복합 질환 관련 SNP 상호작용 예측을 위한 국소탐색 알고리즘

홍원표, 위규범
아주대학교 컴퓨터공학과
e-mail:deblan2@ajou.ac.kr

A local search algorithm for predicting epistatic interactions of SNPs

Won-Pyo Hong, Kyubum Wee
Dept. of Computer Engineering, Ajou University

요 약

최근 GWAS(Genome-wide association study)로 인해 수십만 개의 SNP들이 사용 가능하게 되었다. 그러나 SNP 정보의 양이 방대하여 모든 SNP 조합을 검토하는 방식은 계산 비용이 클 뿐 아니라 오버피팅의 위험이 따른다. 본 논문에서는 필터링 기반 알고리즘인 SNPHarvester의 속도를 개선하고 평가함수를 상호정보량으로 대체하여 실험한다. 기존 SNPHarvester와 비교해 속도면에서 50%가 향상되었고 평가함수 면에서는 기존 SNPHarvester와 동일한 성능을 보였다.

1. 서론

복합 질환(complex disease)이란 다수의 유전자들이 상호작용하여 발병되는 질병을 말한다. 유전정보의 개인차를 나타내는 단일염기다형성(single nucleotide polymorphism; SNP)과 질병의 관련성을 연구하는 SNP 연관성 분석(SNP association study)에서 SNP들의 상호작용은 복합질환의 발병 감수성을 결정하는데 중요한 역할을 한다. 최근 SNP 칩 기술의 발전에 따라 수십만 개의 SNP 유전형 데이터가 사용 가능하게 되었다. 그러나 SNP 정보의 양이 방대하여 모든 SNP 조합을 검토하는 방식은 계산 비용이 클 뿐만 아니라 오버피팅(overfitting)의 위험이 따른다. 따라서 SNP의 개수를 줄여주는 특징 선택 방법(feature selection method)이 필요하다.

특징 선택 방법 중 하나인 필터링 기반 알고리즘은 전처리(pre-processing) 과정이며 상호정보량(mutual information)이나 χ^2-검정값과 같은 간접적인 평가방법으로 유의하지 않은 feature를 걸러낸다. SNP의 개수가 많은 SNP 연관성 분석에서 필터링 기반 알고리즘을 사용할 경우, 첫 단계로 유의하지 않은 SNP을 걸러내 탐색 공간을 줄이고 다음 단계로 병인기전을 모델링한다. 계산 비용이 큰 두 번째 단계에서 SNP의 수가 줄어들기 때문에 결과적으로 총 계산량을 줄일 수 있다. 대표적인 필터링 기반 알고리즘으로는 SNPHarvester[1], Relief[2], ReliefF[3], SURF[4], TURF[5] 등이 있다.

본 논문에서는 필터링 기반 알고리즘의 하나인 SNPHarvester의 속도를 개선하고 평가함수를 상호정보량(mutual information)으로 대체하여 실험한다.

2. 배경

2.1. 상호정보량(mutual information)

랜덤변수(random variable) X의 엔트로피 $H(X)$는 X가 가지는 값에 대한 불확실성을 나타내며 다음과 같이 정의된다:

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

결합 엔트로피(joint entropy) $H(X,Y)$와 조건부 엔트로피 $H(X|Y)$는 각각 다음과 같이 정의된다:

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

$$H(X|Y) = -\sum_{x \in X}\sum_{y \in Y} p(x|y)\log_2 p(x|y).$$

두 랜덤변수 X와 Y가 공유하는 정보의 양을 나타내는 상호정보량 $MI(X;Y)$는 다음과 같이 정의된다:

$$\begin{aligned} MI(X;Y) &= H(X) + H(Y) - H(X,Y) \\ &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= MI(Y;X). \end{aligned}$$

SNP 연관성 분석에서는 SNP의 유전형(genotype)과 질병의 발병 여부와의 관계를 측정하기 때문에 X는 SNP

의 유전형을 나타내고 Y는 질병의 발병 여부를 나타낸다.

집합 S를 전체 데이터라고 하고 $X=\{A_1,A_2,\cdots,A_n\}$을 집합 S의 파티션이라고 하자. 이 때 엔트로피는 다음과 같이 파티션을 이용하여 정의할 수 있다:

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

상호정보량은 다수의 파티션에 대해서 확장하여 정의할 수 있다:

$$MI(X_1,X_2,\cdots,X_k;Y)= H(X_1)+H(X_2)+\cdots+H(X_k)+H(Y) - H(X_1,X_2,\cdots,X_k,Y).$$

여기서 X_i는 i번째 SNP으로 나타내어지는 파티션을 의미하며 $MI(X_1,X_2,\cdots,X_k;Y)$는 k개의 SNP과 발병 여부와의 연관성을 나타낸다.

2.2. SNPHarvester 알고리즘

SNPHarvester는 국소탐색(local search)을 이용하여 유의한 SNP들을 추출하는 알고리즘이다[1]. SNPHarvester는 평가함수의 값이 높은 해를 찾아내는 국소탐색 알고리즘으로 PathSeeker 알고리즘을 사용한다.

PathSeeker는 한 번 호출될 때마다 하나의 경로를 생성하고 하나의 경로는 여러 active set(A)으로 이루어진다. A는 특정 시점에서의 가능성 있는 해(feasible solution) 집합이며 전체 SNP 집합의 부분집합이다. 집합 A의 크기는 PathSeeker의 파라미터인 k에 의해 결정된다. k는 상호작용하는 SNP들의 개수를 나타내며 일반적으로 최대 $\lceil \ln_3 N_d - 1 \rceil$ 까지 고려하는데 이는 데이터 희소성 문제를 방지하기 위해서이다[6].

기존의 SNPHarvester 알고리즘은 다음과 같은 사항을 따른다.

- χ^2-검정값을 평가함수로 사용한다.
- 유의한 SNP들인지의 여부는 Bonferroni correction을 적용한 유의수준으로 판단한다. 이 때 자유도(degree of freedom)는 3^k-1이다.
- 반복적으로 국소 최적치에 도달하는 것을 방지하기 위해 PathSeeker가 종료될 때마다 국소 최적치 집합의 원소 SNP들을 데이터 세트 D에서 제거한다.

본 실험에서는 집합 A의 원소인 SNP들의 상호작용과 질병 간의 관련 정도를 측정하는 평가함수로 상호정보량을 이용하였고 그 값을 $Score(A)$로 표현한다. 집합 A는 평가함수 값이 더 높은 집합을 찾았을 때에만 갱신되기 때문에 $Score(A)$의 값은 점점 증가하고 결국엔 국소 최적치(local optima)에 도달해 PathSeeker 알고리즘이 끝난다.

3. SNPHarvester의 개선

본 실험에서는 기존의 SNPHarvester 알고리즘에서 평가함수를 상호정보량으로 대체한다. 상호정보량을 기준으로 상위 t개의 SNP 집합을 유의한 것으로 본다.

또한 메모이제이션(memoization) 기법을 사용하여 수행 속도를 개선한다. 이는 SNPHarvester의 다음 세 가지 특성에 기인한다.

- 평가함수 값을 계산하는 $Score$ 함수는 호출횟수가 많고 수행 시간의 대부분을 차지한다.
- $Score$ 함수는 파라미터인 SNP 집합에 따라 반환 값이 달라지며 같은 파라미터에 대해서는 같은 값을 반환한다.
- 국소탐색 과정에서 반복해서 여러 번 $Score$값을 계산하는 SNP 집합이 있다.

전체 SNP이 L개이고 두 SNP의 상호작용을 대상으로 하는 SNPHarvester를 생각해보자. SNPHarvester가 고려하는 2-SNP 조합의 참조 횟수를 $L\times L$ 행렬로 나타내면 그림 1과 같은 경향을 띤다.

$$\begin{bmatrix} 0 & & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & & 2 & 1 & 2 & 2 & 2 & 2 & 2 & 2 \\ 0 & & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \cdots & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\ 0 & & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & & 0 & 2 & 6 & 6 & 6 & 6 & 6 & 6 \\ 0 & & 0 & 0 & 4 & 4 & 4 & 4 & 4 & 4 \\ 0 & & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \ddots & & & & & \vdots & & & \\ 0 & \cdots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

(그림 1) 2-SNP 조합의 참조 횟수를 나타내는 $L\times L$ 행렬

모든 가능한 2-SNP 조합의 개수는 $\binom{L}{2}$이므로 그림 1은 상삼각행렬(upper triangular matrix)로 나타낼 수 있다. 원소 a_{ij}는 i번째 SNP과 j번째 SNP으로 이루어진 2-SNP 집합이 SNPHarvester에서 고려되는 횟수를 나타낸다. 그림 1의 행렬에서 많은 원소가 0인데 0이 아닌 원소들 중 2 이상인 원소들이 존재하며 이는 중복 계산을 의미한다. 이와 같은 상황에서 중복된 계산을 줄이기 위해 처음 참조되는 2-SNP 집합의 평가함수 값을 계산하여 메모리에 저장하고, 다음 참조부터는 메모리에 저장된 값을 사용하는 메모이제이션 기법을 사용하였다.

4. 실험

4.1. 데이터

데이터는 Zhang *et. al.*[7]의 보충자료(supplemantary note)를 따라 생성하였으며 가법모형(additive model)에 대해 실험한다.

특성이 다른 세 종류의 데이터에 대해 실험하는데, 각 종류마다 100개의 데이터 세트를 생성하고 각 데이터 세트마다 임의의 두 개 SNP들이 질병과 관련 있도록 생성한다. 모든 데이터 세트는 4000개의 샘플에 대한 2000개 SNP의 유전형(genotype)으로 이루어졌고, 환자군과 대조군의 수는 각각 2000명이다. 각 데이터의 파라미터 설정은 다음과 같다.

- 위험 함수(risk function): SNP의 열성 대립형질(minor allele)이 질병에 영향을 주는 정도를 나타내는 함수이다. 가법모형의 위험 함수는 <표 1>과 같다.
- λ: marginal effect의 크기를 나타내며 0.3로 고정하여 실험한다.
- r^2: SNP간의 연관불균형(linkage disequilibrium; LD)을 나타내며 0.7로 고정한다.
- 열성 대립형질의 발현빈도(minor allele frequency; MAF)는 각각 0.1, 0.2, 0.5로 설정한다.

<표 1> 가법모형(additive model)의 위험 함수(risk function)

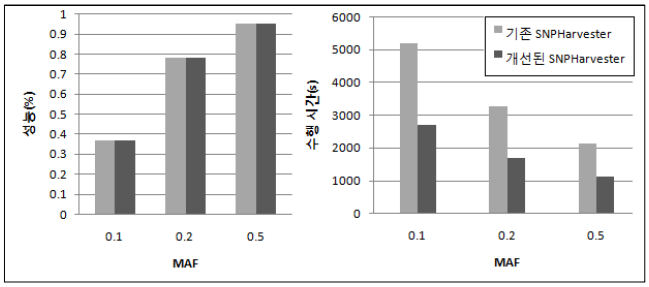
	AA	Aa	aa
BB	1	$(1+\theta)$	$(1+\theta)^2$
Bb	$(1+\theta)$	$(1+\theta)^2$	$(1+\theta)^3$
bb	$(1+\theta)^2$	$(1+\theta)^3$	$(1+\theta)^4$

4.2. 실험 결과

생성 데이터에 대해 다음과 같이 실험한다.

- 연속 20회의 PathSeeker 수행 도중 상위 10개의 SNP 조합에 변화가 없으면 종료한다.
- 기존 SNPHarvester와 개선한 SNPHarvester를 동작시킨다. 여기서 기존 SNPHarvester란 χ^2검정 값을 평가함수로 사용하면서 χ^2검정 값을 기준으로 상위 10개의 2-SNP 조합을 유의하다고 판단하는 것을 말하고, 개선된 SNPHarvester란 기존 SNPHarvester에서 평가함수를 상호정보량으로 대체하고 메모이제이션 기법을 사용한 것을 말한다.

그림 2는 기존 SNPHarvester와 개선된 SNPHarvester의 성능 및 수행 시간 비교를 보여준다. 그림 2에서 성능이란 100개의 데이터 세트 중 질병과 관련 있다고 정해놓은 두 SNP을 SNPHarvester가 유의하다고 판단한 데이터 세트의 개수를 말하고, 수행 시간은 100개의 데이터를 처리하는 과정에서 파일 입력 시간을 제외하고 측정한다.



(그림 2) 기존 SNPHarvester와 개선된 SNPHarvester의 성능 및 수행 시간 비교

성능 면에서는 χ^2와 상호정보량의 차이를 볼 수 없었고 수행 시간에 대해서는 약 50%의 감소를 보여준다. MAF가 증가함에 따라 성능은 증가하고 수행 시간은 감소하는 경향을 보인다. MAF가 낮은 경우 정답인 SNP과 정답이 아닌 SNP간의 차이가 적어 최적 값을 찾아내기가 힘들어 성능이 낮다. 또한 상위 10개의 SNP 집합들이 자주 바뀌기 때문에 SNPHarvester의 종료 조건이 만족되지 않아 수행 시간이 길다.

표 2는 메모이제이션의 효율을 보여준다. 표 2에서 호출횟수는 그림 1에서 모든 원소의 합과 같으며 SNPHarvester가 *Score* 함수를 호출하는 횟수를 말한다. 계산횟수는 그림 1에서 0이 아닌 원소의 개수와 같으며 SNPHarvester가 고려하는 서로 다른 2-SNP 집합의 개수를 말한다. 계산횟수와 호출횟수는 100개의 데이터 세트에 대한 평균값이다. H와 H'는 각각 기존 SNPHarvester와 개선된 SNPHarvester를 의미한다.

표 2와 같이 *Score* 함수의 수행 시간을 45% 내지 50% 감소시킬 수 있으며 이는 SNPHarvester의 수행 시간 감소로 이어진다.

<표 2> 메모이제이션의 효율

MAF	*Score*함수의 호출횟수	*Score*함수의 계산횟수	$\frac{\textit{Score}\text{함수의 계산횟수}}{\textit{Score}\text{함수의 호출횟수}}$	H의 수행시간 (sec.)	H'의 수행시간 (sec.)	$\frac{H'\text{의 수행시간}}{H\text{의 수행시간}}$
0.1	463734.7	238018.6	0.513265	5198.75	2698.28	0.519018
0.2	266694.5	138595.0	0.519677	3273.51	1680.85	0.513480
0.5	189350.2	102963.1	0.543771	2122.50	1143.87	0.538923

5. 결론

본 논문에서는 전체게놈(genome-wide) SNP 연관성 분석을 위해 필터링 기반 알고리즘인 SNPHarvester의 평가함수를 상호정보량으로 대체하고 메모이제이션 기법을 이용하여 속도를 개선하였다. 생성 데이터에 대해 χ^2-검정값과 상호정보량을 각각 평가함수로 썼을 때 비슷한 결과를 얻어 상호정보량 또한 경쟁력 있는 평가함수로 사용할 수 있다는 것을 보였다. 수행 시간에 대해서는 약 50%의 감소를 보임으로써 대용량 SNP 유전형 데이터를 국소탐색으로 처리하는 경우 속도 향상의 방법을 제시하였다.

참고문헌

[1] C. Yang *et. al.*, "SNPHarvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies", Bioinformatics 25, pp. 504-511, 2009.

[2] K. Kira and L. Rendell, "A practical approach to feature selection", Machine Learning: Proceedings of the AAAI '92, 1992.

[3] I. Kononenko, "Estimating attributes: analysis and extension of relief", Machine Learning, ECML-94, pp. 171-182, 1994.

[4] S. Greene *et. al.*, "Spatially Uniform ReliefF(SURF) for computationally-efficient filtering of gene-gene interactions", BioData Mining, 2(5), 2009.

[5] H. Moore and C. White, "Tuning ReliefF for genome-wide genetic analysis" Lect. Notes in Comp. Sci. 4447, pp. 166-175, 2007.

[6] J. Marchini, P. Donnelly and L. Cardon, "Genome-wide strategies for detecting multiple loci that influence complex diseases", Nature Genetics 37, pp. 413-417, 2005.

[7] Y. Zhang and J. Liu, "Bayesian inference of epistatic interations in case-control studies", Nature Genetics 39, pp. 1167-1173, 2007.