부울 대수를 이용한 복합질환의 중요 SNP 찾기

임상섭*, 김승현**, 위규범*
아주대학교 정보통신 전문대학원*
아주대학교 병원 알레내과**
e-mail:{leemss, kimsh, kbwee}@ajou.ac.kr

Detection of SNPs involved in the development of complex diseases with the boolean algebra

Sangseob Leem* , SeunghyunKim** , Kyubum Wee*
Graduate School of Information & Communication, Ajou University, Suwon,Korea*
Department of Allergy & Rheumatology, Ajou University School of Medicine, Suwon,Korea**

요 약

복합질환(complex disease)의 원인과 작용 모델을 찾기 위해 여러 가지 통계적인 방법들과 기계 학습(machine learning)의 방법 등이 사용되고 있다. 소수 SNP의 작용모델을 찾는 방법은 많이 알려져 있지만 다수 SNP의 작용 모델을 효과적으로 찾는 방법은 거의 연구되어 있지 않다. 본 연구에서는 원인 SNP들의 작용을 부울 식(boolean expression)으로 나타내고, 유전 알고리즘(genetic algorithm)을 이용하여 예측 정확도가 높은 부울 식을 구성하였으며 실제 자료와 생성된 자료에 대하여 제안한 모델의 성능을 측정하였다.

1. 서 론

복합질환(complex disease)의 원인과 그 작용 모델을 찾기 위해서 여러 가지 통계적인 방법들과 기계학습(machine learning)등의 방법들이 사용되고 있다[1, 2].

통계적인 방법들은 수학적으로 원리와 타당성이 증명되어 있는 반면에, 분석해야 하는 자료의 수가 많아지면 통계적 모델을 찾는 시간이 늘어 매우 오래 걸린다. 예를 들어 로지스틱 회귀분석(logistic regression)의 방법을 이용할 경우, 입력 인자들의 수가 많아지면 통계적 모델을 찾는 것이 거의 불가능하다[3].

p-value, 엔트로피 등을 이용하여 모든 SNP의 조합을 계산해 보는 것은 후보 SNP의 수가 증가하면 실행시간이 오래 걸리는 어려운 단점이 있다. SVM의 경우에는 최적해를 찾는 것이 수학적으로 증명되어 있지만, 소수의 자료에 의해 판단 모델이 결정될 수 있고, 사용하는 kernel의 종류에 따른 한계가 존재하고, 작용기작을 눈으로 쉽게 알아보기 힘들며, SNP의 수가 늘어나면 모든 조합의 수를 실험해 보기 어려운 단점이 있다[4, 5, 6].

인공신경망(neural network)을 이용한 방법은 SNP의 수가 늘어나도 계산시간이 크게 늘어나지 않고, 예측 정확도가 일반적으로 높은 장점이 있다. 그러나 인공신경망을 이용해 찾은 판단 모델이 겉으로 드러나지 않는 가려진 모델(black box)이기 때문에 작용 기작을 파악하는 것이 쉽지 않다[7, 8, 9].

결정트리(decision tree) 방법은 판단모델을 분석하기 쉽고 계산시간이 비교적 빠르나, 소수의 데이터에 의해서 전체적인 판단 모델이 크게 바뀔 수 있는 문제가 존재한다[10, 11].

MDR(multifactor dimensionality reduction) 방법은 적용모델을 결정하지 않고 사용할 수 있다는 장점이 있지만, 모든 조합을 다 시험해 보기 위해서 많은 시간이 소비되고 다차원 분석을 할 경우 분할표에서 비어 있는 원소가 많아져, 테스트 데이터에 대한 판단기준이 없을 확률이 높아지는 문제가 있다[12].

GABA(genetic algorithm with boolean algebra)에서는 작용기작을 부울 식(boolean expression)으로 나타낼 수 있다는 가정을 하고, 높은 예측 정확도를 가지는 부울 식을 찾기 위하여 유전 알고리즘을 사용하였다. GABA는 작용기작을 부울 식으로 표현하기 때문에 파악하기 쉽다. 그러나 GABA에서 사용하는 부울 식은 괄호가 포함되어 있지 않으므로 나타낼 수 없는 작용기작이 존재하게 된다[13].

본 연구에서는 괄호가 포함된 부울 식을 사용하여, 이를 결정회로(decision logic)라 부르며, 유전 알고리즘을 이용하여 결정회로를 구성하고 성능을 분석하였다.

2. 본 론

2.1 SNP 인코딩

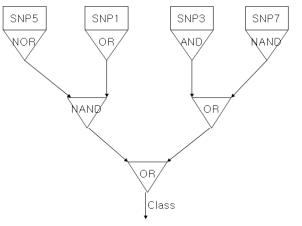
SNP 정보는 대립형질이 결합되어 있는 형태, 즉 유전자형(genotype)으로 입력되고 1, 2, 3 가운데 하나의 값으로 주어진다. 이때 1은 major/major, 2는 major/minor, 3은 minor/minor 대립형질이 결합된 형태를 나타낸다.

2.2 결정회로(decision logic)

결정회로는 개인의 SNP 정보를 입력으로 받아서 환자군과 대조군을 구별하는 회로이다. 일반적인 논리 회로에서 모든 발생 가능한 회로의 경우를 생각해 보면 무한히 많은 경우가 있으므로, 이를 다 실험해 볼 수 없다.

따라서 일반적인 논리 회로에 아래와 같은 제한을 주어서, 이를 결정회로라 부르기로 한다.

(1) gate는 AND, OR, NAND, NOR gate들로 제한된다.
(2) gate는 input의 개수가 2개로 제한된다.
(3) 각 신호들은 input으로는 한 번만 쓰일 수 있다. 즉 두 개 이상의 gate의 input으로 중복되게 사용될 수 없다. 이러한 조건을 만족할 경우, n(SNP의 개수) 개의 입력 신호가 n-1 개의 gate를 지나면 하나의 output으로 만들어 진다.



(그림 1) 결정회로의 예

위의 그림 1은 n(SNP의 개수)이 4인 결정 회로의 한 예이다.

2.2.1 결정회로의 구현

앞에서 언급한 제한 사항을 포함한 실제 결정 회로는 SNP, selector, gate, sequence 로 구성된다.

<표 1> 결정회로의 구성

SNP 1 ~ n	selector 1 ~ n	gate 1 ~ n-1	sequence 1 ~ n-1

n개의 SNP로 이루어진 하나의 결정 회로를 보면 앞의 표 1과 같다.

(1) SNP: 결정회로의 입력으로 들어올 SNP들의 번호이다. 전체의 SNP 가운데서 n 개의 SNP를 선택하는 것이다. 선택된 SNP정보를 1, 2, 3 중 하나의 입력으로 받아오게 된다.

(2) selector: 각 SNP가 dominant/recessive model로 동작함을 나타내기 위한 것이다. gate 중 하나이고, 실제로는 0이 AND, 1이 OR, 2가 NAND, 3이 NOR gate를 나타낸다. 이때 SNP정보가 1은 major/major를 나타내므로 1/1, 2는 major/minor를 나타내므로 1/0, 3은 minor/minor를 나타내므로 0/0으로 바꾼 다음, i번째 SNP정보가 i번째 selector의 input으로 들어가게 된다. 그러므로 OR와 NOR는 dominant model을 나타내고, AND와 NAND는 recessive model을 나타낸다.

(3) gate: sequence와 함께 모델이 작용하는 기작을 나타내고, gate에 2개의 input이 들어와 1개의 output이 나가므로, n-1개의 gate가 필요하게 된다.

(4) sequence: 각 값들은 gate에 입력되는 SNP의 index를 의미한다.

2.2.2 결정회로의 동작

다음은 결정회로가 동작하는 예이다. 입력은 4개의 SNP로 이루어지는 경우를 가정한다.

<표 2> 결정회로 예

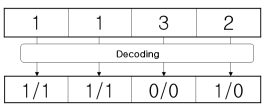
SNP				selector				gate			sequence		
5	3	1	7	3	1	0	2	1	2	1	2	1	1

<표 3> 입력 자료 예

SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9
3	2	1	2	1	2	2	1	3

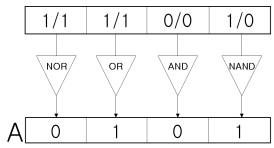
결정회로가 표 2와 같이 표현되고, 이때 입력되는 자료가 표 3과 같다면 결정회로의 동작은 다음과 같다.

(1) 1, 2, 3의 값 중 하나의 값으로 코딩되어 있는 자료를 부울식에 사용할 수 있도록 0, 1의 조합으로 바꾸는 전처리 과정을 수행한다. 각 SNP에 해당하는 index의 genotype을 가져온 후에 decoding을 한다. 이때 1(major/major)일 경우는 1/1, 2(major/minor)일 경우는 1/0, 3(minor/minor)일 경우는 0/0의 입력으로 처리된다.



(그림 2) decoding의 예

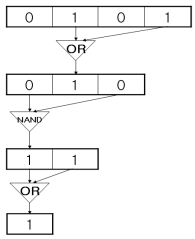
(2) 각 selector를 적용한다. 표 2에서 selector가 3-1-0-2 인데 각 해당하는 gate가 NOR, OR, AND, NAND이므로, 다음 그림과 같이 동작하게 된다.



(그림 3) selector 적용

(3) i번째 sequence가 가리키고 있는 index에 저장된 값(A[sequence[i]])과 마지막 index에 저장된 값(A[n-i+1])이 i번째 gate에 input이 되고, output을 i번째 Sequence가 가리키고 있던 index에 저장한다. 주어진 예에서 i가 1인 경우를 살펴보면, sequence의 첫 번째 값은 2이므로 그림 3의 A에서 index의 2와 4의 값인 1과 1이 첫 번째 gate의 입력이 된다.

(4) (3)의 동작을 n-1번 반복한다. 다음의 그림 4은 (3)의 동작을 반복하는 것을 나타낸다.



(그림 4) sequence(2-1-1)에 따른 gate 적용

(5) 최종 결과가 1이면 환자로 판단하고, 0이면 대조군으로 판단한다.

2.2.3 결정회로의 의미

앞의 예에서 표 2의 결정회로를 그림으로 표현하면 앞의 그림 1과 같다. 앞의 그림 1에서 표 2와 SNP1과 SNP3의 위치가 바뀌었는데 이는 회로의 선들이 겹쳐져 보이지 않게 하기 위함이다.

각 SNP에 해당하는 selector들은 각 SNP들이 작용하는 모델을 결정해 준다. 예를 들어 OR와 NOR는 major allele의 존재여부에 의해 output이 결정되므로 dominant 모델을, AND와 NAND는 minor allele의 존재여부에 의해 output이 결정되므로 recessive 모델을 나타낸다. 여기서 codominant 모델은 하나의 SNP가 중복되어 입력으로 선택되었을 경우에 나타낼 수 있다.

Gate들에 NAND와 NOR가 포함된 이유는 NOT gate를 추가하여 관리하는 것보다 구현하는 것이 간단하였기 때문이다.

그림 1의 결정회로를 분석해 보면 SNP5가 1 또는 2 이거나, SNP1이 3, SNP3이 1, SNP7이 2 또는 3일 경우 환자군으로 판단하는 모델이다. 표 2와 같은 결정회로가 어떤 자료의 실행 결과로 나온다면, SNP1, 3, 5, 7이 환자군과 대조군을 구분하는 정보임을 의미하며, 더 나아가서 각 SNP들의 작용 기작까지 예상해 볼 수 있다.

2.3 유전 알고리즘

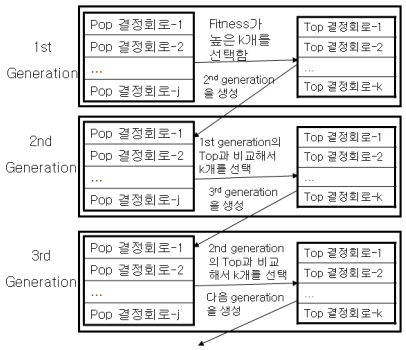
환자와 대조군을 분류하는 정확도가 높은 결정회로를 찾기 위해서 SNP, selector, gate, sequence의 모든 조합을 실험해 보는 것은 거의 불가능하다. SNP의 수가 늘어나게 되면 현실적으로 계산 불가능하다. 본 연구에서는 유전 알고리즘을 사용하였다.

결정회로를 일반적인 유전 알고리즘의 한 개체로

사용하기에는 결정회로의 구성하는 요소들(SNP, selector, gate, sequence)의 특성이 모두 다르기 때문에, 유전 알고리즘을 적합하게 변형하여 사용하였다.

2.3.1 유전 알고리즘의 변형

변형된 유전 알고리즘에서 하나의 세대는 Top과 Pop으로 이루어진다. Top은 우수한 형질의 개체들을 모아 놓은 집단이고 Pop은 일반적인 개체들의 집합이다. 여기서 하나의 개체는 하나의 결정회로를 나타낸다.



(그림 5) 유전 알고리즘의 변형

위의 그림 5에서 Pop 결정회로-j로 표현된 부분이 Pop개체들이고, Top 결정회로-k로 표현된 부분이 Top개체들이다.

첫 세대는 무작위로 생성된 초기집단을 Pop으로 설정하고, 그 가운데 우수한 k개의 개체들을 선택하여 Top으로 설정한다. 다음 세대의 Pop개체들은 랜덤하게 생성된 후에 각 Pop개체들이 일정한 확률을 가지고 이전 세대의 Top개체 속성들을 상속 받는다. 알고리즘을 정리하면 다음과 같다.

(1) Pop개체들을 랜덤하게 생성한다.
(2) 각 Pop개체들의 fitness값을 계산한다.
(3) Pop개체들 가운데 우수한 k개의 개체를 선택하고, 이를 Top이라 부른다.
(4) 다음 세대의 Pop개체들을 생성한다. 하나의 Pop개체가 생성되는 과정은 다음과 같다.
  (a) 하나의 Pop개체를 랜덤하게 생성한다.
  (b) Roulette-wheel selection을 통하여 앞 세대의 Top개체 중 하나의 개체를 선택하고, 선택된 Top개체의 속성들이 일정한 확률에 의해 Pop개체로 유전된다.
  (c) (b)의 과정을 0번, 1번, 또는 2번 실행한다. (b)의 과정을 0번 실행하는 경우는 랜덤으로 생성하는 개체이고, 1번 실행하는 경우는 Top의 개체가 다음 세대에 보존되는 개체이고, 2번 실행되는 경우는 Top의 두 개체를 교배하여 생성하는 개체이다.
(5) (2)~(4)의 과정을 정해진 횟수만큼 반복한다.

2.3.2 적합도(Fitness)

본 연구에서는 적합도 함수(fitness function) F를 다음과 같이 정의하였다.

F=민감도+특이도+(민감도*특이도)

위의 식에서 민감도(sensitivity)는 환자를 환자로 판정하는 정확도를 나타내며, 특이도(specificity)는 환자가 아닌 사람을 환자가 아닌 것으로 판단하는 정확도이다. 위와 같은 식을 사용한 이유는 민감도와 특이도가 일반적으로 실험모델의 성능을 판단하는 데 많이 사용되는 지표이고, positive predictive value, negative predictive value, likelihood ratio등 다른 지표들 역시 민감도와 특이도의 함수 형태로 나타낼 수 있기 때문이다. 마지막 민감도와 특이도의 곱은 두 지표 모두 높은 값을 취하는 모델을 찾기 위함이다[13].

2.4 실험

결정회로를 실제 자료와 생성된 자료에 적용해 보았다. 이때 결정회로의 가장 큰 장점이라 생각되는 sequence 도입의 효과를 입증하기 위해, 유사 GABA 결정회로와 결정회로의 성능을 비교하였다. 유사 GABA 결정회로는 괄호가 없는 부울 논리식을 의미하며, 결정회로에서 sequence를 내림차순 형태(n-1, n-2, …, 2, 1)로 고정한 경우에 해당한다.

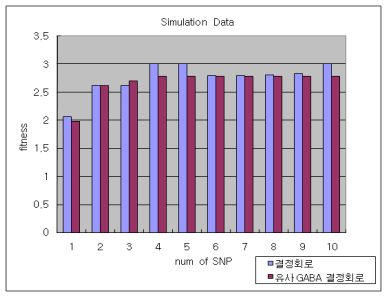
2.4.1 생성 자료를 이용한 실험

자료의 생성에서 실제로 실험할 자료를 목표로 하여 실험하므로, SNP의 수를 25로 실제 자료와 같게 하였고 환자군과 대조군은 각 125명씩 총 250명을 생성하였다. 작용모델은 SNP3, SNP10, SNP19, SNP20, 네 개의 SNP가 서로 상호 작용하도록 설정하였다. 부울 식으로 나타내면 다음과 같다.

환자군 = (SNP10 × SNP19) + (SNP3 × SNP20)

SNP10, SNP19, SNP20은 dominant model로, SNP3은 recessive model로 동작하도록 설정하였다. 자료의 생성은 유전자형을 무작위로 발생시킨 후에 boolean algebra를 적용하여 환자인지 대조군인지를 판단하고, 환자와 대조군이 모두 125명이 될 때까지 자료를 생성한다.

생성된 자료를 대상으로 프로그램을 실험할 때는 한 세대의 개체수를 2000으로 하였고, 세대의 반복수를 n(SNP의 개수) * 200으로 실험하였다. 예를 들어 n이 5일 경우는 세대별 개체수는 2000이고, 세대의 반복수가 1000이다.



(그림 6) 생성 자료 fitness 도표

n(SNP의 개수)이 4, 5, 10 일 때 결정회로는 최대 fitness에 도달하였고, 유사 GABA 결정회로는 도달하지 못하였다.

n이 1, 2, 3 일 때 도달하지 못하는 이유는 4개 SNP의 상호작용을 표현하기 위해 n이 작다고 생각되며, 6~9에서 도달하지 못하는 이유는 세대의 개체수와 반복수가 충분하지 못한 것으로 생각된다.

유사 GABA 결정회로가 도달하지 못하는 이유는 표현할 수 있는 부울 식에 제한이 있기 때문이라고 생각된다.

2.4.2 실제 자료를 이용한 실험

실제 자료는 본 연구자의 소속기관의 부속병원의 천식관련 SNP 데이터를 이용하였다.

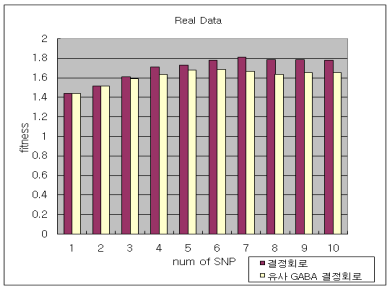
실제 자료는 총 390명의 SNP정보가 주어진다. AIA 94명, ATA

152명, 대조군 (normal control) 144명으로 이루어져 있다.

AIA는 aspirin-induced asthma의 약자로서, 아스피린에 의해 발병한 천식을 의미하며, ATA는 aspirin-tolerant asthma의 약자로서, 아스피린과 관계없는 천식을 의미한다. 그리고 대조군은 천식이 걸리지 않은 사람들이다.

본 연구에서는 AIA와 ATA를 환자군과 대조군으로 선정하여 실험하였다.

실제 자료를 대상으로 실험한 경우는 한 세대의 개체 수는 2000으로, 세대의 반복수는 생성된 자료에 비해 10배 큰 n(SNP의 개수) * 2000으로 실험하였다.



(그림 7) 실제 자료 fitness 도표

전체적으로 보면 유사 GABA 결정회로보다 결정회로가 우수한 성능을 보이고 있고, 실제 자료에서 n이 7일 때 fitness값이 최고임을 알 수 있다.

일반적인 결정회로에서 n이 7일 때 찾은 작용 모델을 보면 다음과 같다.

환자군 = (((SNP21 + SNP17) × $\overline{SNP22}$) + (SNP15 + $\overline{SNP6}$)) × ($\overline{SNP21}$ + $\overline{SNP11}$)

각 selector를 출력하여 확인해 본 결과, SNP15는 AND, SNP17은 NAND로 구성되어 recessive모델로 동작하고, SNP21, SNP22, SNP11은 OR, SNP6은 NOR로 구성되어 dominant model로 동작하고 분할표는 다음과 같다.

<표 4> 분할표(contingency table)

실제 \ 실험	환자군(AIA)	대조군(ATA)
환자군(AIA)	58	36
대조군(ATA)	40	112

이때 나타나는 여러 지표들의 값은 민감도(sensitivity)는 0.617, 특이도(specificity)는 0.7368이다. 정확도는 0.691이고, Chi-square 값은 30.8386로 1/1000 보다 작은 확률 값을 나타내므로 통계적으로 상당히 유의하다고 보인다.

3. 결론

본 연구에서는 복합질환 작용 기작을 찾기 위하여 결정회로라는 모형을 제안하였으며 생성 자료와 실제 자료에 적용하여 실험하였다.

결정회로를 일반적인 형태와 일차원적인 형태로 동작시켜 보았을 때 일반적인 형태가 우수한 성능을 보여주었다. 결정회로에 sequence가 포함될 때 높은 성능을 보인다는 것을 알 수 있었으며 sequence가 boolean algebra에서 괄호를 표현해 주기 때문이라고 생각된다.

결정회로는 GABA의 괄호가 없는 일차원적인 부울 논리식을 일반적인 부울 논리식으로 확장한 것이다. 실험을 통하여 결정회로가 더 좋은 성능을 보임을 확인하였다.

참고 문헌

[1]. 이재원, 박미라, 유한나, "생명과학연구를 위한 통계적 방법", 자유 아카데미, 2005.

[2]. A.L. Tarca, V.J. Carey, X. Chen, R. Romero, S. Drăghici, "Machine Learning and Its Applications to Biology", PLoS Computational Biology 7:953-963, 2007.

[3]. D.W. Hosmer and S. Lemeshow, "Applied Logistic Regression", John Wiley & Sons, New York, 2000.

[4]. B. Scholkopf, K. Tsuda, J.P. Vert, "Kernel Methods in Computational Biology", The MIT Press, 2004.

[5]. M. Yong, Z. Xiao-bo, P. Dao-ying, S. You-xian, W. Stephen, "Parameters Selection in Gene Selection using Gaussian Kernel Support Vector Machines by Genetic Algorithm", Journal of Zhejiang University SCIENCE 2005 6B(10):961-973.

[6]. D.P. Lewis, T. Jebara, W.S. Noble, "Support vector machine learning from heterogeneous data: an empirical analysis using protein sequence and structure", Bioinformatics 22:2753-2760, 2006.

[7]. S. Papadokonstantakis, A. Lygeros, S. Jacobsson, "Comparison of Recent Methods for Inference of Variable Influence in Neural Networks", Neural Net. 19(4):500-513. 2006.

[8]. Y. Tomita, S. Tomida, Y. Hasegawa, Y. Suzuki, T. Shirakawa, T. Kobayashi, H. Honda, "Artifitial Neural Network Approach for Selection of Susceptible Single Nucleotide Polymorphisms and Construction of Prediction Model on Childhood Allergic Asthma", BMC Bioinformatics 5:120-132, 2004.

[9]. 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 Disease", BMC Genetics 7:23-37, 2006.

[10]. K. Miyak, K. Omae, M. Murata, N. Tanahashi, I. Saito, K. Watanabe, "High Throughput Multiple Combination Extraction from Large Scale Polymorphism Data by Exact Tree Method", Journal of Human Genetics 49(9):455-462, 2004.

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

[12]. J. Moore, L. Hahn, M. Ritchie, "Power of Multifactor Dimensionality Reduction for Detecting Gene-gene Interactions in the Presence of Genotyping Error, Missing Data, Phenocopy, and Genetic Heterogeneity", Genet. Epidem. 24(2):150-157, 2003.

[13]. K.H. Liang, Y. Hwang, W.C. Shao, E.Y. Chen, "An algorithm for model construction and its applications to pharmacogenomic studies", Journal of Human Genetics 51:751 - 759, 2006.