베이지안 네트워크에 기초한 백혈병 유전자데이터의 분석

Bayesian Network-Based Analysis on Gene Data of Leukemia Patients

지도교수 : 장병탁

이 논문을 공학학사 학위 논문으로 제출함.

2005년 12월 26일

서울대학교 공과대학
컴 퓨 터 공 학 부
김 경 헌

2005년 12월

베이지안 네트워크에 기초한 백혈병 유전자데이터의 분석

(Bayesian Network-Based Analysis on Gene Data of Leukemia Patients)

김 경 헌
서울대학교 컴퓨터공학과

요약

최근 유전자데이터를 분석할 수 있는 기술이 개발되면서 유전자간의 관계와 유전자의 고유 기능에 대한 관심이 증폭되고 있다. 하지만 수만 개에 이르는 유전자를 아직까지는 완벽하게 분석하지 못했고 실험-대조군이 발생한다고 하더라도 바뀐 환경에 의해 유전자가 변경이 되었는지 확인하는 것은 쉽지 않다. GeneChip을 통하여 얻어진 Microarray 형태의 유전자데이터를 백혈병이 걸린 상태와 치료 후의 상태의 데이터를 얻어내어 이것을 분석하도록 한다. 기계학습의 한 종류로 베이지안 네트워크는 확률에 근거하여 각 노드간의 관계를 규정할 수 있는 것으로서 유전자데이터를 분석하는 데에는 가장 적합한 방법이라고 할 수 있다. 이렇게 유전자데이터와 베이지안 네트워크를 사용하여 베이지안 네트워크 분류기의 성능을 종류별로 알아보고 다른 기계학습 방식과 비교해보도록 한다.

I. 서론

1.1 연구의 목적

인류나 다른 생물학적 개체들이 오랫동안 진화라는 것을 할 수 있었던 것은 체내에 정보를 저장하는 무엇인가가 있기 때문에 가능하였다. 이것을 유전자(Gene)라고 부른다. 범위를 좀 더 좁혀서 인간의 질병을 치료하기 위해서 유전자라는 것에 대해서 오랫동안 연구를 하였고 그 결과들이 조금씩 보이고 있다. 실제로 유전자 분석은 후천적인 질병이 아닌 태어나면서부터 결정이 되는 선천적인 질병에 대해서는 근본적인 치료의 방법이 될 수도 있다고도 논의되고 있다. 하지만 인간의 유전자를 분석하기에는 그 양이 너무나 많고 그 유전자 하나하나가 어떤 역할을 하는지 그리고 그 유전자 간의 관계가 어떻게 되는지 알아내는 과정은 정량적으로도 쉽지 않을 것으로 예상이 된다. 유전자 간의 관계를 알아낼 수 있는 통계적인 정보를 가지고 사전 작업을 통하여 유전자들을 선별하고 베이지안 네트워크를 사용하여 유전자 간의 관계를 알아내어 인간의 유전정보를 분석하고 나아가 유전자를 통하여 질병을 예측할 수 있는 기반을 만들어 보도록 한다.

1.2 연구의 내용

백혈병의 원인이 무엇인지는 아직도 완벽히 밝혀지진 않았다. 그러나 방사선, 유전적인 요인, 화학물질, 그리고 바이러스가 백혈병 발생과 관련이 있다고 보고 정도는 되고 있다. 동물 실험이나 연구로는 바이러스 감염이 가장 가능성이 많은 것으로 나타나고 있는데, 이는 특정한 염색체의 전좌를 일으킬 수 있고, 세포성 면역의 감시로부터 회피할 수 있기 때문이라고 생각된다. 예를 들면, HTLV-1과 같은 바이러스는 성인에서 백혈병을 일으키는 데 직접적인 역할을 하는데, 아직까지 소아 백혈병의 경우는 그 원인이 확실하게 알려지고 있지는 않다. 환경적 요인으로는 방사선 조사, 감염, 벤졸, 클로람페니콜 같은 약물 등이 의심되고 있으나 이에 대한 확증은 아직 없다.

유전적인 원인을 드는 학설은, 백혈병을 앓고 있는 환자의 형제자매 중에서 그렇지 않은 경우 보다 4배의 높은 발생률을 보이며 가족적으로 백혈병의 집결 빈도가 높은 것, 또 일란성 쌍생아 중의 한 명이 백혈병을 앓을 때 남은 쌍둥이가 같은 병을 앓을 가능성은 다섯 중 하나인 것, 선천적으로 염색체가 붕괴되기 쉬운 질환(Bloom 증후군, Fanconi 재생 불량성 빈혈, Ataxia telangiectasia), 후천적으로 염색체 붕괴를 일으키는 상태(방사선 조사, 벤젠),과다 염색체(다운 증후군), 선천성 면역 결핍증이 있는 사람들에서 보통 사람보다 백혈병의 발생빈도가 높다는 사실을 들어 유전적인 원인을 드는 학설도 있다. 이번 논문에서 다른 것 보다는 유전적인 이유에 대한 자료를 바탕으로 백혈병의 원인에 관련된 유전자와 관련이 많을 수 있다고 판단되는 것들에 대해서 조사를 하기로 한다.

1.3 연구의 범위

인간의 유전자의 개수는 2만에서 6만개 사이라고 알려져 있다. 이것으로 나올 수 있는 인간의 형질은 대략 10의 9,000제곱 정도 된다고 가정을 한다면 이러한 데이터는 현존 하는 슈퍼컴퓨터라도 만만치 않은 계산양이라는 것을 알 수 있다. 특히나 이러한 수치는 하나의 유전자가 0또는 1을 갖는다는 가정 하에 추정된 것이고 실제로는 더 많은 범위의 성질을 가지고 있다고 예상된다. 여기서 이런 모든 데이터를 사용하여 백혈병과 유전데이터의 관계를 밝히기는 불가능하다. 1차적으로 백혈병과 관련이 없을 것 같은 유전자를 분류하고 최종적으로는 유전자의 양을 베이지안 네트워크에 적용할 수 있을 정도의 양으로 줄인 후에 각 유전자데이터와 백혈병이 발병할 수 있는 가능성에 대해서 관계를 찾아내어 백혈병의 유전적인 요소와 그 데이터를 바탕으로 베이지안 네트워크의 본연의 특성에 대해서 앞으로 알아보기로 한다.

II. 베이지안 네트워크를 이용한 유전자데이터의 분석

2.1 베이지안 네트워크의 기본개념

베이지안 네트워크(Bayesian Network)는 특정 분야의 영역 지식(Domain Knowledge)을 확률적으로 표현하는 대표적인 수단으로, 변수들 간의 확률적 의존관계(Probabilistic Dependency)[5][7]를 나타내는 그래프와 각 변수별 조건부 확률로 구성된다<그림1>. 따라서 하나의 베이지안 네트워크는 각 노드마다 하나의 조건부 확률표(Conditional Probability Table)를 갖는 하나의 비순환 유향 그래프(Directed Acyclic Graph:DAG)로 정의할 수 있다. 이때 각각의 노드는 이벤트의 발생을 의미하고 노드를 이어주는 DAG는 이벤트 간의 관계(Relationship)를 의미한다.

$$A \rightarrow B : A \; causes \; B \tag{1}$$

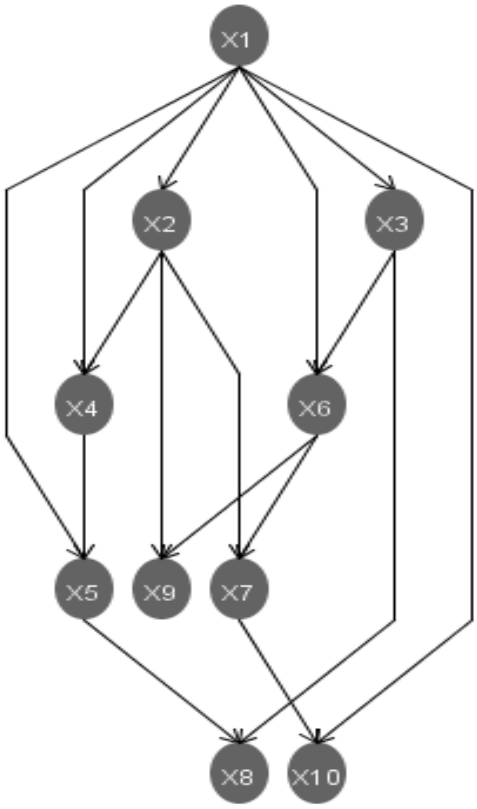
일반적으로, 하나의 베이지안 네트워크는 다른 노드들에 배정된 값들을 기초로 특정 노드가 가질 값에 대한 조건부 확률을 계산하는데 이용할 수 있다. 따라서 하나의 베이지안 네트워크는 한 개체의 다른 속성들의 값이 주어졌을 때 분류 클래스 노드(Classification Node)의 사후 확률 분포(Posterior Probability Distribution)를 구해줌으로써 개체들에 대한 하나의 자동 분류기(Classifier)로 이용될 수 있다[9,11]. 즉 하나의 데이터 집합으로부터 베이지안 네트워크를 학습할 때 베이지안 네트워크의 각 노드는 데이터 집합의 각 속성을, 각 아크는 속성들 간의 의존성을 표현하게 되며, 이렇게 학습된 베이지안 네트워크를 기초로 분류 클래스를 확률적으로 예측 할 수 있다. 예를 들어 변수 X1~X10 이 있고 각 변수들이 TRUE(1) 와 FALSE(0) 의 값을 갖는다고 할 때 베이지안 네트워크는 <그림1>과 같이 각 변수들에 대한 의존성을 그래프로 표현할 뿐 아니라 각 변수별로 <표1>과 같은 조건부 확률도 함께 표현할 수 있다. <표1>은 10개의 변수 중, 변수 X7이 가지고 있는 X2과 X6에 대한 조건부 확률표(Conditional Probability Table)를 나타내고 있다. <그림1>에서 보듯이 변수 X7은 변수 X2와 X6에 대해서만 종속성을 가지며 따라서 조건부 확률표의 각 행(Row)은 이 두 변수 X2와 X6에 배정 가능한 값들을 나타내며 각 열(Column)은 변수 X7이 가질 수 있는 값들을 나태난다. 결국 표상의 각 셀은 두 변수 X2와 X7의 값들에 의해 X7이 가질 수 있는 값들의 확률을 표현하고 있는 것이다.

$$\begin{aligned} P(X7 \mid X2, X6) &= 0.198 \\ P(\sim X7 \mid X2, X6) &= 0.802 \end{aligned} \tag{2}$$

와 같은 의미를 지닌다.

<표1> CPT

Probability Distribution Table For X7			
X2	X6	0	1
0	0	0.102	0.898
0	1	0.335	0.665
1	0	0.568	0.432
1	1	0.802	0.198



<그림1> 일반 베이지안 네트워크

베이지안 네트워크는 조건부 확률 계산에 식(1)과 같은 베이즈 정리(Bayesian Theorem)를 이용한다. 베이즈 정리는 관측된 데이터 D로부터 가설 h 가 옳을 확률 P(h|D)과 가설 h의 사전 확률(prior probability) P(h)을 기초로 계산할 수 있는 방법을 제시한다[8][14].

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} \quad (3)$$

2.2 베이지안 네트워크 분류기의 유형

대표적인 3가지 유형의 베이지안 네트워크 분류기에는 <그림2>,<그림3>,<그림1>에서 보는 바와 같이 순서대로 나이브 베이지안 네트워크(NBN;Naïve Bayesian Network), 트리-확장 베이지안 네트워크(TAN;Tree-Augmented Network), 일반 베이지안 네트워크(GBN;General Bayesian Network)[2]등이 있다. NBN은 <그림2>와 같이 제일 위의 클래스 노드를 제외한 다른 모든 속성노드들이 클래스 노드에만 의존적이고, 그들 간에 는 서로 독립적이라는 가정을 바탕으로 만들어진 베이지안 네트워크이다. 수식으로 설명을 하자면

조건부 확률의 성질에 의해서 다음이 성립하고,

$$P(A|B) = \frac{P(A,B)}{P(B)} \quad (4)$$

Product Rule에 의해서

$$P(A,B) = P(A|B)P(B) = P(B|A)P(A) \quad (5)$$

가 성립한다.

결과적으로 체인룰(Chain Rule)을 사용하면,

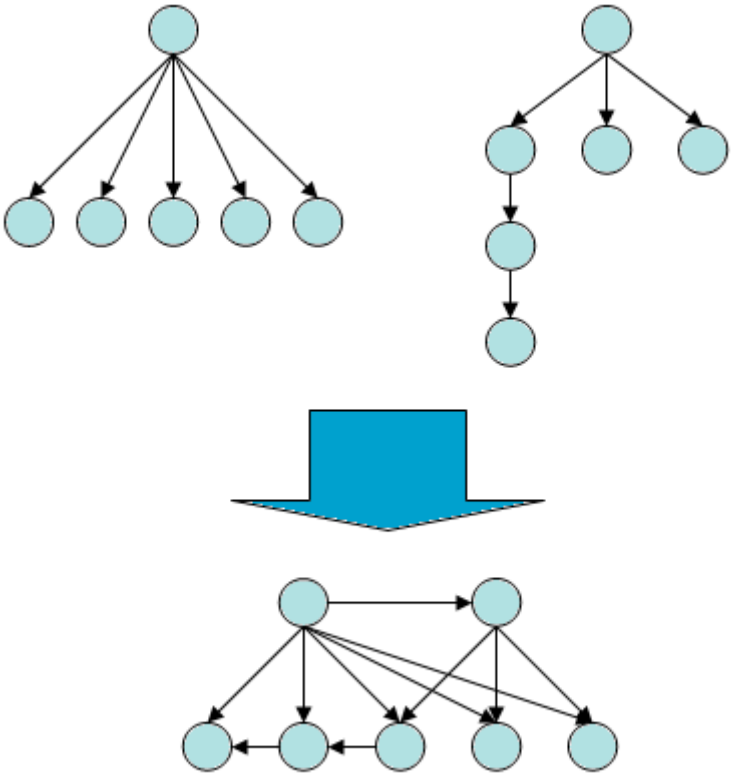
$$P(X_1, X_2, \ldots, X_n) = P(X_1)\prod_{i=2}^{n} P(X_i|X_1, \ldots, X_{i-1}) \quad (6)$$

의 식이 만들어진다.

이러한 NBN은 가정의 단순함에도 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 속성들 간의 독립성 가정을 완화함으로써 NBN의 분류 성능을 높여 보려는 많은 시도들이 이루어지고 있다.

TAN은 이와 같은 시도의 하나로서, NBN과는 달리 속성 노드들 간에도 상화의존성이 존재한다고 가정하고 이러한 속성 간 상호의존성을 하나의 일반 베이지안 네트워크 형태로 표현 가능하도록 NBN을 확장한 것이다. 즉 TAN은 <그림3>과 같이

NBN을 기반으로 하여 트리구조의 네트워크를 결합하는 형태이다. 베이지안 네트워크 분류기중 가장 일반화된 형태는 <그림1>과 같은 GBN으로서, GBN에서는 기존의 다른 베이지안 네트워크 분류기들과는 달리 클래스 노드 조차일반 속성노드와 차이를 두지 않고 모든 노드들 간의 상호의존성을 하나의 베이지안 네트워크로 표현한 것이다[1]. 따라서 GBN에서는 클래스 노드도 부모 노드들을 가질 수 있다. GBN은 속성 간 상호의존성을 표현할 수 있는 가장 자연스러운 방법이지만 어떠한 제약도 갖지 않은 상태에서 이러한 베이지안 네트워크를 학습하는 데는 매우 높은 학습비용이 소요된다[16].



<그림3> TAN의 예[16]

2.3 베이지안 네트워크의 학습

베이지안 네트워크를 학습하는 과정은 크게 베이지안 네트워크 그래프를 학습하는 과정과 그것을 바탕으로 각 변수의 조건부 확률들을 계산하는 과정으로 나누어 볼 수 있다[4][5]. 사람이 배경지식을 가지고 베이지안 네트워크 그래프를 직접 수작업으로 그려주거나 편집해주면 이를 바탕으로 훈련데이터들을 분석하여 조건부 확률들을 자동으로 계산해주는 방식의 많은 베이지안 네트워크 학습 알고리즘과 프로그램들이 존재한다. 하지만 베이지안 네트워크 그래프로부터 각 변수의 조건부 확률을 계산하는 과정은 매우 단순한 과정인데 반해 훈련 데이터로부터 베이지안 네

트워크 그래프를 학습하는 과정은 매우 복잡하고 어려운 과정이다. 따라서 기존의 많은 연구들은 바로 이러한 베이지안 네트워크 그래프 학습에 초점이 맞추어져 왔다[2][3][5][12][13]. 베이지안 네트워크 그래프 학습을 위한 기존의 방법들은 크게 점수 기반학습 알고리즘들과 조건부 독립성 기반 학습알고리즘들로 나눌 수 있다. 점수 기반의 학습 알고리즘은 영역지식을 바탕으로 임의의 초기 베이지안 네트워크를 만들고 일정한 평기기준을 이용하여 가장 좋은 점수를 받는 양질의 베이지안 네트워크가 만들어질 때 까지 계속해서 이 베이지안 네트워크로 고쳐가는 일종의 Heuristic Search의 한 방법이다. 점수를 산출하는데 이용되는 평가기준에 따라서 엔트로피 기반의 방법, 베이지안 점수 방법, MDL 방법 등이 제안되었다. 한편 조건부 독립성 기반 학습 알고리즘에서는 노드 간 조건부 독립성 테스트(CI test)를 시행하여 임계값 이상의 독립성을 갖는 노드들 간에 아크를 삭제하거나 혹은 임계값 이하의 독립성을 갖는 노드들 간의 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 네트워크를 학습할 수 있는 것으로 알려져 있다[3][6].

2.4 베이지안 네트워크를 위한 특징 축소

일반적으로 한 개체를 표현하는 중요한 속성들을 특징이라고 한다. 한 개체의 분류 클래스를 판단하는데 큰 영향을 미치지 못하는 특징들은 삭제하고 반대로 중요도가 높은 특징만을 골라 이들로 분석 데이터를 표현하는 처리과정을 특징 축소(Feature reduction), 특징 부분 집합 선택(Feature subset selection), 차원 축소(Dimension reduction)등으로 부른다. 일반적으로 이와 같은 특징 축소를 통해 처리대상 데이터의 양을 줄임으로써 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)로 나누어 볼 수 있다[10]. 여과 방법은 정보 획득양, 상호 정보량 등의 척도를 이용하여 각 특징의 중요도를 개별적으로 평가하고 이것이 일정한 수준에 미치지 못하는 특징들을 삭제하는 방식이다. 이에 반해 포장 방법은 가능한 특징 집합으로부터 실제로 특정 분류기를 생성하여 이 분류기의 분류 성능을 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.

2.5 데이터 수집

60명의 백혈병 환자에게서 유전정보를 추출한다. 골수 샘플을 추출하여서 Affymetrix GeneChip array기술을 사용하여 유전자 데이터를 얻어 낸다.

GeneChip Array 기술은 1995년 미국의 스탠포드대학에서 탄생되었다. 당시 스탠포드대학은 자체 제작한 DNA칩 어레이를 이용, DNA를 약1.8㎠의 유리판 위에 고정시킨 다음, 형광물질을 이용해 유전자 발현 정도를 한 번에 측정할 수 있게 만들었다. 곧 이어 실리콘밸리 소재 Affymetrix도 'Oligo Chip'을 개발했다. Affymetrix는 컴퓨터 칩 제작에 사용되는 '포토리소그래피' 기술을 응용, 40만개의 올리고염기들을 1.28㎠되는 유리 기판 위에서 직접 합성해 칩을 제작했다. DNA 칩은 격자 모양으로 촘촘한 구멍이 있는 칩 위에, 여러 종류의 유전자 염기서열을 격자에 하나씩 집어넣은 것으로 여기서 얻은 data를 Micro Array Data라고 하며, DNA chip 혹은 gene chip이라고도 한다. 이것은 시료를 DNA 칩 위에 떨어뜨려, 시료 안에 존재하는 유전자들을 분석하거나 발현정도를 측정하는 것을 목적으로 한다[15].

데이터의 총 개수는 120개로 치료 이전의 데이터 60개와 치료이후의 60개의 데이터로 이루어져 있다. 유전자의 종류는 12,600개가 얻어지는데 Affymetrix HGU95A Array가 사용되었다. 각 유전자는 두 가지로 표현이 된다. Signal의 세기를 측정하여 수치로 표현하기도 하고 A/P Call을 사용하여 표현하기도 한다.

<표2> Gene Chip으로부터 얻어진 데이터

post	HDMTX-1	HDMTX-1	HDMTX-1	HDMTX-1	HDMTX-1	HDMTX-1
AFFX-MurIL	131	A	0.814869	114.8	A	0.99156
AFFX-MurIL	195.5	A	0.897835	330.8	A	0.945787
AFFX-MurIL	37.7	A	0.986189	113.4	A	0.993129
AFFX-MurF/	368.7	A	0.737173	401.4	A	0.979978
AFFX-BioB-	6191.1	P	0.000662	27842.1	P	0.00034
AFFX-BioB-	15308.8	P	0.00007	49224.5	P	0.00006
AFFX-BioB-	6877.9	P	0.000044	22224.5	P	0.00006
AFFX-BioC-	22128.7	P	0.00006	71208.3	P	0.000044
AFFX-BioC-	18752.3	P	0.000044	56108.3	P	0.000052
AFFX-BioDı	27926.1	P	0.000044	79416.9	P	0.000044
AFFX-BioDı	110043.7	P	0.000044	292185.6	P	0.000044

2.6 사전처리

처음에 모아진 데이터의 크기는 측정이 된 개체의 수 120개, 유전자의 개수 12,600개이다. 120개의 데이터 집단은 문제가 크게 되지 않지만 12,600개는 노드의 개수를 의미하게 된다. 즉, 모든 필드를 사용할 수 없어진다. 그래서 이 숫자를 줄이기 위해서 사전처리(Preprocessing)의 과정을 거치게 된다.

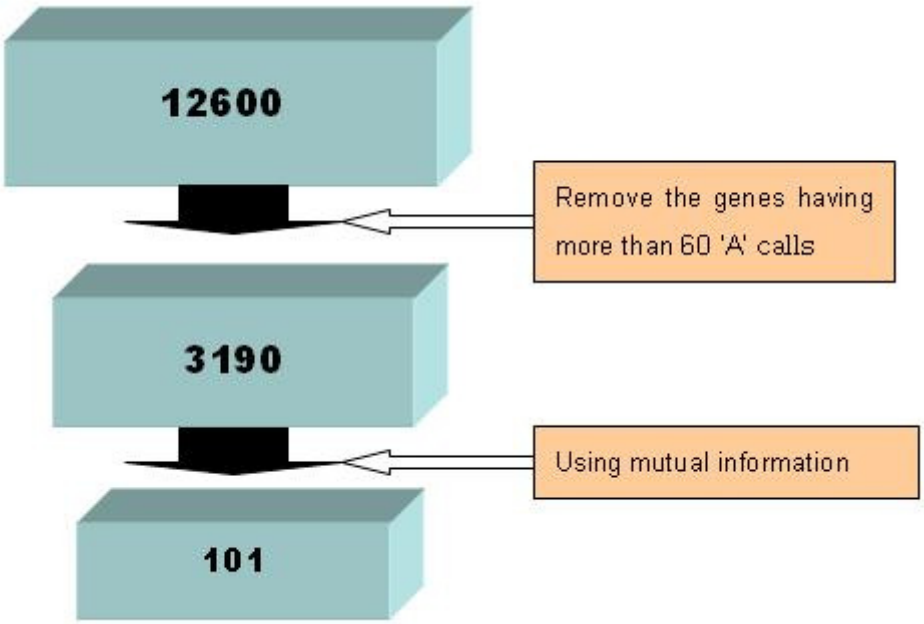
일단 60개 이상의 A 값을 가진 유전형질은 개체의 특성을 구분하는데 큰 도움이 되지 않는다는 가정 하에 제거를 하고, 각각의 형질에서 중간 값을 설정한 후에 그 값보다 큰 것을 1(High) 낮은 것을 0(Low)라고 하여 다시 데이터를 정리한다.

2.7 Using Mutual Information

이후에 데이터의 수를 줄일 수 있는 방법으로는 Mutual Information(이하 MI)을 사용해서 진행한다. MI는 Information Gain(이하 IG)안에 포함이 되는 형태로 나타난다. 식(7)은 Average Mutual Information을 표현한 식이고 log안에 들어있는 부분이 MI를 표현한다.

$$I(G;C) = \sum_{G,C} P(G, C)\log\frac{P(G, C)}{P(G)P(C)} \tag{7}$$

MI를 사용함으로서 많은 데이터에서 랜덤하게 데이터의 샘플을 재구성할 때 무작위로 선택하면서 생기는 불확실성을 줄이는 효과를 볼 수 있다.



<그림4> Preprocessing

III. 실험 및 결과

3.1 실험 목표

본 논문에서는 앞서 전처리 작업을 통해 정제된 백혈병과 유전자데이터의 집합에 대해, 베이지안 네트워크를 기초로 다양한 분석 실험을 전개하기로 한다. 특별히 의료 데이터 집합에 대해 베이지안 네트워크를 적용해보려는 이유는 이 영역 데이터가 갖는 몇 가지 특성 때문이다. 먼저 백혈병 치료를 비롯한 대부분의 의료 영역의 경우 원인-결과 관계, 요인 간 연관 관계 등 진단과 처방에 필요한 의료 지식이나 이론에 많은 불확실성과 가변성을 내포하고 있어 명확한 사전 지식을 확보하거나 단언적 추론을 전개하기 어렵다. 따라서 불확실성을 고려할 수 있는 확률적 학습과 추론 방법이 반드시 필요하다. 그리고 대부분의 유전 데이터들은 필연적으로 관측기의 오차와 관측 환자의 생리생태 및 관측 시점 등에 따라 많은 잡음과 특이값을 포함할 수밖에 없다. 따라서 이러한 데이터로부터 유전자 간의 영향과 관계를 예측하기 위해서는 특별히 잡음과 특이값에 견고한 분류 학습법을 적용하여야 한다.

그래서 다음과 같은 목표를 설정하였다.

- 베이지안 네트워크 분류기들 간의 분류 성능 비교
  주어진 유전자데이터 집합으로부터 NBN, TAN, GBN 등 제약조건이 다른 다양한 유형의 베이지안 네트워크 분류기들을 생성하고 이들이 이 데이터에서 보여주는 분류성능을 서로 비교해본다.

- 베이지안 네트워크 분류기에서 자료의 크기에 따른 성능 비교
  120개의 자료 숫자를 조절해가면서 결과를 살펴본다. 분류기의 성능도 어떻게 변화하는지 살펴본다.

- Neural Net과의 분류 성능 비교
  유전자데이터의 특성을 고려할 때 베이지안 네트워크가 다른 분류기와 비교하여 어느 정도 효과가 있는지를 알아보기 위해서 Neural Net을 사용하여 분류 성능을 비교해본다. 그리고 각 분류기의 특성을 비교해본다.

3.2 실험 방법

앞서 설명한 실험목표를 위해 전처리 작업이 끝난 데이터를 사용하여 몇 가지 유형의 베이지안 네트워크를 생성한다. 본 연구에서 실험할 베이지안 네트워크 분류기 유형은 NBN(Naive Bayesian Network), TAN(Tree-Augmented Network), GBN(General Bayesian Network) 등으로 각 분류기가 내포하고 있는 가정과 제약이 분류성능에 미치는 효과를 보기 위함이다. 또 유전자데이터에서의 베이지안 네

트워크 분류기의 강점을 확인하기 위하여 동일한 실험 조건 하에서 다른 대표적인 Neural Network 방법을 적용하고 분류성능을 비교해본다.

3.3 Using WEKA

WEKA는 실세계의 Data Mining 문제를 해결하기 위해 만들어진 기계학습 알고리즘을 모아놓은 툴이다. 자바로 만들어 졌고 거의 모든 플랫폼에서 돌아간다. 우리는 이번 실험에서 WEKA를 사용한다. 일단 사전 처리로 모아진 데이터는 WEKA에 적합한 ARFF 파일로 변환되어야 한다. 변환된 파일은 X1에서부터 X100까지의 필드와 마지막 Class 필드를 더한 101개의 필드를 가지고 데이터의 양은 120개 이다.

3.4 실험 상세구성

실험은 Weka를 사용하여 101개의 속성을 가지는 데이터의 양을 120개, 60개, 30개로 나누어 베이지안 네트워크를 구성한 후에 초기의 120개의 데이터를 Validation Set으로 사용하여 분류기의 정확성을 시험한다. 그리고 이 실험은 NBN, TAN, GBN에 모두 시행하도록 한다. NBN의 경우 가장 Naive한 형태로 어떤 옵션도 사용하지 않는다. TAN의 경우 Search Algorithm을 TAN으로 세팅하였고 GBN에서는 NBN, TAN과 차이를 주기 위하여 Search Algorithm을 HillClimber를 사용하였다. 그리고 Markov Blanket을 허용하였다.

120개의 데이터를 가지고 Neural Net에 적용하여 결과를 비교해본다.

<표3> 네트워크 분류기 종류별 Setting

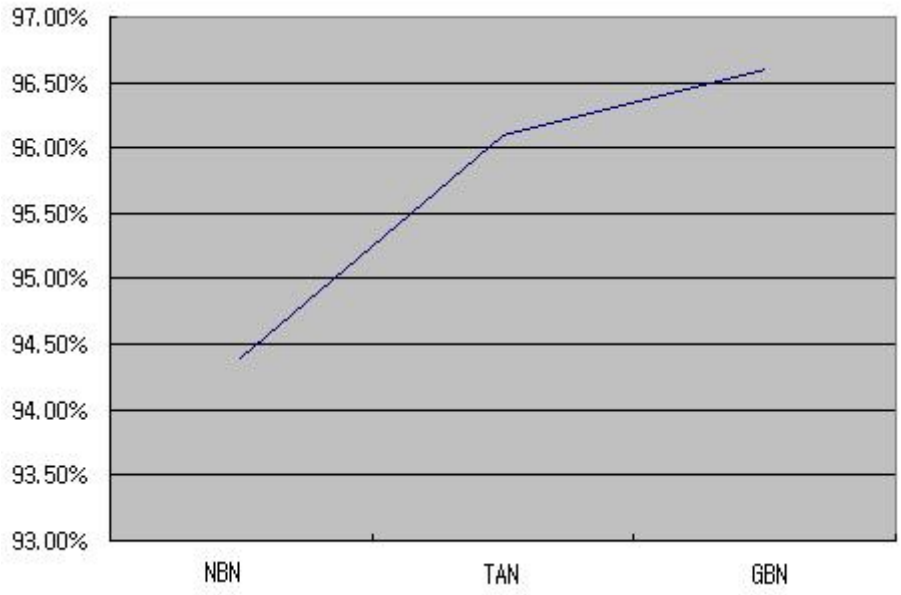
NBN	Filter : None Classifier : NaiveBayes Use Kernel Estimator : False Use Supervised Discretization : False
TAN	Filter : None Classifier : BayesNet Estimator : SimpleEstimator -A 0.5 Search Algorithm : TAN User ADTree : False
GBN	Filter : None Classifier : BayesNet Estimator : SimpleEstimator -A 0.5 Search Algorithm : HillClimber -N -P 1 -S BAYES User ADTree : False
Neural net	Filter : None Hidden Layer : half of attributes Learning Rate : 0.3 Momentum : 0.2 Training Time : 100

3.5 실험 결과

<표4> 베이지안 네트워크의 종류와 Training Set 크기에 따른 결과

	Training Set 개수	Collect 개수	퍼센트	평균
NBN	120	114	95%	94.4%
	60	117	97.5%	
	30	109	90.8%	
TAN	120	119	99.1%	96.1%
	60	116	96.6%	
	30	111	92.5%	
GBN	120	118	98.3%	96.6%
	60	117	97.5%	
	30	113	94.1%	
Neural Net	120	120	100%	100%

실험의 결과는 위의 <표4>와 같다. 예상을 했던 대로 가장 원시적인 형태의 NBN의 정확도가 가장 낮았고 GBN이 가장 높았다. 물론 GBN의 경우 정확도를 높이기 위해서 Markov Blanket을 사용하기도 하였다. 눈에 띄는 부분은 NBN에서 Training set의 개수가 120일 때 보다 60일 때 더 좋은 결과가 나왔다는 것인데 아마도 120개로 네트워크를 구성할 때 120중의 특이값 등에 의해서 형태가 무너지지 않았나 생각된다. 뉴럴넷의 경우 성능면에서는 가장 좋은 100%를 보였다.

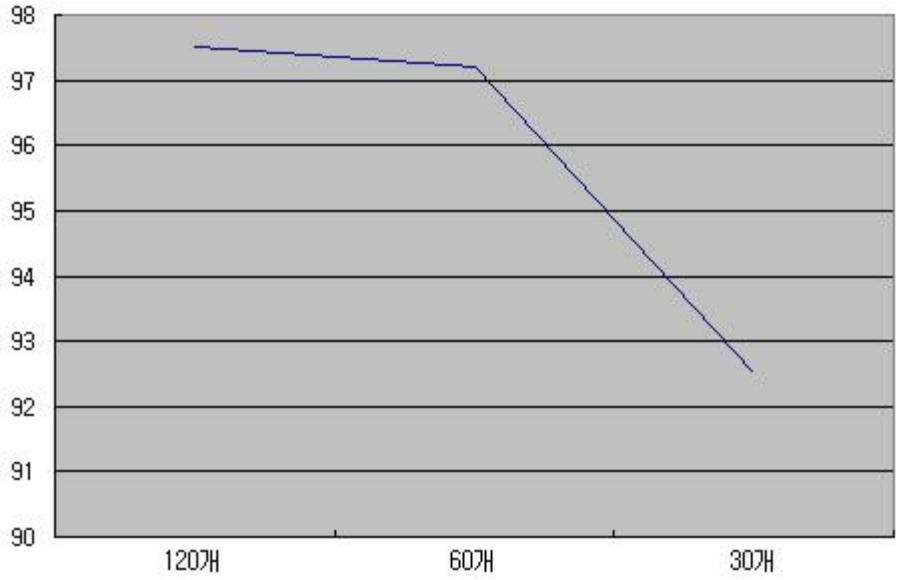


<그림5> 베이지안 네트워크 종류에 따른 정확도

<표5> Training Set의 크기에 따른 결과

집단 개수	120개	60개	30개
평균	97.5%	97.2%	92.5%

이 표는 집단의 개수를 기준으로 정확도의 평균을 낸 것이다. 예상대로 120개의 Training set으로 네트워크를 구성하고 120개의 Validation set으로 바로 테스트 하는 것이 가장 좋은 결과를 보였다.



<그림6> 베이지안 네트워크 Training set에 따른 정확도

IV. 결론 및 논의

베이지안 네트워크는 노드들 간의 관계를 규정하면서 각 관계를 네트워크 형태로 표현을 할 수 있는 방식이다. 이러한 점은 유전자데이터 간의 관계를 표현하는 데는 가장 적절한 방식이라고 생각된다. 이번 모든 실험은 GBN이 가장 적절한 성능을 보였고 Traning set역시 120개의 가장 많은 양이 좋은 결과를 보여줬다. 예상컨대 유전자의 특성상 유전자 간에 영향이 전혀 존재하지 않는다고 생각된다. NBN의 경우 노드들 간의 영향이 거의 없을 때 괜찮은 결과가 나올 수 있다고 예상이 되지만 실제 유전자데이터에서는 좋은 결과가 나오지 않았고 노드들 간의 관계를 자유롭게 만들어줄 수 있는 다른 형태의 네트워크가 좋은 결과로 연결된 것을 보면

알 수 있다. Neural Net과의 비교에서 성능면에서는 Neural Net이 뛰어났지만 앞에서 언급했듯이 베이지안 네트워크의 특성상 노드들 간의 관계를 알 수 있다는 장점을 Neural Net은 지니지 못한다. 나중에 치료목적으로 유전자데이터간의 관계를 규명해야 하고 자료로 활용을 할 경우 Neural Net은 단지 T/F의 성능에서는 좋은 모습을 보이겠지만 분석적인 면에서는 베이지안 네트워크를 따라 올 수 없지 않을까 생각이 된다.

베이지안 네트워크의 한계라고 하면 노드의 개수가 많아질수록 계산양이 기하급수적으로 증가한다는 것이다. 실제로 연구된 유전자데이터의 개수는 적게는 2만개 많게는 6만개 정도로 추정이 된다. 현실적으로 이 유전자들을 모두 네트워크에 넣어 분석을 한다는 것은 불가능하다. 베이지안 네트워크에서 확장하여 유전자중 형질과 자극에 반응하는 형식이 일정한 것들을 그룹으로 모아서 최대한 많은 유전자들을 네트워크에 포함시키고 그 그룹을 노드로 시약을 투여하기 전과 후를 비교하는 네트워크를 제작하는 방식도 현재의 베이지안 네트워크의 한계를 벗을 수 있는 하나의 방법이라 생각이 된다. 또 하나 최대한 많은 군집을 모으는 것도 하나의 과제이다. 일반 실험과는 달리 사람을 대상으로 하는 것이라 비교를 할 수 있는 대상이 많지 않고, 혹은 군집이 커진다고 하더라도 유전자데이터를 분석하는 것에 대한 비용과 시간은 기하급수적으로 증가하는 컴퓨터의 성능과 메모리의 적은 단가가 일정부분 해결해 줄 수 있더라도 쉽게 해결할 수 있는 문제가 아니다. 근본적으로 유전자데이터를 좀 더 완벽하게 분석을 하기 위해서는 새로운 형식이나 뛰어난 성능의 베이지안 네트워크와 좀 더 정형화 된 형식을 만들어 낼 수 있을 만한 양의 군집이 필요할 것이라 생각된다.

Acknowledgement

이 논문은 서울대 컴퓨터공학부 바이오지능(bioinformatics) 연구실의 장병탁 교수님의 지도를 받아 작성하였습니다.

참고 문헌

[1] Bouchaert, R., "Bayesian Belief Networks : From Construction to Inference," *Doctoral Dissertation*, University of Utrecht, The Nethelands, 1995

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

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

[4] Gorrill, Marshal-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

[5] Hecker, D., "A Tutorial on Learning Bayesian Networks," *Technical Report MSR-TR-95-06,* Microsoft Research

[6] Heckerman, D., Meek, C. and Cooper, G., "A Bayesian Approach to Causal Discovery," *Technical Report MSR-TR-75-05,* Microsoft Research, 1997

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

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

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

[10] Kohavi, R. and John G., "Wrappers for Feature Subset Selection," *Artificial Intelligence, Special issue on Relevance*, Vol.97, No.1-2 pp273-324, 1997

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

[12] 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

[13] Singh, M., "Learning Bayesian Networks fro Incomplete Data," *Proceedings of AAAI-97,* the MIT Press, pp.534-539, 1997

[14] Tom M. Mitchael, Machine Learning, *McGrow-Hill*, 1977

[15] DNA Chip. "http://blog.naver.com/ar6com?Redirect=Log&logNo=20003228980," *Naver Blog*

[16] Kyu-Baek Hwang and Byoung-Tak Zhang, "An Introduction to Bayesian Networks : Concepts and Learning from Data,"
"http://bi.snu.ac.kr/Courses/4ai05f/introBN.pdf," *SNU Biointelligence Lab*