효율적인 의료데이터마이닝을 위한 특징축소와 베이지안망 학습

정용규* · 김인철
서울보건대학 · 경기대학교
ygjung@shjc.ac.kr* · kic@kyonggi.ac.kr

Features Reduction and Baysian Networks Learning for Efficient Medical Data Mining

Yong Gyu Jung* · In Cheol Kim
Dept. of Computer Information Processing, Seoul Health College
Dept. of Computer Science, Kyonggi University

요 약

베이지안망은 기존의 방법에 비해 불확실한 상황에서도 지식을 표현하고 결론을 추론하는데 유용한 것으로 알려져 있다. 본 논문에서는 대표적인 베이지안망 분류기들을 제시하고, 동일 임상데이터에 대해 서로 다른 유형별 베이지안망 분류기들을 학습하였다. 베이지안망을 적용할 때 변수의 수가 많아짐에 따라 베이지안망의 구조를 학습하는데 탐색공간이 넓어져 어려움이 있다. 본 연구에서는 이런 탐색공간을 효율적으로 줄이기 위하여 클래스 노드의 Markov blanket에 속한 특징들로 집합을 축소하는 것을 제안하고, 실험을 통해 이 특징 축소방법이 베이지안망 분류기들의 성능을 높여 줄 수 있는지 알아 보았다. 분류기들의 성능에서는 축소한 특징집합으로부터 얻은 베이지안망으로 확장한 나이브 베이지안망 분류기가 가장 우수한 정확도를 가짐을 실험을 통해 알 수 있었다.

Key words : 베이지안망(Baysian Networks), 특징축소(Features Reduction), 불임(Infertility), Markov blanket

1. 서 론

의료분야에서의 환자에 대한 정확한 데이터 분석은 치료의 방법을 결정해야 하는 의사들에게 매우 중요한 일이다. 많은 경우 의사들은 과거의 환자기록을 토대로 치료의 방법을 선택하게 된다. 의사의 판단력을 기초로 치료의 방법이 결정하게 되는데, 불임시술의 경우 치료기간이 길고 임신에 이르기 위한 변수가 많은 관계로 의사의 판단력이 무엇보다 더 중요하다. 특히 불임에 영향을 미치는 요인들은 대부분 원인불명 또는 복합적인 경우가 많기 때문에 최적의 치료방법을 선택하기가 더욱 어렵다. 이런 불확실한 상황에서 확률적 판단을 해야 하는 의료분야에서 적용할 수 있는 확습 방법을 찾기 위하여 불임환자의 임상 데이터에 대한 다양한 분석 실험을 전개하였다. 본 연구에서 사용한 데이터는 임상에서 얻은 실제의 불임환자들에 대한 검사기록 및 시술과정이 기록된 데이터이다.

베이지안망은 여러 가지 변수들간의 확률적 관계를 표현하는 그래픽 모델이다. 우리는 베이지안망을 의료분야, 그 중에서도 불임과 관련된 요인들간의 의존성을 표현하고 분석하는데 이용하고자 한다. 일반적으로 베이지안망은 전산학적인 측면에서 여러 관심 대상들 간의 의존성에 대해 확률적으로 표현함으로써 그 구조 및 관계를 잘 표현할 뿐만 아니라 수학적인 면에서도 견고성이 입증된 통계적인 도구라고 할 수 있다. 그리고 출력 값에 대하여 단언적인 결과를 내는 타 알고리즘들과는 달리 확률을 기반으로 하여 입력 값 및 출력 값들을 조건적인 의존관계로 표현함으로써 누구나 쉽게 그 구조와 관계를 파악할 수 있는 장점이 있다. 또한 해당 영역 지식의 입력으로 사용할 수 있어 이미 많은 연구에서 그 우수성을 입증한 도구라고 할 수 있다[4].

이러한 베이지안망은 우리가 적용하려는 의료분야에 그 특징이 잘 맞는 것으로 판단이 된다. 의료데이터의 분석은 수많은 원인과 결과들의 인과관계를 설명할 수 있어야 하며 이는 과거로부터 가장 많이 사용되어온 방법이라 할 수 있다. 결국 원인들간에 서로 상관관계는 통계적인 분포를 갖고 있으며 우리가 결국 밝혀 내려는 분류 클래스들도 여러 특징에 의해 통계적 분포를 갖는다. 이런 의사들의 진료행위와 치료결과에 근거하여 가장 효율적인 치료패턴을 찾아주므로, 이에 근거하여 진료의 방법을 결정할 수 있고, 또한 의료의 질 향상을 도모할 수 있게 한다.

본 연구에서는 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. 실험을 통하여 특징 축소가 학습의 결과와 성능에 영향을 긍정적으로 미치고 있음을 알 수 있다. 이러한 연구를 통하여 베이지안망이 변수들간의 확률관계를 축약된 형태로 표현

하는데 최적의 모델로서, 확률적 추론, 예측, 의사결정을 요하는 분야에 잘 적용될 수 있는 분류기임을 의료영역에서 확인하기 위하여 임상 데이터를 분석함으로써 이를 제시하고자 한다.

2. 관련연구

2.1 베이지안망

베이지안망(Bayesian network)은 특정 분야의 영역 지식(domain knowledge)을 확률적으로 표현하는 대표적인 수단으로서, 변수(특징)들간의 확률적 의존 관계(probabilistic dependency)를 나타내는 그래프와 각 변수별 조건부 확률들로 구성된다. 따라서 하나의 베이지안망은 각 노드마다 하나의 조건부 확률표(conditional probability table, CPT)를 갖는 하나의 비순환 유향 그래프(directed acyclic graph)로서 G=<N, A>, B=<N, A, θ>으로 정의할 수 있다[6]. 이 때 각 노드 N은 하나의 영역 변수를, 각 아크 A는 두 변수간의 확률적 의존성을 나타내며, θ는 조건부 확률들의 집합을 나타낸다. 일반적으로, 하나의 베이지안망은 다른 노드들에 배정된 값들을 기초로 특정 노드가 가질 값에 대한 조건부 확률을 계산하는데 이용할 수 있다. 따라서 하나의 베이지안망은 한 개체의 다른 특징들의 값이 주어졌을 때 분류 클래스 노드(classification node)의 사후 확률분포(posterior probability distribution)를 구해 줌으로써 개체들에 대한 하나의 자동 분류기(classifier)로 이용될 수 있다[10]. 즉 하나의 데이터 집합으로부터 베이지안망을 학습할 때 베이지안망의 각 노드는 데이터 집합의 각 특징을, 각 아크는 특징들간의 의존성을 표현하게 되며, 이렇게 학습된 베이지안망을 기초로 분류 클래스를 확률적으로 예측할 수 있다[5].

2.2 베이지안망 분류기의 유형

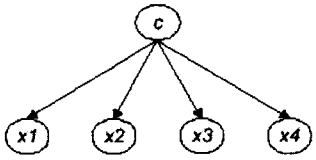
베이지안망 분류기는 한 개체 j가 클래스 Ci에 속할 확률을 계산함으로써 그 개체를 분류하는 방법이다. 그러한 확률은 [식1]와 같이 계산되며, 이 때 개체 j는 $Ai = Vi$ 형태로서 특징과 값의 쌍으로 표현된다.

$$P(C_i \mid A_1 = V_1, \ \& \ldots \& \ A_N = V_{N_j}) \qquad [식1]$$

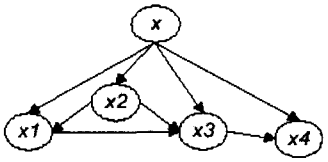
그러나 베이지안망에서 변수의 수가 많아지면 각 변수들간의 관계에 해당하는 값들도 많아져서 계산량을 급격히 요구하게 된다. 그래서 이를 현실적인 문제에 적용하기 위해서는 제약조건을 여러 가지로 두어 계산량을 줄이게 되는데 이런 여러 가지의 제약조건 형태가 베이지안망의 유형이 된다. 베이지안망의 유형은 학습성과를 결정하는데 매우 중요한 요소이다. 제약조건과 관련하여 대표적인 유형들을 살펴보면, 각 변수들이 다른 변수와의관계는 무시하고 분류 클래스와의 관계만을 갖는 제약조건을 인정한 나이브 베이지안망 (Naive Bayesian Network, NBN)이 있다. 그리고 트리구조로 확장한 나이브 베이지안망 (Tree Augmented Nave-Bayes, TAN)과 베이지안망으로 확장한 나이브 베이지안망 (Bayesian Network Augmented Nave-Bayes, BAN)이 있는데, 이는 나이브 베이지안망에서 제한된 조건인 변수들간의 의존성을 추가한 형태이다. 분류 클래스별로 망을 달리 가질 수 있는 베이지안 다중망 (Baysian Multi-Net), 그리고 분류 클래스를 별도로 두지 않는 형태인 일반 베이지안망 (general Bayesian Network, GBN)을 들 수 있다[5][6]. 이와 같은 대표적인 베이지안망의 유형중 본 연구에서는 [그림1]과 같은 3가지 유형에 대하여 실험하고 분석한다.

[그림1] 베이지안망의 유형

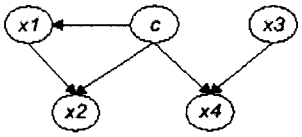
(a)나이브베이지안망



(b)베이지안망으로 확장한 나이브 베이지안망



(c)일반 베이지안망



2.3 베이지안망의 학습

베이지안망은 각 변수들과 그 변수들간의 관계를 표현하는 것이기 때문에 망을 결정하는데 있어서 가장 중요한 것은 망이 적용될 분야에 대한 영역지식이라 할 수 있다. 문제해결에 이용될 베이지안망을 학습하는 방법에는 크게 두 가지가 있다.

첫째는 사람이 직접 각 변수들간의 인과관계(causal relationship)를 이용하여 유형을 결정하는 방법이다. 베이지안망에서 각 노드는 변수에 해당한다. 사람은 우선 문제해결에 중요한 변수들을 결정한다. 그 뒤에 그 변수들 중 인과관계가 있다고 생각되는 노드들을 화살표로 연결하는 것이다.

두 번째 방법은 대량의 데이터를 이용하여 망을 결정하는 것이다. 이 경우에는 우선 필요한 변수들을 설정한다. 그리고 이 변수들간의 인과관계를 대량의 데이터를 이용해서 찾아내는 것이다. 여기에

는 각 유형들의 적합도(fitness)를 결정할 수 있는 기준(criterion)이 필요하며 이 기준을 이용해서 필요한 유형들을 찾아낼 수 있는 탐색기법(searching method)이 필요하다[12].

베이지안망을 학습하는 과정은 크게 베이지안망 그래프를 학습하는 과정과 그것을 바탕으로 각 변수의 조건부 확률들을 계산하는 과정으로 나누어 볼 수 있다[7]. 사람이 배경지식을 가지고 베이지안망 그래프를 직접 수작업으로 그려 주거나 편집해 주면 이를 바탕으로 훈련 데이터들을 분석하여 조건부 확률들을 자동으로 계산해주는 방식의 많은 베이지안망 학습 알고리즘과 프로그램들이 존재한다. 하지만 베이지안망 그래프로부터 각 변수의 조건부 확률을 계산하는 과정은 매우 단순한 과정인데 반해 훈련 데이터로부터 베이지안망 그래프를 학습하는 과정은 매우 복잡하고 어려운 과정이다. 따라서 기존의 많은 연구들은 바로 이러한 베이지안망 그래프 학습에 초점이 맞추어져 왔다. 베이지안망 그래프 학습을 위한 기존의 방법들은 크게 점수 기반 학습 알고리즘(scoring based learning algorithm), 조건부 독립성 기반 학습 알고리즘(conditional independency based learning algorithm)이 있다

점수 기반의 학습 알고리즘은 영역 지식을 바탕으로 임의의 초기 베이지안망을 만들고 일정한 평가 기준을 이용하여 가장 좋은 점수를 받는 양질의 베이지안망이 만들어 질 때까지 계속해서 이 베이지안망을 고쳐 가는 일종의 휴리스틱 탐색 방법이다. 이 경우 베이지안망의 구조가 데이터에 적합한 정도를 나타내는 점수를 산정한 후 데이터에 가장 적합한 베이지안망의 구조를 탐색하게 된다. 베이지안망의 적합도, 즉 점수를 산출하는데 이용되는 평가기준(scoring criteria)에 따라 엔트로피 기반의 방법(entropy-based method), 베이지안 점수 방법(Bayesian scoring method), MDL (minimum description length) 방법 등이 제안되었다. 점수가 선정되면 베이지안망 구조학습은 가능한 탐색공간에서 점수가 가장 좋은 망 구조를 찾는다.

조건부 독립성 기반 학습 알고리즘에서는 각 노드간의 의존성을 측정하기 위하여 조건부독립성을 나타내는 값(Conditional Independence, CI)을 사용한다. 노드간 조건부 독립성 테스트(CI test)를 시행하여 임계값 이상의 독립성을 갖는 노드들간에 아크를 삭제하거나 혹은 임계값 이하의 독립성을 갖는 노드들간에 아크를 추가해 가는 방법이다. 이는 두 변수간의 독립성 여부를 측정하기 위하여 정보 함수(Information Function)을 사용하는데 정보 함수중 가장 많이 사용하는 것이 상호 정보량(Mutual Information)이다. 기존의 많은 연구들을 통해 조건부 독립성 기반의 학습 알고리즘들이 점수 기반 학습 알고리즘들에 비해 비교적 학습시간은 오래 걸리지만, 보다 우수한 베이지안망을 학습할 수 있는 것으로 알려져 있다.

본 논문에서는 의료 데이터 마이닝을 위한 베이지안망 학습으로 조건부 독립성 기반 학습 알고리즘의 하나인 Jie Cheng의 CBL 알고리즘을 이용한다. CBL 알고리즘의 학습과정은 크게 3단계로 이루어진다. 초안 작성 단계 (drafting phase) 단계, 아크 추가 단계(thickening phase), 아크 삭제 단계 (thinning phase)로 나눌 수 있다

먼저 초안 작성 단계에서 베이지안망의 구조를 생성하게 되는데 이때 사용되는 측정치 (measurement)로는 [식2]로 정의되는 두 노드간의 단순 상호정보량(mutual information)이 사용된다. 이를 이용하여 임계값 이상을 갖는 노드들간에 아크들을 추가함으로써 개략적인 베이지안망 그래프의 초안을 작성한다[4].

$$I(X_i, X_j) = \sum_{x_i, x_j} P(x_i, x_j) \log \frac{P(x_i, x_j)}{P(x_i)P(x_j)} \quad [식2]$$

아크 추가 단계에서는 현재 초안 그래프에 포함되어 있는 아크들 외에 [식2]과 같이 정의되는 두 노드간의 조건부 상호정보량(conditional mutual information)이 임계값 이상인 경우들을 찾아 해당하는 새로운 아크들을 그래프에 추가한다.

$$I(X_i, X_j \mid C) = \sum_{x_i, x_j, c} P(x_i, x_j, c) \log \frac{P(x_i, x_j \mid c)}{P(x_i \mid c)P(x_j \mid c)} \quad [식3]$$

아크 삭제 단계에서는 이미 그래프에 포함되어 있는 기존 아크들에 대해 [식3]의 조건부 상호정보량이 충분한지 검사하여, 그렇지 못한 아크들을 모두 제거하여 최종 베이지안망 그래프를 완성한다. 이와 같은 CBL 알고리즘은 노드들의 완전한 순서(total ordering)가 주어진 경우에는 $O(N^2)$, 어떠한 순서도 주어지지 않는 경우에는 $O(N^4)$에 비례하는 시간만큼의 조건부 독립성 테스트를 요구하는 매우 효율적인 알고리즘이다.

2.4 베이지안망을 위한 특징 축소

베이지안망은 많은 변수들간의 확률관계를 가시적으로 표현하는 모델이긴 하지만 변수의 갯수가 많아지면 탐색공간이 넓어 학습하는데 시간이 많이 걸린다. 이러한 측면에서 탐색공간을 줄이기 위하여 덜 중요한 특징들을 판단하여 축소하는 방법을 사용한다. 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 분석 데이터를 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[9].

여과방법(filtering method)은 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. 이는 특징을 선택(selection)하는데 사용될 뿐 아니라 특징들을 융합(joining)하는데 유용하게 사용이 된다.

Pazzani의 연구에서는 나이브 베이지안망 분류기의 분류 성능을 높이기 위한 방법으로서, 특징 부분집합 선택뿐만 아니라 특징 융합(feature joining)도 적용해 보았다. 베이지안망의 구조가 주어진 경우 노드 x_i의 Markov blanket은 쉽게 구해진다[11].

$$P(X_i \mid BL(X_i)) = P(X_i \mid X - \{X_i\}) \qquad [식4]$$

즉, x_i의 부모노드, 자식노드, x_i를 제외한 자식노드의 부모노드가 $BL(x_i)$이다. 학습 데이터를 이용하면 조건부 상호 정보량 등을 이용해서 $BL(x_i)$를 구할 수 있다. 본 연구에서는 베이지안망 분류기의 성능을 높이기 위한 특징 부분집합 선택 방법으로 분류 클래스 노드의 Markov blanket에 속한 특징들을 선택하였다. 베이지안망에서 한 노드의 Markov blanket은 그 노드의 부모 노드들과 자식 노드들, 그리고 자식 노드들의 또 다른 부모 노드들을 포함하는 노드들의 부분 집합이다. 따라서 베이지안망에서 클래스 노드의 Markov blanket은 분류에 직접 영향을 주는 특징들만을 포함함으로써 자연스러운 또 하나의 특징 부분집합 선택 방법이 될 수 있다. 즉, Markov blanket 밖의 노드들로부터 어떠한 영향도 받지 않도록 차폐할 수 있다. 하지만 이러한 방법으로 선택된 특징들만으로 재구성된 베이지안망 분류기들이 실제로 어떤 분류 성능의 변화를 가져오는지를 밝힌 실험연구는 많지 않다. 그러나 나이브 베이지안망처럼 제약조건이 많은 경우 이를 개선하기 위하여 여과방법을 사용하기도 한다.

포장 방법(wrapper method)은 가능한 특징집합으로부터 실제로 특정 분류기를 생성하여 이 분류기의 분류 성능을 검사해 봄으로써 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. 이러한 포장방법은 상호정보량 테스트에 재사용할 수 있어 효율적이다. 베이지안망 학습과정에서 95%이상의 상호정보량 테스트가 사용이 된다.

Langley와 Sage의 연구에서는 양질의 특징 부분집합을 찾기 위해 전향선택(forward selection)방법을 적용하였고 이렇게 선택된 특징들만을 포함하는 나이브 베이지안망(NBN)을 생성하여 분류에 적용하였다. Kohavi와 John의 연구에서는 특징 부분집합 선택을 위해 최적 우선 탐색 방법(best-first search)을 적용하였다. 이 방법은 결정트리(decision tree)나 나이브 베이지안망 분류기 등 임의의 분류기를 포함하는 일종의 포장 방법이다. 포장방법에서는 전체 변수들의 모든 부분집합에 대해 탐색하여 예측율을 가장 좋게 하는 최적 부분집합을 찾아야 하기 때문에, 포장기법에서 전체 부분집합에 대해 탐색하는 경우 변수의 수가 늘어나게 되면 풀기 어려운 문제가 된다. 따라서 탐색공간을 줄이면서 효과적으로 탐색하기 위한 방법이 필요하다. 일반적으로 여과기법이 계산의 효율성이 좋으나 사용하려는 기계학습 기법과 독립적으로 주요 변수를 선택함으로써 실제로 해당 기계학습 기법을 적용시 성능은 포장기법이 더 낫다고 평가되고 있다

3. 실험

3.1 실험목표

본 연구에서 실험할 베이지안망 분류기 유형은 NBN, BAN, GBN으로 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 임신여부와 직접 연결 아크를 갖는 임신 요인들을 구하고 또 이들간의 상호의존성을 분석해 본다. 특징 축소 실험을 위해서는 GBN 그래프 상에서 분류 클래스 노드의 Markov blanket에 속한 특징들만을 골라 낸 뒤, 이 특징들만을 포함하도록 실험 데이터를 축소하고 이 축소된 데이터로부터 다시 NBN, BAN, GBN 등을 학습해 냈다. 이와 같이 축소된 특징집합으로부터 얻어진 NBN, BAN, GBN을 각각 NBNSF (NBN with Selected Features), BANSF (BAN with Selected Features), GBNSF (GBN with Selected Features)로 부른다. 특징 축소의 효과를 알아보기 위한 방법으로 특징 축소 이전의 NBN, BAN, GBN과 특징 축소 이후의 NBNSF, BANSF, GBNSF의 분류성능을 서로 비교해 본다.

주어진 데이터집합으로부터 베이지안망의 대표적인 유형인 NBN, BAN, GBN 등 제약조건이 다른 다양한 유형의 베이지안망 분류기들을 생성하고 이들이 이 의료분야 데이터에서 보여주는 분류성능을 서로 비교해본다. 각각 제약조건을 달리하고 있고 가장 실재와 유사한 GBN과 가장 제약조건이 많은 NBN과의 차이를 실험결과로서 비교한다. 일반적으로 제약조건이 많은 NBN에서도 성능이 우수함을 많은 연구결과에 의해 밝혀졌지만 의료분야에서도 NBN의 성능을 실험해 본다.

베이지안망 분류기의 학습 효율과 분류성능을 높이기 위한 특징 축소의 한 방법으로서, 분류 클래스노드의 Markov blanket으로 특징을 축소하는 것이 얼마만큼 효과가 있는지 베이지안망 분류기별로 분류성능을 비교해 본다. 특징 축소의 방법으로는 분류 클래스를 임상적 임신여부를 정하고 이를 Markov blanket방법으로 종속성이 적은 특징을 제거함으로써 축소하게 된다.

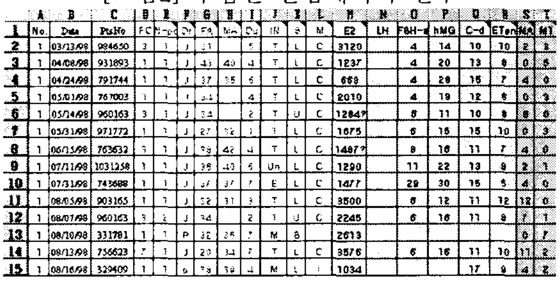
본 논문에서는 하나의 베이지안망에서 클래스 노드의 Markov blanket에 속한 특징들로 특징집합을 축소하는 것이 베이지안망 분류기들의 성능을 높여 줄 수 있는지를 실험을 통해 알아본다. 먼저 269 개의 전체 실험 데이터집합을 244개의 훈련 데이터와 25개의 테스트 데이터 집합으로 구성한다. 그리고 훈련 데이터 집합에 대해 각 베이지안망 분

류기들과 여타 분류기들을 학습한 뒤, 이 분류기들을 각각 훈련 데이터 집합과 테스트 데이터 집합으로 분류성능 테스트를 시행해 본다. 끝으로 전체 실험 데이터 집합을 가지고 10회 교차검정방법(10-fold cross validation)으로 각 분류기의 평균 분류성능을 알아본다. 베이지안망 학습을 위해서는 조건부 독립성 기반의 알고리즘인 Jie Cheng의 CBL 알고리즘과 그것을 구현한 BN Power Constructor 1.0 프로그램을 이용한다[5]. 다른 분류기와의 비교를 위해서는 Java로 구현된 Weka 패키지를 이용한다. 모든 실험은 Intel Pentium III 700MHz, 256MB Memory 사양의 컴퓨터에서 진행하였다.

3.2 실험데이터의 수집

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 불임 시술의 방법을 선택하게 된다. 임상의 특성상 기간이 가장 짧은 경우에도 1년이 걸리고 보통의 경우 실패를 포함하여 수년이 걸리는 상황을 고려할 때 많은 실험 데이터의 확보가 상당히 어려웠다. 더구나 동시에 많은 환자를 시술할 수 없는 임상적 특징으로 인하여, 많은 환자의 정보를 한꺼번에 얻기가 어려웠다. 이런 환자의 정보들은 차트에 의해 자세히 기록되어 있으나, 우리는 실험을 위하여 관련된 정보를 요약하여 정리하게 되었는데, 그 결과로 [그림2]과 같은 Excell파일은 얻게 되었다.

[그림2] 수집된 실험데이터 일부



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	No.	Date	PtsNo	FC	N-pd	Dr	FA	MA	Du	IN	?	M	E2	LH	FSH-d	hMG	C-d	ETen	MA	MI
2	1	03/13/98	984650	3	1	J	31		5	T	L	C	3120		4	14	10	10	2	3
3	1	04/08/98	931893	1	1	J	43	40	4	T	L	C	1237		4	20	13	8	0	5
4	1	04/24/98	791744	1	1	J	37	35	6	T	L	C	668		4	28	15	7	4	0
5	1	05/01/98	767003	1	1	I	[illegible]		4	T	L	C	2010		4	19	12	6	0	3
6	1	05/14/98	960163	3	1	J	34		2	T	U	C	1284?		8	11	10	8	8	0
7	1	05/31/98	971772	1	1	J	27	32	1	T	L	C	1675		6	15	15	10	0	3
8	1	06/15/98	763632	3	1	J	39	42	4	T	L	C	1487?		8	16	11	7	4	0
9	1	07/11/98	1031258	1	1	J	36	40	6	Un	L	C	1290		11	22	13	8	2	1
10	1	07/31/98	743688	1	1	J	37	37	7	E	L	C	1477		29	30	15	5	4	0
11	1	08/05/98	903165	1	1	J	32	31	3	T	L	C	3500		8	12	11	12	12	0
12	1	08/07/98	960163	3	2	J	34		2	T	U	C	2245		6	16	11	8	7	1
13	1	08/10/98	331781	1	1	P	32	35	7	M	8		2613						0	7
14	1	08/13/98	756623	7	1	J	29	34	7	T	L	C	3576		6	16	11	10	11	2
15	1	08/16/98	329409	1	1	6	39	39	4	M	L	[illegible]	1034				17	9	4	2

이와 같이 수집된 정보중에는 환자 개인의 신상정보까지 포함하고 있었으며, 많은 경우 진료 및 시술을 위한 의사들의 불규칙적인 형태의 정보 표현이 많았다. 이를 해당 의사들의 자문을 얻어 규칙적인 코드값으로 표현하도록 노력하였으며 해당 분야의 전문가 조언을 통하여 400여 명의 환자에 대하여 총 40여개의 검사항목이 기록되어 있는 실험 데이터를 확보하게 되었다. 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[7]. 본 연구에서는 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 특징들은 결과적으로 <표1>과 같이 9개가 선택되었다[1][3].

<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

3.3 전처리

3.3.1 정제작업

수집된 원시 데이터집합에서 분석대상인 불임과 가임에 영향을 미치는 중요한 항목을 영역 지식을 갖고 있는 전문가를 통하여 주요 특징을 선정하게 되었다. 또한 검사항목이 일부환자에게만 적용이 되어서 동일한 조건에서 비교할 수 없는 항목들과 임신과 관련성이 거의 없다고 전문가가 판단하는 항목 등을 제거하여, 실제 데이터 분석에 사용할 항목을 정했다. 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.

<표2> 정제된 데이터 항목

Attribute	설명	값	설명	값
증상 (IND)	Endometriosis	A	P and T	G
	Immunological	B	Tubal	H
	Ovarian	C	T and U	I
	O and T	D	Uterine	J
	O and U	E	Unexplained	U
	Peritoneal	F		
약물치료법 (Stim)	Long Protocol	L	Parlodel	P
	ShortProtocol	S	Follicular	RF
	Ultra Short "	U	Null	N
	HMG only	H	Clomiphene	C
	FSH	F	FSH-HMG	FSH-H
시술방법 (IVF)	IVF-ET	C	ICSI	I
Wallace 사용(ETM)	Wallace	W	Default	T

누락 항목이 많은 미세 조작술, 보조도구 사용 유무, 총이식 수정란수 등은 정상인 경우 기입하지 않은 경우들이 대부분을 차지하고 있어, 정상값 또는 평균값으로 적용하였다. 이식일수, 총이식 수정

란수의 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 연속 수치 데이터 값들을 많이 포함하고 있어 적절한 이산화작업이 필요하였다[2].

3.3.2 이산화작업

대표적인 이산화 방법에는 Equal Width Interval Binning 방법, Holte's 1R(One-Rule) Discretizer 방법 그리고 Recursive Minimal Entropy Partitioning 방법이 있다[8]. Equal Width Interval Binning 방법은 하나의 특징이 갖는 값의 범위에 대해 단순하게 데이터의 최대값과 최소값을 이용해서 그 사이의 값들을 동일한 크기의 구간으로 구분하여 대표값을 할당하는 방법이다. 이 방법에서는 [식5]과 같이 측정된 최대값과 최소값의 차이를 사용자에 주어지는 변수 k값으로 균등하게 나누는 방법이다.

$$\delta = \frac{x_{\max} - x_{\min}}{k} \qquad [식5]$$

Holte's 1R Discretizer 방법은 값의 넓이와 깊이에 구애받지 않고 영역 지식에 의하여 구간을 정하고 각 구간에 속한 값을 대표값을 정하는 방법이다. 이는 1차 레벨만을 갖는 결정트리와 같은 방법을 사용하는 간단한 분류방법으로 decision stumps라고도 불리워 지기도 한다. 영역 지식의 전문가를 통하여 이산화 대상이 되는 특징의 연속된 값들을 결정트리 방법에 의해 이산화하는 방법이다. 그리고 마지막으로 각 특징값들을 임의의 범위로 구분한 후 엔트로피 계산을 통해 가장 작은 수치를 나타내는 구간을 선택하여 사용하는 Recursive Minimal Entropy Partitioning 방법을 들 수 있다. 주어진 인스턴스의 집합을 S라 하고 특징의 집합을 A, 그리고 이산화할 구간을 T라 하면 엔트로피는 [식6]에서와 같이 계산이 된다.

$$E(A, T; S) = \frac{|S_1|}{|S|} Ent(S_1) + \frac{|S_2|}{|S|} Ent(S_2) \qquad [식6]$$

Gain값이 최소가 될 때까지 반복적으로 구간을 나누게 되는데 [식7]과 같은 조건이 만족하면 반복을 멈추게 되고 구간이 나누어지게 된다.

$$Gain(A, T; S) < \frac{\log_2(N-1)}{N} + \frac{\Delta(A, T; S)}{N} \qquad [식7]$$

본 연구에서는 여성의 나이, 미세 조작 난자수, 총 이식 수정란수의 3가지 항목에 대해서 이산화 작업을 하였으며, 적용된 이산화 방법은 Holte's 1R Discretizer 방법이다. 이는 각 항목들이 나타내는 값들의 분포가 일정치가 않고 변화가 심한 특성으로 인해 영역 지식을 가장 잘 반영한 방법을 택한 것이다. 예를 들면 많은 수정란을 얻는 것과 좋은 수정란으로 키우는 것이 매우 중요하므로 여성의 나이가 35세 이상이면 자연임신이 급격히 감소하게 된다. 이러한 경우 나이의 값이 이산화되는 구간은 35세로 판단하는 것이 바람직하다고 볼 수 있다.

<표3> 이산화한 데이터 항목

Attribute	설명	값	설명	값
여성의나이 (FA)	20 ~ 34 35 ~ 40	L M	over 40	H
미세조작 난자수(ICT)	Null 1 ~ 5 6 ~ 10	N L M	11 ~ 15 16 ~ 20 over 20	MH H HH
총이식 수정란수(TO)	Null 1 ~ 5	N Normal	6~10 over 10	SHIGH HIGH

4. 실험결과

훈련 데이터 집합으로부터 학습하게 되면 기본 유형의 베이지안망이 얻어 지게 되는데, 이는 각 베이지안망 특성에 따른 각 특징들의 종속성을 잘 표현해 주고 있다. 실험의 대상은 임신여부(clin), 증상(IND), 약물치료법(Stimulation), 여성의 나이(FA), 미세조작 난자수(ICT), Wallace사용여부(ETM), 총 이식 수정란수(TO), 이식일수(ETD), 시술방법(IVF) 등과 같이 9개의 특징들을 변수로 사용하여 실험을 하였다. 각 유형별로 제약조건이 달라 베이지안망의 형태가 달리 학습되었다. 각각에서 생성된 베이지안망은 영역지식을 그래프로 표현하여 특징들간의 종속성 특히 인과관계로 잘 나타내고 있으며 인과관계뿐만 아니라 각각의 아크(arc)에 대해서 조건부 확률표가 생성이 되어 특정조건 여기서는 주로 분류 클래스인 임신여부에 대하여 확률적 분포로서 종속성을 표현한다.

4.1 특징축소 성능개선

본 연구의 관심대상인 3가지 베이지안망 분류기 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 <표4>에서 보여주고 있다. <표4>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 비교해 보면, NBN, BAN, GBN의 순으로 분류 성능이 증

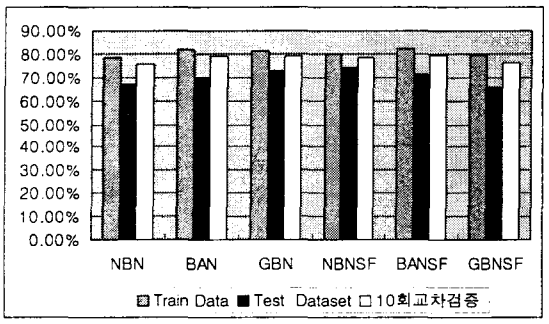
가하였다는 것을 알 수 있다. 특히 특징들간의 자유로운 의존관계를 허용하는 BAN과 GBN이 그렇지 못한 NBN에 비해서 상당히 우수한 성능을 나타냄을 알 수 있다. 하지만 특징들간의 직접적인 의존관계를 무시하고 모두 서로 독립성을 가정하는 NBN도 다른 일반 분류기에 비해 상당히 높은 성능을 나타낸 것은 주목할 만 하다.

<표4> 분류기별 분류성능 비교

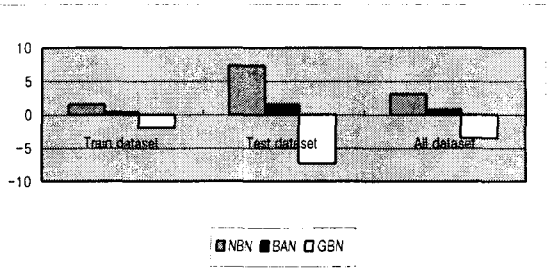
Classifiers	By Train Dataset	By Test Dataset	10-Fold Cross Validation
NBN	78.4%	67.1%	75.5%
BAN	81.9%	70.0%	78.8%
GBN	81.4%	72.9%	79.2%
NBNSF	79.9%	74.3%	78.4%
BANSF	82.4%	71.4%	79.6%
GBNSF	79.4%	65.7%	75.9%

베이지안망 유형별로 분류 성능을 살펴보면 [그림3]에서 보는 바와 같이 BAN이 가장 높게 나타났으며 NBN이 가장 낮게 나타났다. 가장 제약조건이 많은 NBN이 분류성능이 낮은 것은 당연한 것으로 판단할 수 있으나 당초의 예상보다는 대체로 높게 나타났다. 데이터 집합간의 분석결과는 10회 검증방법으로 실험한 결과가 가장 높게 나타난 것은 반복적인 학습의 효과를 실험의 결과로 얻은 결과로 볼 수 있다.

[그림3] 베이지안망 유형별 분류성능비교



이제까지 우리는 특징 축소 이전의 NBN, BAN, GBN의 분류 성능과 더불어 Markov blanket내의 5개 특징들로 특징집합을 축소한 후 얻어진 NBNSF, BANSF, GBNSF 등의 분류성능을 살펴보았다. 특징 축소 이전의 NBN과 BAN에 비해 각각 특징 축소 이후의 NBNSF와 BANSF의 분류성능이 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.



[그림4] 특징축소 성능개선 효과 (단위:%)

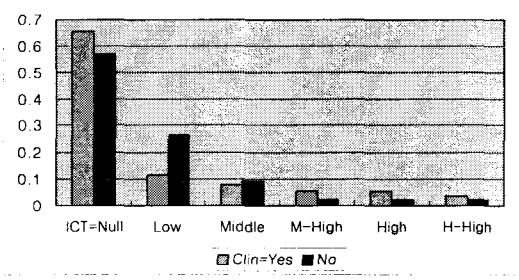
4.2 특징축소 의존성 변화

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 나이브베이지안망에서는 특징의 축소가 학습의 결과에 영향을 주지 않은 듯 보이지만, 성능에서는 많은 개선을 보였다. <표5>와 [그림5]은 NBN과 NBNSF에서 동일한 학습결과를 생성한 하나의 예를 보여주고 있다. 특징을 축소한 학습의 결과가 기본유형과 동일한 결과를 갖은 것은 학습결과의 성능이 좋았음을 알려주고 있다.

<표5> 미세조작 난자수가 임신에 미치는 영향

ICT / Clin	Null	Low	Middle	M-High	High	H-High
Yes	0.6566665	0.1166667	0.0766667	0.0566667	0.0566667	0.036667
No	0.5712789	0.2631027	0.0932914	0.024109	0.024109	0.024109

[그림5] 미세조작 난자수와 임신여부의 종속성



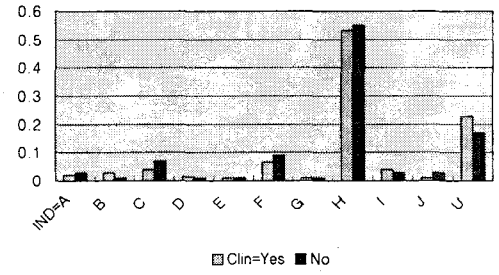
BAN에서의 실험의 결과를 살펴보면 <표6>과 [그림6]에서 나타내 주고 있듯이 증상과 임신의 종속성을 나타내는 학습의 결과에서 증상이 H(Tubal, 난관이상)인 경우 가장 의존성이 높았다. 이러한 난관이상도 체외수정을 통하여 보조도구를 이용한 이식방법을 사용하면 임신에 이르게 되어, 불임환자

중 가장 의존성이 높게 나타난 것으로 보인다. 특징을 축소한 BANSF의 경우에도 비슷한 결과를 얻을 수 있었다.

<표6> BAN에서 증상과 임신여부의 종속성

IND / Clin	A	B	C	D	E
Yes	0.0217267	0.028016	0.0405946	0.0154374	0.0091481
No	0.0290909	0.0090909	0.0690909	0.0090909	0.0090909
F	G	H	I	J	U
0.0657519	0.0091481	0.5311607	0.0405946	0.0091481	0.2292738
0.0890909	0.0090909	0.5490909	0.0290909	0.0290909	0.169091

[그림6]BAN에서 증상과 임신여부의 종속성



5. 결론

본 연구에서는 산부인과의 래원 환자들에 대한 실제의 임상 데이터로부터 불임과 관련된 특징들간의 의존성을 표현하고 분석하는데 베이지안망을 적용해 보았다. 베이지안망은 특정 분야의 영역지식을 표현하는 대표적인 수단으로서, 특징들간의 확률적 의존 관계를 나타내는 그래프와 각 변수별 조건부 확률들로 표현한다. 하나의 데이터 집합으로부터 베이지안망을 학습할 때 베이지안망의 각 노드는 데이터 집합의 각 특징을, 각 아크는 특징들간의 의존성을 표현하게 되며, 이렇게 학습된 베이지안망을 기초로 분류 클래스를 확률적으로 예측할 수 있다. 이에 근거하여 의료의 의사결정을 최적의 방법으로 선택하게 되고, 환자들의 검사결과와 각종 진단자료에 근거하여 의사들의 정확한 진단을 도와 줄 수 있도록 한다. 본 연구에서는 실험을 통해 베이지안망에 드러난 임신여부에 영향을 주는 특징들간의 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, 약물치료법, 여성의 나이, 미세 조작 난자의 수, Wallace 사용여부 등 5개의 특징들을 가려낼 수 있었고, 이 특징들간의 상호 의존성도 찾아 낼 수 있었다. 또 본 연구에서는 실험을 통해 서로 다른 유형의 베이지안망 분류기들 중에서 특징들간의 상관관계를 더 자유롭게 표현할 수 있는 BAN과 GBN들이 그렇지 못한 NBN에 비해 상대적으로 더 높은 분류 성능을 보여준다는 것을 확인하였다. 또한 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 그 효과를 입증할 수 있었다. GBN의 경우는 예상외로 성능이 저하되는 결과를 보여 주었다.

본 연구를 통하여 의료영역에서의 지식표현을 실제의 임상데이터를 활용하여 베이지안망을 학습시켜 얻어냈고, 학습결과에 나타난 의학적 해석을 시도하였다. 기존의 통계적 해석에 불과한 의료영역의 지식표현을 인공지능의 기법을 이용하여 학습하는 과정으로 얻어냈고, 새로운 환경과 데이터를 통한 지속적인 영역지식의 학습 결과가 이루어 질 것이다. 학습된 베이지안망이 의료 영역지식을 잘 표현하고 있음을 확인할 수 있었고, 특히 특징을 축소하여 만든 유형들이 성능에 있어서 NBN, BAN의 경우 개선됨을 실험을 통하여 확인할 수 있었다.

참고문헌

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

[2] 정용규, 진훈, 김인철, 베이지안망을 이용한 불임요인 분석 및 가임예측, 정보과학회 추계학술대회, 2001.

[3] 정혁, '불임, 무엇이 문제인가 -그 원인과 치료', 우리출판사, 1997

[4] Cheng, J., Bell, D.A. and Liu, W., An Algorithm for Bayesian Belief network construction from data, Proceedings of AI & STAT'97 pp83-90, 1997.

[5] Cheng, J., Greiner, R.., Comparing Bayesian network classifiers, Proceedings of UAI-99, 1999.

[6] Cheng, J. and Greiner, R., Learning Bayesian Belief network classifiers: Algorithms and system, Proceedings of the fourteenth Canadian conference on artificial intelligence, 2001.

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

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

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

[10] Kevin Patrick Merphy, A Brief Introduction to graphical Models and Baysian networks, Berkley

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

[12] Tom M. Mitchel, Machine Learning, McGrow-Hill, 1997