의사결정규칙을 이용한 복합 질환의 예측

김명기*, 김승현**, 위규범*
*아주대학교 정보통신전문대학원
**아주대학교 의과대학
e-mail:{zephyran, kimsh, kbwee}@ajou.ac.kr

Prediction of complex disease using Decision Rules

Myoungki Kim*, Seung-hyun Kim**, Kyubum Wee*
*Graduate School of Information & Communication, Ajou University
**School of Medicine, Ajou University

요 약

복합 질환과 관련된 임상데이터에 대한 예측 모델을 회귀분석, 신경망, 또는 MDR과 같은 방법을 이용하여 분석할 경우 데이터의 차원 문제(Dimensionality Problem)가 발생할 수 있다. 엔트로피(Entropy)를 이용한 의사결정규칙 방법은 이러한 데이터의 차원 문제를 줄이고 의사결정규칙의 결과를 바로 해석할 수 있다는 점에서 질환 예측 모델을 만드는데 유용하다. 본 논문에서는 천식과 관련된 임상데이터를 사용하여 예측 모델을 구성하고 결과를 분석한다.

1. 서론

인간의 질병을 유발하는 원인 중 하나는 유전자 변이로서, 일반인이 가지고 있는 유전체 정보에 변이가 생겨 이것이 질병을 유발하는 경우를 말한다. 이러한 유전변이의 원인을 살펴보면 유전자의 염기서열이 다른 사람과 다르다는 것을 알 수 있으며 이러한 변이를 단일염기다형성(Single Nucleotide Polymorphism)이라고 부른다. 유전자의 SNP이 일어나면 원래 유전자의 기능을 제대로 수행하지 못할 가능성이 있으며, 질병의 원인을 밝혀내는 한 가지 방법으로서 이러한 기능 장애를 일으키는 SNP 혹은 SNP 그룹을 찾는 노력이 진행되고 있다.

질병과 연관이 있을 것으로 예상되는 후보 유전자를 골라서 이 유전자들의 SNP을 모으면 수십에서 수백, 혹은 SNP 칩을 이용하면 수 만개 이상의 SNP 데이터가 만들어진다. SNP의 수가 증가할 경우에, 이 중에서 질병과 연관성이 높은 SNP들을 찾아내고 각 SNP이 어떤 유전형(genotype)일 때 질병이 발생할 확률이 높은지 판단하는 것은 쉽지 않은 작업이다.

다량의 SNP 데이터에서 질병과 연관된 SNP들을 골라내는 방법은 고전적인 방법인 회귀분석(Regression) 방법과 신경망(Neural Network)을 이용한 PDM(Parameter Decreasing Method) 방법[1], HWE(Hardy Weinberg Equilibrium)과 AA(Allelic Association) 값을 이용한 Set Association 방법[2], 그리고 가장 널리 사용되고 있는 MDR(Multifactor Dimensionality Reduction) 방법[3]과 함께, 이 논문에서 실험한 의사결정나무(Decision Tree)[4] 및 의사결정규칙(Decision Rule)에 이르기까지 다양한 방법이 제안되었다[5]. 이 중 회귀분석, 신경망, MDR을 이용한 방법은 데이터의 차원 문제와 블랙박스 모델에 따른 각 SNP간의 상호 연관성을 알기 어렵다는 문제점이 있다. Set Associaion 방법은 유전형의 분포가 HWE를 따르지 않는 성염색체 상에 존재하는 경우 AA와 HWE의 통계량을 대신하는 다른

이 논문은 2006년도 정부의 재원으로 한국과학재단의 지원을 받아 수행된 연구임 (No. R01-2006-000-10775-0)

통계량을 도입해야 하며, 각 SNP이 어떤 유전형일 때 질병과 연관이 있는지 알 수 없는 문제점이 있다. 그러나 의사결정규칙 방법은 이러한 데이터의 차원 문제와 SNP간 연관성 해석 그리고 유전형의 분포에 상관없으며 처리속도가 빠르고 결과를 쉽게 해석할 수 있는 방법이다. 본 논문에서는 이 방법을 이용하여 천식과 관련된 임상 데이터를 분석한다.

2. 규칙기반 분류 방법

규칙 기반 분류 방법(Rule Based Classification Approach)은 If-Then과 연산자 and로 이루어진 규칙(Rule)을 모은 규칙 세트(Rule Set)를 만들고 이 규칙 세트의 규칙에 따라 데이터를 맞춰보고 맞으면 데이터의 결과를 해당 규칙의 결과로 예측하는 방법으로 그림1과 같은 형태를 가지고 있다.

```
If A = 'a' and B = 'b' Then Result = 'r1'
If A = 'b' Then Result = 'r2'
```

(그림 1) 규칙 세트의 구조

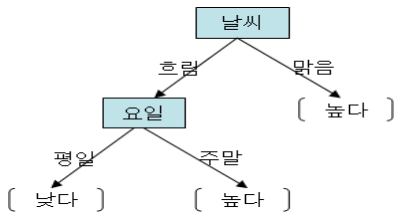
이러한 규칙 세트의 규칙이 상호 배제(Mutually Exclusive)한 경우에는 규칙끼리 서로 상반된 결과를 가지게 되는 경우가 없으나, 규칙이 서로 다른 결과를 나타내는 경우에는 가장 적용이 많이 된 규칙을 먼저 사용하는 방법과 중요도에 따라 규칙 순서를 정하여 적용하는 방법을 이용한다. 또한, 규칙 세트에 규칙이 복잡해지고 많아지는 것을 방지하기 위해 기본 분류를 정하기도 하는데, 이러한 경우 어떠한 데이터가 규칙 세트에 해당하지 않는 경우에는 기본 분류에 따라 결과를 결정한다.[6]

규칙 기반 분류 방법은 회귀 분석(Regression)이나 신경망(Neural Network)과 같은 방법들과는 달리 명목형(Categorical) 변수 처리에 알맞고 블랙박스(Blackbox) 모델이 아니므로 규칙 세트만으로도 사람이 그 내용을 이해하기 쉽다는 장점 때문에 다양한 분야에서 널리 사용되고 있다.

3. 의사결정나무와 의사결정규칙

의사결정나무(Decision Tree)는 노드(Node)와 에지(Edge)로 연결된 뿌리나무(Rooted Tree)로 각 노드는 데이터의 값에 따라 분류를 하는 분류기 역할을 하며 이 분류 결과에 따라 상위에서 하위로 에지를 따라 진행하면서, 단말 노드에 이르면 단말 노드의 분류 결과 값으로 데이터의 결과를 결정하는 방법이다. 이러한 의사결정나무를 생성하는 방법은 크게 두 가지로 Gini Index를 이용하는 CART 방법과 Information Gain Ratio를 이용하는 C4.5 방법이 있다. 이 C4.5는 Quinlan[7]에 의해 고안되었으며 ID3 알고리즘에 바탕을 두고 있다. C4.5를 이용하여 의사결정나무를 만드는 방법은 두 가지로, 우선 배치 모드(batch mode)라고 하여 전체 학습 데이터(learning data)를 모두 사용하여 한 개의 의사결정나무를 만들어 내는 방법이며, 다른 한 가지 방법은 반복 모드(iterative mode)라고 하여 전체 학습 데이터를 모두 사용하지 않고 일부를 랜덤하게 선택하여 이를 가지고 의사결정나무를 만들고 나머지 데이터를 추가하면서 의사결정나무를 만드는 방법이다. 이때, 랜덤하게 선택되는 학습 데이터가 매번 다르므로, 매번 다른 의사결정나무를 만들어 내게 된다. 그러므로 같은 데이터로 많은 의사결정나무를 만들고 이중 가장 좋은 성능을 보이는 의사결정나무를 선택하게 되므로 테스트 데이터로 실험하는 경우 더 높은 정확도를 보인다. 이 논문에서는 의사결정나무 생성방법으로 C4.5의 반복 모드를 사용한다.

이렇게 위와 같은 의사결정나무를 만들어내면 각각의 단말 노드는 최종 분류 결과 값을 가지게 되는데, 이 결과 값이 같은 것끼리 루트 노드부터 단말 노드까지의 분류하는 방법을 and 연산자로 묶어 내면, 규칙이 만들어지고 이 규칙들을 모으면 규칙 세트가 만들어 진다. 이렇게 만들어진 의사결정규칙은 각 규칙이 상호 배제한 결과를 가지게 되어 쉽게 분류가 가능하지만, 규칙의 길이가 길고 비슷한 규칙이 여러 번 나오게 되므로 이를 바로 한눈에 알아보기 쉽지 않다. 따라서 이를 규칙 생성(Rule Induction) 방법을 통해 다시 정리를 하면 간단한 규칙을 만들 수 있다. 규칙 생성 방법을 통해 정리된 규칙은 의사결정나무에서 바로 만들어진 규칙과는 달리 각 규칙사이에 상호 배제한 관계를 가지고 있지 않아 규칙을 적용 하였을 때 서로 반대되는 결과를 나타낼 수도 있다. 이때에는 결과 클래스의 정확도를 확인하여, 정확도가 높은 쪽으로 결정한다.

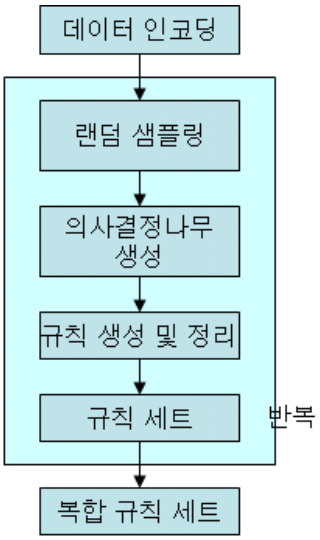


(그림 2) 교통체증에 대한 의사결정나무

If 날씨=흐림 and 요일=주말 Then 교통체증=높다
If 날씨=흐림 and 요일=평일 Then 교통체증=낮다
If 날씨=맑음 Then 교통체증=높다

(그림 3) 그림 2로 생성한 의사결정규칙 세트

이 논문에서는 반복 모드를 사용하여 여러 개의 의사결정나무를 만들게 되며, 각 의사결정나무마다 각각 다른 의사결정규칙 세트를 만들어내기 때문에 각기 다른 의사결정규칙 세트를 조합하여 복합규칙세트(Composite Rule Set)를 만들고 이 규칙세트로 결과를 뽑아내게 된다.



(그림 4) 실험 방법

3. 임상 데이터 인코딩 방법

SNP 유전형과 그에 따른 질병 데이터를 의사결정나무에 적용하기 위해 우선 각 SNP 유전형과 질병 여부를 인코딩을 한다. 이때 인코딩은 SNP이 질병의 표현형(Phenotype)이 어떤 유전 모델인지에 따라 다른 값을 가지게 되는데 우성우위모델(Dominant)인 경우 유전형이 우성+우성, 우성+열성 인 경우 1을 열성+열성인 경우 3으로 인코딩하였다. 공동우성모델(Codominant)은 우성+우성인 경우 1을 우성+열성인 경우 2를 열성+열성인 경우 3으로 인코딩하고 열성우위모델(Recessive)은 우성+우성을 1로 우성+열성, 열성+열성의 조합인 경우를 3으로 인코딩하였다. 이렇게 인코딩된 데이터를 가지고 반복 모드로 10번 의사결정나무를 생성하고 이 의사결정나무에서 의사결정규칙을 생성하여 최종 복합 규칙 세트로 테스트 데이터의 정확도를 측정한다.

4. 실험 결과

AIA형 천식과 ATA형 천식이 있는 환자들을 대상으로 ATA형 천식을 유발하는 것과 관련이 있을 것으로 생각되는 53개의 SNP 유전형 데이터를 수집하였으며, 이 SNP 데이터로부터 어떤 SNP이 어떤 유전형일 때 ATA형 천식에 더 잘 걸리는지 분석해 보았다. AIA형 천식을 가지고 있는 환자는 총 160명이며, ATA형 천식을 가지고 있는 환자는 총 222명이었다. 각 SNP이 어떤 모델로 질병에 관여하는지 알 수 없기 때문에 우성우위, 공동우성, 열성우위 모델을 함께 한꺼번에 의사결정규칙을 생성하였으며, 같은 데이터를 우성우의 모델만 이용해서도 실험하였다. 실험 결과 모든 모델을 함께 사용한 경우 3개의 규칙을 가진 정확도가 72.5%인 규칙세트를 찾아내었으며, 우성우위만 사용한 경우에는 정확도 73.3%인 3개의 규칙 세트를 찾아내었다. 이 실험에서의 정확도는 학습 데이터에 대한 정확도이다.

SNP의 이름은 우성우위(Dominant), 공동우성(Codominant), 열성우위(Recessive) 모델의 첫 두 글자와 SNP의 ID 번호를 이용하여 조합하였다.

AIA
RE_11=3 & DO_17=1
DO_11=1 & CO_15=1 & DO_36=1
ATA
DO_0=1 & DO_21=1
Default : ATA

(그림 5) 모든 모델에 대한 의사결정규칙 세트

<표 1> 모든 모델에 대한 예측 결과

예측 \ 실제	AIA	ATA	합계
AIA	$104^{(a)}$	$49^{(b)}$	153
ATA	$56^{(c)}$	$173^{(d)}$	229
합계	160	222	382

$$\text{정확도} = \frac{a+d}{a+b+c+d} = \frac{104+173}{382} = 72.5\%$$

(수식 1) 모든 모델에 대한 예측 결과의 정확도

AIA
DO_8=1 & DO_11=3 & DO_11=1
ATA
DO_21=3
DO_0=1 & DO_17=1 & DO_21=1 & DO_49=1
Default : ATA

(그림 6) 우성우위 모델에 대한 의사결정규칙 세트

<표 2> 우성우위 모델에 대한 예측 결과

예측 \ 실제	AIA	ATA	합계
AIA	157	99	256
ATA	3	123	126
합계	160	222	382

$$\text{정확도} = \frac{157+123}{382} = 73.3\%$$

(수식 2) 우성우위 모델에 대한 정확도

5. 결론

의사결정나무를 이용하여 생성한 의사결정규칙은 데이터의 수나 SNP 수 그리고 질병을 일으키는 원인 SNP 수에 따라 계산의 복잡도의 영향을 많이 받지 않는다. 따라서 다량의 SNP 데이터를 분석하여 다수의 원인 SNP을 찾는데 효과적인 방법이다. 천식과 관련된 임상데이터를 통해 실험해 본 결과 정확도 73.3%의 양호한 결과를 얻었다. 이렇게 찾아진 의사결정 규칙세트의 규칙이 실제 병과 연관이 있는지에 대한 의학적인 검증을 수행하고 있다.

향후 연구로는 발병율의 높고 낮음에 따라 의사결정규칙이 얼마나 다른 성능을 나타내는지에 대한 연구가 필요하다. 또한, 의사결정규칙에서 사용하는 판단 기준 대신에 교차비(odds ratio)와 같은 다른 판단 기준을 사용할 경우에는 어떻게 달라지는지에 대해서도 연구가 필요하다.

참고문헌

[1] Y. Tomita, S. Tomida, Y. Hasegawa, Y. Suzuki, T. Shirakawa, T. Kobayashi, H. Honda, "Artificial neural network approach for selection of susceptible single nucleotide polymorphisms and construction of prediction model on childhood allergic asthma", Bioinformatics 5:120, 2004.

[2] J. Hoh, A. Wille, J. Ott, "Trimming, weighting, and grouping SNPs in human case-control association studies", Genome Research 11:12, 2001.

[3] L. W. Hahn, M. D. Ritchie, J. H. Moore, "Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions", Bioinformatics 19:3, 2003.

[4] D. H. Kim, S. Uhmn, S. W. Cho, K. B. Hahm, J. Kim, Predictive Model for Chronic Hepatitis Susceptibility from Single Nucleotide Polymorphisms, BIOINFO 2006, pp. 35-38.

[5] A. G. Heidema, J. M. Boer, N. Nagelkerke, E. C. Mariman, D. L. van der A, E. J. Feskens, "The challenge for genetic epidemiologists: how to analyze large numbers of SNPs in relation to complex diseases", BMC Genetics 7:23, 2006.

[6] J. Han, M. Kamber, "Data Mining: Concepts and Techniques", Second Edition, Morgan Kaufmann, 2006.

[7] R. Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann, 1993.