. 하지만 특이하게 GBN의 경우만 특징이 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 특징들의 배제가 어느 정도의 영향을 주고 있었던 간접적 영향을 배제함으로써 결과에 부정적으로 영향을 미쳐 그 결과로 정확도를 떨어뜨렸다고 설명이 된다.



[그림2] 특징축소 성능개선 효과 (단위:%)

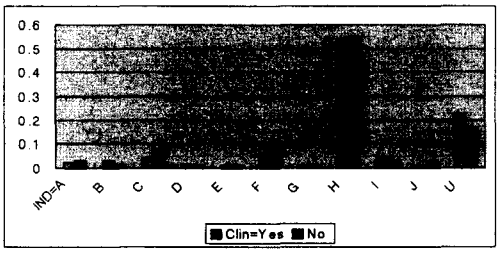
5. 특징축소를 통한 의존성 변화

실험의 결과 중 몇 가지 예를 들면 미세조작 난자수가 임신에 미치는 특징을 분석한 결과 NBN과 NBNSF의 결과값은 같게 나타났다. 이는 실제의 상황을 많이 고려치 않은 NBNSF의 유형에서는 NBN유형에서의 성능 개선을 기대하기가 어렵다. 제약조건이 많은 나이브 베이지안 망에서는 특징의 축소가 학습의 결과에 영향을 주지 않은 듯 보이지만, 성능에서는 많은 개선을 보였다.



[그림3] 미세조작 난자수와 임신여부의 종속성

[그림3]은 NBN과 NBNSF에서 동일한 학습결과를 생성한 하나의 예를 보여주고 있다. 특징을 축소한 학습의 결과가 기본유형과 동일한 결과를 갖은 것은 학습결과의 성능이 좋았음을 알려주고 있다.



[그림4] BAN에서 증상과 임신여부의 종속성

BAN에서의 실험의 결과를 살펴보면 [그림4]에서 나타내 주고 있듯이 증상과 임신의 종속성을 나타내는 학습의 결과에서 증상이 H(Tubal, 난관이상)인 경우 가장 의존성이 높았다. 이러한 난관이상도 체외수정을 통하여 보조도구를 이용한 이식방법을 사용하면 임신에 이르게 되어, 불임환자중 가장 의존성이 높게 나타난 것으로 보인다. 특징을 축소한 BANSF의 경우에도 비슷한 결과를 얻을 수 있었다.

6. 결론

본 연구에서는 실험을 통해 베이지안망에서 임신여부에 영향을 주는 특징들 간의 상호 의존성을 분석해 보았고, 또 NBN, BAN, GBN 등 제약조건이 다른 다양한 유형의 베이지안망 분류기들의 분류성능을 서로 비교해 보았다. 그리고 이와 같은 실험을 통해 임신여부에 보다 직접적으로 영향을 미치는 특징들로 증상, 약물 치료법, 여성의 나이, 미세 조작 난자의 수, Wallace 사용여부 등 5개의 특징들을 가려낼 수 있었고, 이 특징들 간의 상호 의존성도 찾아 낼 수 있었다. 실험을 통해 서로 다른 유형의 베이지안망 분류기들 중에서 특징들 간의 상관관계를 더 자유롭게 표현할 수 있는 BAN과 GBN들이 그렇지 못한 NBN에 비해 상대적으로 더 높은 분류 성능을 보여준다는 것을 확인하였다.

또한 본 연구에서는 하나의 베이지안망에서 클래스 노드의 Markov blanket에 속한 특징들로 축소하는 것이 베이지안망 분류기들의 성능을 높여 줄 수 있는지를 알아보기 위한 실험을 전개하였고 이를 통해 NBN과 BAN의 경우 성능개선이 뚜렷하게 나타남을 확인하였다. GBN의 경우는 특징을 축소한 학습이 오히려 성능을 저하시키는 결과를 보여 주었다. 이는 간접적 의존관계도 전체 성능에 영향을 주고 있음을 보여준 결과로 해석된다.

참 고 문 헌

[1] Cheng, J., Bell, A. and Liu, W., "Learning Belif Networks from Data: An Information Theory Based Approach", Proceedings of ACM CIKM-97, 1997

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

[3] Jung, Yong Gyu and Kim, In-Cheol, "Learning Bayesian Network for Medical Data Analysis", Proceedings of 2nd ICIS-02, pp.307-312, 2002

[4] Mitchell, T, Machine Learning, McGraw-Hill, 1997

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