상호정보량과 MDR을 이용한 대용량 단일염기다형성 연관성 분석

정현환, 위규범
아주대학교 컴퓨터공학과
e-mail : libe@ajou.ac.kr

An large scale single nucleotide polymorphism analysis method using mutual information and MDR

Hyun-hwan Jeong, Kyubum Wee
Dept. of Computer Engineering, Ajou University
e-mail : libe@ajou.ac.kr

요 약

단일염기다형성 유전형 자료에 대한 유전자형을 얻어내는 기술(genotyping)이 발달함에 따라 분석해야 하는 SNP의 개수가 수십만 개로 증가하였다. 따라서 기존의 연관성 분석(association study)연구 방법을 그대로 적용시키기는 어렵다. 본 논문에서는 상호정보량(mutual information)과 Multifactor dimensionality reduction을 이용하여 대용량의 SNP 유전형자료를 분석하는 방법을 제안하였고, 이 방법을 toluene diisocyanate-induced asthma에 대해 실험해본 결과 높은 판별력을 보이는 모델을 찾을 수 있었다.

1. 서론

단일 염기 다형성(Single Nucleotide Polymorphism : 이하 SNP)과 같은 유전자형(genotype)으로 나타낸 자료를 이용해서 유전자와 질병의 관계를 밝히기 위한 연관성 분석(association study)을 수행하는 연구가 활발히 이뤄지고 있다. 이러한 연구를 통해 단일 SNP이 marginal effect를 가지며 질병과 연관성을 보이는 경우 보다 여러 개의 SNP이 상호 연관성(gene-gene interaction)을 가져 질병이 발생하는 경우가 일반적이라고 보고 있다[1].

불과 몇 년 전만 하더라도 분석해야 하는 SNP의 개수는 수십 개 가량이었으나, SNP 유전자형을 얻어내는 기술(genotying)이 발달함에 따라 분석해야 하는 SNP의 개수는 수십만 개로 증가하였다. 이로 인해 연관성 분석을 할 때 소요되는 계산 양이 많아짐에 따라 기존에 사용하던 방법들이 많은 계산 양을 필요로 하기 때문에 실제 연구에 적용하기가 어려워졌다. 따라서 계산 양을 줄이면서 좋은 성능을 보이는 방법을 개발하고자 하는 연구들이 진행 중에 있다[2].

본 논문에서는 대용량 SNP 유전형자료가 췄을 때, 가능한 모든 SNP 쌍과 질병에 대한 상호정보량(mutual information)의 계산을 통해 연관성이 높은 SNP 쌍들을 뽑은 다음 이에 대해 Multifactor Dimensionality Reduction(이하 MDR) 기법[3]을 통해 발병 여부를 잘 판별(classify)할 수 있는 SNP 쌍을 찾는 연구를 수행하였다.

2. 실험 방법

2.1. 자료의 인코딩

분석을 수행하기 위해 다음과 같이 인코딩 하여 사용하였다. 데이터에 포함된 개인(individual)의 수가 P이고 주어진 SNP의 수가 N일 경우, $P \times N$ 크기의 행렬 G와 P 크기의 배열 C로 표현된다. $G_{i,j}$는 i번째 개인의 j번 SNP의 유전형을 나타내며, -1, 0, 1 혹은 2의 값을 가진다. 만약 -1일 경우 해당 유전형이 결측(missing value)되었음을 뜻한다. 이외의 경우는 관측된 유전형을 뜻하며, 0은 우성-우성 대립유전자(major-major allele), 1은 우성-열성 대립유전자(major-minor allele), 2는 열성-열성 대립유전자(minor-minor allele)를 뜻한다. C_i는 i번째 사람의 발병 여부를 뜻하며, 발병하였을 경우 1, 그렇지 않을 경우 0의 값을 가진다.

2.2. 상호정보량(mutual information)

임의의 SNP쌍과 질병에 대한 연관성을 측정하기 위해서 상호정보량을 사용하였다. 상호정보량을 구하기 위해서는 엔트로피(entropy) 값을 구해야 한다.

한 개의 랜덤변수 X에 대한 엔트로피는 다음과 같이 정의 된다.

$$H(X) = -\sum_{x \in X} p_x \log_2 p_x$$

2개의 랜덤변수 X, Y에 대한 결합 엔트로피(joint

entropy)는 다음과 같이 정의 된다.

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} P_{x,y}\log_2 P_{x,y}$$

위의 식을 조합하여 랜덤 변수 X와 Y에 대한 상호정보량은 다음과 같이 정의 된다.

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

여기서 $I(X;Y)$는 X와 Y가 공유하는 정보의 양을 나타낸다.

엔트로피는 위와 같은 랜덤변수 뿐만 아니라 파티션(partition)된 집합에 대해서도 정의할 수 있다. 본 연구에서는 랜덤변수가 아닌 자료를 이루는 개인들의 자료가 발병 여부나 유전형 값을 통해 나뉠 수 있기 때문에 파티션에 대한 엔트로피를 사용하는 것이 적합하다.

$X=\{A_1,A_2,...,A_n\}$를 집합 S의 파티션이라고 하자. 여기서 $S=A_1\bigcup A_2\bigcup ...A_n$이고, 서로 다른 i와 j에 대해 $A_i\bigcap A_j=\varnothing$ 이다. 파티션 X에 대한 엔트로피 $H(X)$는 다음과 같이 정의된다.

$$H(X) = -\sum_{i=1}^{n}\frac{|A_i|}{|S|}log_2\frac{|A_i|}{|S|}.$$

위의 식에서 $|S|$는 집합의 전체 원소의 개수, $|A_i|$는 파티션에 포함된 원소의 개수를 뜻한다. 결합 엔트로피 역시 같은 방식으로 파티션을 사용하여 정의할 수 있다. 또한 상호정보량은 다수의 파티션에 대해 다음과 같이 확장하여 정의할 수 있다.

$$I(X_1,X_2,...,X_k|Y) = H(X_1)+H(X_2)+...+H(X_k)+H(Y) - H(X_1,X_2,...,X_k,Y)$$

SNP_1, SNP_2, …, SNP_k를 통해 나뉜 파티션을 각각 $X_1,X_2,...,X_k$라 하고, Y를 발병 여부로 나뉜 파티션이라고 할 경우 위의 식은 선택된 SNP들의 상호 작용과 질환의 발병에 대한 연관성을 나타낸다.

본 실험에서는 [4]의 연구에서 사용한 방법과 동일하게 X_1, X_2는 임의의 SNP 2개의 조합으로, Y는 발병 여부로 간주하여 이에 대한 상호정보량을 구하였다. 여기서 $I(X_1,X_2;Y)$의 값이 클수록 발병에 연관성이 있는 SNP들의 조합임을 나타낸다.

2.3. MDR(Multifactor Dimensionality Reduction)

상호정보량은 유전자간의 조합과 질병의 연관성을 측정할 수 있으나 유전형의 조합에 대한 질병 여부를 판별할 수가 없기 때문에 판단 모델(classification model)로 사용할 수 없다. 판단 모델을 만들기 위해서 사용하는 방법인 MDR은 유전자간의 조합과 질병의 연관성에 대한 연관성 분석을 하는데 있어서 많이 사용되는 도구이며 질병을 판별하는 모델을 만들 수 있는 방법이다[3]. MDR은 N개의 SNP을 분석하고자 할 때, 가능한 모든 SNP의 조합들을 헤아린다.

MDR을 통해 SNP칩 자료를 분석하기 위해 사용할 경우 많은 시간이 필요하기 때문에 적합하지 않다. 따라서 본 논문에서는 2.2에서 소개 된 상호정보량을 통해 연관성이 높은 것으로 측정되는 K개의 SNP 조합을 찾아내고, 이들 조합에 대한 MDR 기법을 수행하여 질병을 잘 판별할 수 있는 SNP쌍을 찾아내는 알고리즘을 수행하였다.

2.4. 알고리즘

2.2와 2.3에 소개한 방법들을 이용하여 고안한 알고리즘은 <표1>과 같다. <표1>에서 MDRBA는 주어진 조합에 대한 balanced accuracy[5]를 뜻한다. balanced accuracy(BAACC)는 다음과 같이 정의된다.

$$BAACC = \frac{(sensitivity + specificity)}{2}$$

Balanced accuracy는 MDR 방법을 통해 만들어진 모델을 testing data로 판별 검사를 한 결과에 대해 민감도(sensitivity)와 특이도(specificity) 모두를 고려하는 측정함수(measurement function)이다[5].

N : SNP의 수
P : 개인의 수
G : 개인에 대한 유전형 자료 행렬, $G_{i,j}\in\{-1,0,1,2\}$
C : 질병의 여부에 대한 배열, $C_i\in\{0,1\}$
K : 선택할 SNP조합의 수
top : 상호정보량(MI) 순으로 상위 K개를 저장하는 집합

for i := 1 to N do
  for j := $i+1$ to N do
    mi := i번 SNP과 j번 SNP, 질병에 대한 MI 값
    $top = top\bigcup\{(i,j)\}$
    if $|top| > K$ then
      최하위 MI 조합 삭제

$result = \arg max_{x\in top}(MDRBACC(x))$

<표 1> 분석방법에 대한 의사코드

3. 실험 결과

위에서 제안한 분석법이 좋은 판별 모델을 찾는지 보기 위해 아주대학교 임상시험센터에서 제공한 toluene diisocyanate-induced asthma 질병에 대한 상염색체와 성염색체의 SNP들의 유전형자료를 이용하였으며 각 자료들에 대한 정보는 <표2>와 같다.

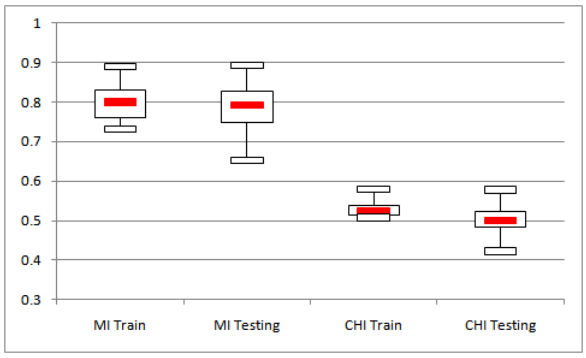
종류	상염색체	성염색체
SNP의 개수	238,304	262,264
환자의 수	76	76
일반인의 수	263	237

<표2> 사용한 SNP유전형 자료의 정보

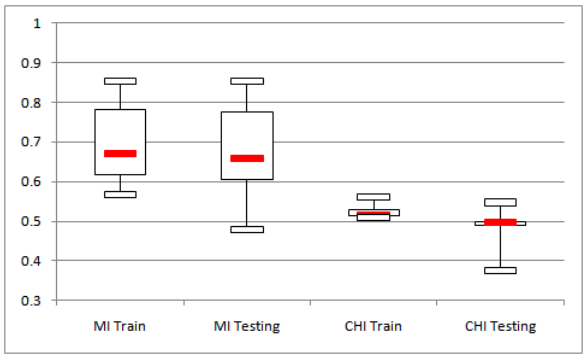
각 자료들에 대해 <표1>에서 제안한 알고리즘을 linux 운영체제, Pentium4 2.8GHz, Ram 1.5GB가 설치된 기계에

GNU C++ 로 프로그램을 구현하여 실험을 수행하였다. 여기서 선택하는 유전형 조합의 개수는 $K=100$으로 설정하였다. 상염색체 유전형 자료의 경우 이를 수행하는데, 18시간이 걸렸고, 성염색체 유전형 자료의 경우 21시간이 소요 되었다.

상호정보량이 MDR방법의 balanced accuracy가 높은 유전형 조합을 찾는지 확인하기 위해서 연관성을 따질 수 있는 기본적인 방법 중 하나인 χ^2 값을 통해 연관성을 따질 경우의 결과와 비교하였다. <그림1>, <그림2>는 앞서 언급한 상호정보량, χ^2로 연관성을 따져 봤을 때 연관성이 높은 것으로 나타나는 상위 100개의 SNP 쌍의 조합에 대해 MDR 방법을 수행한 결과를 박스 플롯(box plot)의 형태로 나타낸 것이다. <그림1>, <그림2>에서 MI는 상호정보량을 뜻하며, CHI는 χ^2를 뜻한다. 또한 Train은 training balanced accuracy를, Testing은 testing balanced accuracy를 나타낸다. 본 결과에서 볼 수 있듯 MI를 사용하여 상호정보량 상위 K개에 대한 MDR을 수행한 결과는 χ^2를 이용하여 실행한 결과 보다 높은 balanced accuracy를 보이는 SNP쌍의 조합들을 찾는 것을 알 수 있다.



<그림 1> 상염색체 유전형 자료의 MDR 결과에 대한 박스 플롯



<그림 2> 성염색체 유전형 자료의 MDR 결과에 대한 박스 플롯

본 실험을 통해서 상호정보량이 높고 MDR방법 에서도 유의하게 나타난 2개 SNP 쌍의 조합들은 <표3>과 같으며, 성염색체 자료의 경우 높은 상호정보량을 보이며 MDR방법에서 유의하게 나타난 SNP 쌍은 존재하지 않았다.

SNP 쌍	상호정보량	MDR Test BA
rs7704589,rs6694813	0.2337(1)	0.8926(1)
rs8037627,rs12460160	0.2337(1)	0.8886(2)
rs7155376,rs1093992	0.2337(1)	0.8881(3)
rs694502,rs9541665	0.2337(1)	0.8847(4)

<표3> 발병에 연관이 있을 것으로 추정되는 상염색체 내의 SNP 쌍

4. 결론

실험 결과를 통해 상호정보량을 이용하여 연관성이 높은 것으로 측정되는 SNP쌍을 찾아냄으로써 모든 조합을 헤아리지 않아도 정확도가 높은 판별 모델을 찾아낼 수 있다는 것을 보였다. 그러나 제시한 알고리즘은 모든 2개의 SNP 조합에 MDR 방법을 적용 시킨 것 보다 빠른 시간을 보이지만 역시 많은 실행 시간을 필요로 한다. 향후 CUDA(Compute Unified Device Architecture)[6]와 같이 병렬 처리 연산을 수행할 수 있는 도구를 통하여 상호정보량을 구하는 부분을 개선시킬 경우 수행시간을 많이 단축시킬 수 있을 것이며, 더불어 헤아릴 수 있는 SNP의 조합의 수를 증가시킬 수 있을 것으로 기대된다.

참고문헌

[1] Wan, X., *et al.,* "Predictive rule inference for epistatic interaction detection in genome-wide association studies.", Bioinformatics, vol. 26, no. 1, pp. 30-37, 2009.

[2] Moore, J., *et al,.* "Bioinformatics challenges for genome-wide association studies.", Bioinformatics, vol. 26, no. 4, pp. 445-455, 2010.

[3] Hahn, L., *et al.,* "Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions.", Bioinformatics, vol. 19, no. 3, pp. 376-382, 2003.

[4] 이중섭 외, "정규상호정보와 지지벡터기계를 이용한 천식관련 단일염기다형성 유전형 자료 분석", 정보과학회 논문지, 제 36권, 제 9호, pp. 691-696. 2009.

[5] Velez D., *et al.*, "A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction", Genetic Epidemiology, vol. 31, no. 4, pp. 306-315, 2007.

[6] Owens J. *et al.*, "GPU Computing", Proceedings of IEEE, vol. 96, no. 5, pp. 879-899, 2008.