Journal of The Korea Society of Computer and Information
Vol. 17, No. 11, November 2012
www.ksci.re.kr
http://dx.doi.org/10.9708/jksci/2012.17.11.141

SNP와 양적 표현형의 연관성 분석을 위한 분류기

엄 상 용*, 이 광 모*

A Classifier for the association study between SNPs and quantitative traits

Saangyong Uhmn*, Kwang Mo Lee*

요 약

인간 유전체 정보와 관련된 기술이 발전함으로 인하여 이를 이용한 질환 또는 질병에 대한 연관성을 분석하여 그 위험도나 치료 예후 등에 대한 예측하기 위한 연구가 활발히 진행되고 있다. 이러한 연구의 대부분은 대표적인 질적 표현형을 대상으로 하는 환자-대조군 연구 (case-control study) 방법을 이용하고 있으며 양적 표현형에 대해서는 개별 단일 염기 변이의 연관성을 회기 분석 방법을 이용하여 규명하는 연구가 주로 수행되고 있다. 특히 복합 질병(complex disease)에 대한 위험도를 예측하기 위한 연구의 경우 흔한 변이 흔한 질환 (common variants common disease)의 가정아래 주로 각각의 단일 염기 변이가 보이는 연관성 정보를 기반으로 진행되고 있으며 여러 변이의 상호 작용에 의한 영향을 분석한 결과는 상대적으로 미비하다.

이 논문에서는 양적 표현형에 대한 SNP의 연관성을 분석하고 그 결과로 발견된 SNP을 이용하여 대상 표현형의 값을 예측하기 위한 분류기를 구성하고 그 성능을 평가하였으며 분류기의 단일 염기 변이의 선택에 있어서 각각의 단일 염기 변이의 연관성을 고려할 때와 단일 염기 변이의 쌍이 보이는 연관성을 고려할 때의 분류 성능을 비교하였다.

▸ Keywords : SNP, 양적 표현형, 결정 트리

Abstract

The advance of technologies for human genome makes it possible that the analysis of association between genetic variants and diseases and the application of the results to predict risk or susceptibility to them. Many of those studies carried out in case-control study. For quantitative traits, statistical analysis methods are applied to find single nucleotide polymorphisms (SNP) relevant to the diseases and consider them one by one. In this study, we presented methods to select informative single nucleotide polymorphisms and predict risk for quantitative traits and compared their performance. We adopted two SNP selection methods: one considering single SNP only and the other of all possible pairs of SNPs.

▸ Keywords : SNP, quantitative trait, decision tree

•제1저자 : 엄상용 •교신저자 : 이광모
•투 고 일 : 2012. 05. 14, 심사일 : 2012. 07. 31, 게재확정일 : 2012. 09. 17.
* 한림대학교 컴퓨터공학과(Dept. of Computer Science, Hallym University)
•이 논문은 2011년도 한림대학교 교비연구비(HRF - 201109 - 052)에 의하여 연구되었음.

I. 서 론

인간 유전체 프로젝트(The Human Genome Project)[1]를 통해 사람의 염기 서열이 알려진 후 유전 정보를 이용한 다양한 질환 또는 질병에 대한 연관성 연구와 이를 통한 진단이나 치료, 나아가 치료 예후 등을 예측하기 위한 많은 연구가 진행되었다[2-7]. 이러한 연구의 대부분은 흔한 변이 흔한 질병(common variants common disease)의 가정 하에 회기 분석 방법 등에 의해 수행되고 있으며, 분석 결과에 대해 p-value를 기준으로 연관성을 보이는 단일 염기 변이(Single Nucleotide Polymorphism, SNP)를 선별하고 있다. 근래에 와서는 앞선 연구에서 발견된 SNP를 이용하여 질병에 대한 위험도를 예측하기 위한 연구가 진행되고 있다[7-10]. 대부분의 위험도 예측은 질병 유무를 다루는 환자 대조군 연구(case-control study)로 연구되고 있으나 현재까지 발견된 SNP의 질병에 대한 유전연관성으로 설명할 수 있는 영향이 미비함으로 인하여 유전정보를 이용한 질환 또는 질환 연관성 연구에 대한 부정적 의견도 일부에서 제기되고 있다. 한편으로는 기존의 흔한 변이(common variants) 외에 희귀 변이(rare variants)를 포함하여 좀 더 많은 SNP를 연구에 이용하기 위한 차세대 시퀀싱(Next Generation Sequencing. NGS) 기법이 이용되고 있다[11-14]. 반면에 양적 표현형(quantitative trait)인 혈압이나 비만도 등 다른 질병에 대해 많은 영향을 주는 주요 지수에 대한 위험도를 예측하기 위한 연구 결과는 상대적으로 많지 않다.

이 논문에서는 양적 표현형에 대한 SNP의 연관성 분석을 통하여 연관성을 보이는 SNP를 발견하고 이들 SNP들과 대상 표현형간의 관계를 결정트리(decision-tree-like)형 분류기로 구성함으로써 SNP 정보로부터 대상 표현형의 값을 예측하도록 하고 그 성능을 실험을 통하여 확인하고자 한다. 또한 각 SNP 뿐 아니라 SNP 쌍이 보이는 연관성을 이용한 방법의 비교를 통하여 SNP의 질병 연관성 분석이 있어서 SNP 상호간의 관계도 고려되어야함을 확인하고자 한다.

이 논문은 다음과 같이 구성되어 있다. 2 장에서는 관련 용어와 연구 사항을 알아보고, 3 장에서는 분류기를 구성하기 위한 SNP 분류 성능의 평가 방법과 분류기 구성 방법 및 분류기의 평가 방법을 설명한다. 그리고 4 장에서는 실험에 사용된 유전형질 및 표현형 자료에 대해 기술하고 실험 결과를 제시한다. 마지막으로 결론은 5 장에서 기술하였다.

II. 관련 연구

1. 단일 염기 변이(SNP)와 전장 유전체 연관성 분석(Genome-Wide Association Study, GWAS)

인간의 유전자는 A, T, G, C 네 가지의 염기의 서열로 되어 있으며 이 유전자 염기서열 중에서 개인 또는 집단 사이의 차이를 가지며 질환에 대한 민감도나 피부 색, 약물에 대한 반응 등에 있어서 차이를 나타내는 것을 SNP라고 한다. SNP의 유전형질 판독(genotyping)이 빠르게 수행될 수 없었던 초기에는 미리 선정된 수 개에서 수십 개의 SNP를 이용하여 대상 질병과의 연관성을 분석하였으나[15] 판독 기술이 발전함에 따라 상대적으로 저렴한 비용으로 단시간 내에 많은 전장 유전체(whole genome)에 대한 정보를 얻게 되었으며 이를 처리하기 위한 기술의 발전과 더불어 전장 유전체에 대한 질병 연관성 연구(Genome-wide Association Study, GWAS) 방법이 가능하게 되었다. 그 결과 유전자의 특정한 부위의 변이에 의해서 발생되는 단일 유전자 질환(single gene disorders 또는 Mendelian disorder)에 대해서는 연구를 통해 많은 사실이 밝혀진 반면 여러 유전자 또는 환경 요인에 의해 발생되는 것으로 추측되는 복합 질병(complex disease)에 대해서 많은 GWAS를 통하여 연구가 진행되고 있으나 괄목할만한 성과를 이루지 못하고 있다[11, 12, 16]. 이에 대한 대안으로 흔한 변이 외에 희귀 변이를 포함하여 연관성을 분석하기 위한 방법이 사용되고 있으나 희귀 변이를 포함함으로써 분석 대상이 되는 자료 양의 증가가 또 다른 문제로 대두되고 있다. 자료의 양에 따르는 문제는 GWAS와 함께 직면한 문제로 기존의 GWAS의 경우도 수십만 개의 SNP를 분석함에 있어 주로 각각의 SNP에 대한 연관성을 분석하는데 그치고 있으며 SNP 쌍이 가지는 정보를 이용한 연구는 아직 많은 결과를 보이지 못하고 있다.

2. 위험도 예측

인간 유전체 정보에 대한 연구가 가능해지기 전에도 다양한 방법을 이용한 위험도 예측 방법이 발표되었다[17,18]. 유전체 정보를 이용하게 된 이후에도 GWAS 이전에는 질환 또는 질병에 대해 미리 선정된 수 개 또는 수십 개의 SNP 중에서 일부 SNP를 선별하여 위험도를 예측하기 위한 연구가 진행되었다. 그러나 GWAS가 가능해짐에 따라 GWAS를 통하여 발견된 많은 새로운 SNP에 대한 정보를 위험도 예측에 이

용하는 것이 시도되고 있다. 그러나 앞선 연구 결과로 발표된 SNP로 설명할 수 있는 질병의 유전적 영향력이 미비함으로 인하여 연구에 대한 부정적 의견도 제기되고 있다[11, 12]. 특히 위험도 예측의 경우는 주로 환자-대조군 연구 또는 질적 표현형(qualitative phenotype)에 대해 수행되었으며 양적 표현형에 대한 위험도 예측에 관해서는 많은 연구가 수행되지 못하고 있다.

III. 평가 함수와 분류기 구성 알고리즘

이 연구에서는 양적 표현형을 대상으로 SNP를 이용한 분류기를 구성하고 그 성능을 평가하고자 한다. 이때 분류기에 사용되는 SNP를 선별함에 있어 단일 SNP의 분류 능력을 이용하는 경우와 SNP 쌍 (pair of SNPs)이 가지는 분류 능력을 이용하는 경우로 나누어 구성하고 그 결과를 비교함으로써 SNP 연관성 분석에서 여러 SNP이 동시에 고려되어야함을 확인하고자 한다. 따라서 상황에 따라 SNP의 분류 능력을 평가하기 위한 함수가 필요하고 이를 바탕으로 SNP를 선별하여 분류기를 구성하고 그 성능을 비교하는 과정을 수행한다.

1. 정의

자료는 N 개의 샘플에 대해 L 개의 SNP, $S_i = \{s_{i1}, s_{i2}, \cdots, s_{iN}\}$ $(1 \le i \le L)$과 한 개의 양적 표현형 p_j $(1 \le j \le N)$로 구성되며, 입력 자료는 p_i의 오름차순으로 정렬되어 있다고 가정한다. 이때 i번 SNP S_i는 {A, T, G, C} 중 두 개의 대립유전자(allele) M_i와 m_i를 가지며, 각 샘플 j에 대해 s_{ij}는 $\{M_iM_i, M_im_i, m_im_i\}$ 중 하나의 유전형질을 가진다. 또한 샘플 j 의 양적 표현형 p_j는 숫자로 표현된다.

2. SNP 평가 함수

이 논문의 목표는 상기한 것처럼 주어진 SNP 중에서 대상 표현형을 두 개 이상의 중첩되지 않는 영역으로 구별하기 위한 분류기를 구성하는 것이다. 이때 입력으로 주어진 모든 SNP를 사용하는 것이 아니라 SNP의 대상 표현형에 대한 영역 구분 능력을 평가하여 일부를 선별하여 사용하고자 한다. 여기서 SNP S_i의 대상 표현형에 대한 영역 구분 능력은 각 유전형질 M_iM_i, M_im_i, m_im_i의 대상 표현형 값들이 가능한 한 중첩되지 않고 분포될수록 높아지며 이를 수치로 계산하기 위한 함수가 SNP 평가 함수로 사용되어야 한다.

2.1 단일 SNP를 위한 평가 함수

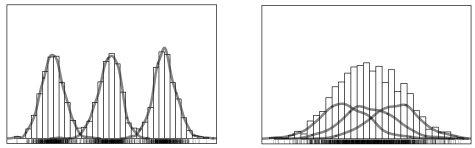

그림 1 SNP의 유전형질별 표현형 히스토그램
Fig. 1. Histogram for a SNP

하나의 SNP S_j는 두 개의 대립유전자 M_j와 m_j로 표현되므로 가능한 유전형질(genotype)은 $\{M_jM_j, M_jm_j, m_jm_j\}$의 세 가지이다. 만약 주어진 표현형에 대해 하나의 SNP S_j가 좋은 분류 성능을 보인다면 그림 1 (a)와 같이 세 유전형질은 서로 겹치는 영역이 많지 않음을 의미하고 이들 영역은 경계선 p_a와 p_b을 이용하여 나눌 수 있다. 각 유전형질이 나타내는 표현형의 범위를 $p_{M_jM_j} < p_a \le p_{M_jm_j} < p_b \le p_{m_jm_j}$라고 가정하면 다음의 식 (1)에 의해 해당 SNP의 분류 성능을 평가할 수 있다.

$$C(S_j) = \frac{N_{M_jM_j} + N_{M_jm_j} + N_{m_jm_j}}{N} \qquad (1)$$

이때 $N_{M_jM_j}$는 유전형질 M_jM_j를 갖는 샘플 중에서 $p_i \le p_a$인 샘플의 수를, $N_{M_jm_j}$는 유전형질 M_jm_j를 갖는 샘플 중에서 $p_a \le p_i \le p_b$인 샘플의 수를, $N_{m_jm_j}$는 유전형질 m_jm_j를 갖는 샘플 중에서 $p_i > p_b$인 샘플의 수를 의미한다. $C(S_j)$가 가질 수 있는 최대값은 1이며 1에 가까울수록 높은 구별능력을 갖는다고 할 수 있다. 반면에 그림 1 (b)와 같은 분포를 보이는 SNP는 많은 영역이 중첩되어 있어 상대적으로 낮은 $C(S_j)$ 값을 갖게 된다.

2.2 SNP 쌍을 위한 평가 함수

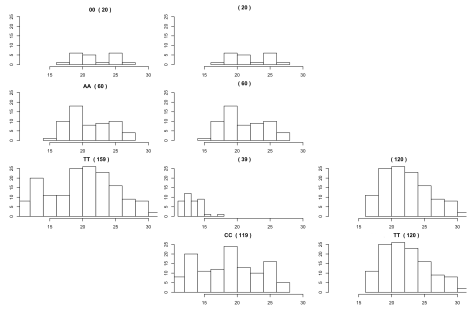

그림 2 SNP 2개에 의한 분류
Fig. 2. Classification with 2 SNPs

그림 1 (b)와 같은 분포를 보이는 경우 하나의 SNP로는 주어진 샘플의 구별이 어렵고 SNP 두 개 이상을 조합하여야 주어진 샘플을 구별할 수 있는 경우를 생각해 볼 수 있다. 이 경우 그림 2와 같이 하나의 SNP에 대해서는 특정 유전형질을 이용하여 표현형의 영역을 구분하는 것이 거의 불가능하지만 두 SNP를 이용하는 경우 영역을 구별하는 것이 가능할 수 있다. 이러한 SNP 쌍을 평가하기 위해서는 SNP S_i의 유전형질과 SNP S_j의 유전형질을 동시에 고려하여 분류 능력을 평가하여야 한다. 이 논문에서 사용한 SNP 쌍 평가 방법은 하나의 SNP에 대한 방법과 유사하며 차이는 두 SNP의 유전형질의 곱집합(product)을 이용한다는 점이다. 즉, 두 SNP S_i와 S_j가 각각 $G_i = \{M_iM_i, M_im_i, m_im_i\}$와 $G_j = \{M_jM_j, M_jm_j, m_jm_j\}$의 유전형질을 가진다면 집합 $\{g_{ij} | g_i \in G_i g_j \in G_j\}$에 대해 그림 1 (a)처럼 영역을 구분하고 각 영역에 대한 샘플 수의 비율로 평가한다. 이를 수식으로 나타내면 다음의 수식 (2)와 같다.

$$C(S_{*i}, S_{*j}) = \frac{\sum_{\substack{g_i \in \{M_iM_i, M_im_i, m_im_i\} \\ g_j \in \{M_jM_j, M_jm_j, m_jm_j\}}} N_{g_ig_j}}{N} \quad (2)$$

이때 $N_{g_ig_j}$는 SNP S_i과 S_j에 대해 유전형질이 $g_i \in G_i$, $g_j \in G_j$인 샘플 중에서 해당 표현형의 값이 설정 영역에 포함되어 있는 샘플의 수를 의미한다.

3. 분류(classification)의 구성

이 논문에서 구현한 양적 표현형을 위한 분류기를 구성하기 위한 기본 흐름은 알고리즘 1같다. 이때 단계 2와 단계 11에서 앞에 기술한 SNP 평가 함수에 의해 주어진 샘플 영역에 대해 가장 좋은 분류 성능을 보이는 SNP를 선택한다. 단계 10에서는 (1) 사용자에 의해 정의된 n_T 개 이하의 샘플만 남았거나, (2) 주어인 영역에 있어서 하나의 유전형질이 90%이상을 차지하는 SNP가 있을 경우 해당 노드를 터미널 노드로 설정한다. 각 터미널 노드는 해당 노드에 의해 분류되는 샘플의 표현형 값의 범위를 저장하고 있으며 입력된 샘플의 유전형질을 이용하여 탐색된 터미널 노드의 표현형 범위와 샘플의 표현형 값을 비교하여 맞고 틀림을 판단한다.

알고리즘 1 분류 알고리즘
Algorithm 1. Classification

1 : $Q \leftarrow \varnothing$
2 : $s \leftarrow$ best SNP based on criteria function
3 : **for** each genotype g_s **do**
4 : $R_{g_s} \leftarrow$ find samples with g_s
5 : Root.addChild((s, g_s, R_{g_s}))
6 : $Q \leftarrow$ addQ(Q, [s, R_{g_s}])
7 : **end for**
8 : **while** not empty(Q) do
9 : [s, R_{g_s}] $\leftarrow$ deleteQ(Q)
10 : if R_{g_s} does not satisfy terminal conditions then
11 : $s' \leftarrow$ best SNP for R_{g_s}
12 : **for** each genotype $g_{s'}$ **do**
13 : $R_{g_{s'}} \leftarrow$ find samples with $g_{s'}$
14 : s'.addChild((s', $g_{s'}$, $R_{g_{s'}}$))
15 : $Q \leftarrow$ addQ(Q, [s', $g_{s'}$, $R_{g_{s'}}$])
16 : **end for**
17 : **end if**
18 : **end while**

표 1 실험 자료의 baseline characteristics
Table 1. Baseline characteristics

Group	Strain	샘플 수	몸무게[2], g
Castle's	A/J	20	23.6±0.74
	129S1/SvImJ	22	22.8±0.71
	DBA/2J	20	23.8±0.39
	AKR/J	20	27.6±0.60
	C3H/HeJ	20	24.3±0.67
	CBA/J	20	26.7±0.54
	BALB/cByJ	20	24.9±0.33
C57	C57BL/6J	20	22.9±0.65
	C57L/J	20	24.4±0.44
Swiss	SJL/J	19	21.3±0.21
	FVB/NJ	20	25.5±0.49
Wild derived	CAST/EiJ	19	12.9±0.42
	PWK/PhJ	20	14.6±0.31
Average			23.73±1.2

VI. 실험

1. 자료

실험을 위한 자료는 포유동물 유전학 연구를 위한 독자적인 비영리 기관인 미국의 The Jackson Laboratory[20]에서 공개한 유전형 자료와 표현형 자료 중에서 Deschepper1 프로젝트 자료를 이용하였다[21]. 표현형 자료는 표 1에서 보이듯이 13 strain에 대해 총 260 마리의 샘플로 구성되어 있으며 각 샘플에 대해 30개의 표현형 정보를 제공한다. 유전형 자료는 The Jackson Laboratory에서 제공하는 575,412개의 SNP로 구성된 자료를 이용하였다. 이 논문의 실험을 위해서는 전체 유전형 자료 중에서 임의로 추출한 300개의 SNP로 구성된 파일을 10개 작성하여 사용하였으며, 대상 표현형으로는 몸무게(body weight)를 이용하였다. 모든 SNP를 실험에 포함하지 않은 이유는 이 논문의 목적이 특정 표현형에 대하여 많은 SNP을 이용하여 의미 있는 분류기를 구성하는 것이 아니라 SNP 평가에 있어서 여러 SNP를 동시에 고려하여야 하는 근거를 확인하고 SNP 평가 함수에 의한 분류기의 성능 비교, 양적 표현형에 대한 분류기의 구현 및 평가이기 때문이다.

자료에 대한 strain과 샘플 수, 몸무게에 대한 기본 정보는 표 1과 같다.

2) 평균±표준편차를 의미함.

2. 실험 방법

앞에서 설명한 10개의 자료 파일 각각에 대해 위에서 설명한 두 가지의 SNP 평가함수를 이용하여 분류기를 구성하고 그 성능을 평가하였다. 이때 분류기의 성능 평가를 위한 평가 수치로 정확도(accuracy)를 사용하였다. 이때 정확도란 입력으로 주어진 샘플 중에서 그 샘플의 표현형 값이 분류기에 의해 분류하였을 때 선택된 터미널 노드가 나타내는 표현형 값 범위에 포함된 경우의 비율로 한다. 이를 수식으로 표현하면 수식 (3)과 같다.

$$\text{Accuracy} = \frac{\sum_{i=1}^{K} CORRCT_i}{N} \quad (3)$$

이때 각 자료 파일에 샘플 수가 많지 않기 때문에 10 fold 교차 검증(cross validation)을 실행하여 성능을 평가하였다. 즉, 입력 파일의 샘플을 10개의 부분집합으로 나눈 후 매 단계에서 9개의 부분집합을 이용하여 분류기를 구성하고 이때 사용되지 않은 샘플 집합을 이용하여 분류기를 평가하는 방법을 이용하였다. 따라서 위 식 (3)에서 K값은 10이며 전체 샘플 수 N은 260이고 $CORRECT_i$는 i 검증 단계에서 바르게 분류된 샘플의 수이다.

3. 실험 결과

이 논문에서 제안한 분류기를 이용하여 위에서 기술한 자료를 분류한 결과는 표 2와 같다. 실험을 위하여 사용된 자료가 특정 질환·질병에 대한 특이성을 가지도록 하여 동종교배한 쥐의 것이기 때문에 그런지 상당히 높은 정확도를 얻을 수 있었다. 또한 원본 파일에서 임의로 추출된 전혀 다른 SNP로 구성된 10개의 자료 파일임에도 불구하고 거의 동일한 실험 결과를 볼 수 있다. 자료 파일의 추가 비교 분석에 의하면 모든 샘플에 대해 동일하게 평가될 수 있는 SNP가 상당히 많기 때문인 것으로 생각된다. 동일한 분류 성능을 가지는 SNP가 다수 존재하고 이들이 각 파일에 나뉘어 포함됨으로 매 선택 단계에서 동일하거나 상당히 비슷한 분류 성능을 제공하는 SNP가 선택되어 나타난 결과라고 할 수 있다. 또한 위에서 기술한 것처럼 단일 SNP에 대한 평가 함수를 이용하여 SNP를 선택한 경우와 비교할 때 SNP 쌍을 기반으로 SNP을 평가하여 선택한 경우에 있어서 이미 단일 SNP 평가 함수를 이용한 실험에서 상당히 높은 정확도를 보였음에도 불구하고 그 정확도가 향상됨을 볼 수 있다. 즉, 단일 SNP에서 평가할 수 없는

SNP와 표현형 사이의 연관성이 있음을 볼 수 있으며 이를 탐색하기 위해서는 둘 이상의 SNP를 동시에 고려하는 평가 방법이 사용되어야 한다.

표 2 분류 정확도(accuracy)
Table 2. Classification accuracy

자료 파일	평가 함수	
	단일 SNP	SNP 쌍
1	99.23	99.61
2	99.23	92.66
3	99.23	99.61
4	99.23	99.61
5	99.23	99.61
6	99.23	99.61
7	99.23	99.61
8	99.23	99.61
9	99.23	99.61
10	99.23	99.61

V. 결 론

이 연구에서는 양적 표현형에 대한 SNP 연관성 분석을 위한 분류기를 구성하여 그 성능을 실험하였다. 이 분류기에서 SNP를 평가하기 위하여 단일 SNP를 위한 평가 함수와 SNP 쌍을 위한 함수를 제안하고 그 성능을 실험을 통하여 비교하였다. 이때 분류 성능은 정확도에 의하여 평가하였으며 이때 정확도는 테스트 샘플의 표현형 값의 범위를 분류기를 이용하여 예측하고 실제 테스트 샘플의 표현형의 값이 이 범위에 포함되는 경우의 비율을 이용하였다.

먼저 실험 결과를 바탕으로 SNP 선택을 위한 평가 함수의 성능을 비교하면 단일 SNP 정보에 기반 한 경우보다 SNP 쌍 정보에 기반 한 경우 정확도가 향상됨을 볼 수 있다. 즉, SNP 쌍을 이용하여 SNP를 선택할 경우 단일 SNP에 의한 정보에서 찾을 수 없는 SNP와 표현형의 연관성을 발견할 수 있다는 것이다. 그러나 주어진 자료에 대하여 모든 SNP 쌍을 평가할 경우 새로운 문제에 직면하게 된다. 즉, 차세대 시퀀싱으로 인하여 상당히 많은 수의 SNP 자료를 이용한 분석에 있어서 모든 SNP 쌍을 처리하기 위하여 필요한 계산 시간과 정보를 저장하기 위한 저장 공간에 의한 문제이다. 이러한 차원의 저주(curses of dimensionality) 상황을 위한 여러 탐색 기법도 제안되었다[22]. 이러한 방법들은 기존의 순차 특징 선택 방법(sequential feature selection)의 문제점, 즉, 이미 선택된 특징이 이후 선택 과정에 미치는 영향을 제거하기 위한 방법으로 제안되었으나 휴리스틱 알고리즘의 특성상 최적의 해를 찾을 수 없는 단점이 있다. 이 논문의 분류기와 관련해서는 SNP 사이의 유사성을 이용한 전처리 과정을 통하여 SNP 수를 줄이는 방법과 그 유사성 평가 방법, SNP 평가 과정에서 SNP 사이의 연관성을 고려하는 방법 등을 생각해 볼 수 있다.

두 번째로 실험 결과를 정확도만으로 판단한다면 상당히 높은 정확도를 확인할 수 있다. 그러나 무감독 분류기의 특성상 대상 표현형의 분류에 대한 정보가 없음으로 해서 특정 유전형질의 비중이 높은 SNP가 주로 선택되는 상황을 결과 분석과정에서 발견할 수 있었다. 이러한 현상은 하위 범주에 대한 SNP 선별과정에서도 반복되는 문제로 적은 샘플 수 와 실험에 사용한 샘플이 가지는 유전형질 분포의 특성에 기인한 것으로 생각된다. 샘플 수가 상당히 증가하거나 샘플간의 이종성(heterogeneity)이 증가되면 발생되는 문제도 감소할 것으로 판단된다. 반면에 분류기의 정확도 측면에서는 그 성능이 떨어질 것으로 생각된다. 이 문제를 해결하기 위한 방법으로 (1) SNP 사이의 유사성 정보를 이용하여 SNP 선별에 고려하거나 (2) 대상 표현형 정보를 기준으로 샘플사이의 연관성을 제공하는 SNP 정보를 상향방식(bottom-up approach)으로 구성한 후 하향방식(top-down approach)에 의해 궁극적인 분류기를 구성하는 방법 등을 고려할 수 있으며, 이러한 해결 방법 등은 향후 연구로 진행하고자 한다.

세 번째로 유사한 SNP에 선택되는 현상이다. 즉, 하나의 SNP에 의해 분류된 샘플 부분집합에 대해 분류 성능을 평가할 경우 이미 선택된 SNP와 유사할수록 하나의 유전형질을 가진 샘플 집합으로 평가되고 이는 높은 분류 성능으로 평가될 수 있다는 것이다. 이 문제를 위하여 SNP 사이의 유사성을 평가하고 모든 샘플에 대해 동일하거나 일정 비율이상 동일한 유전형질을 가지는 SNP를 제거하는 과정이 필요하다.

마지막으로 이 논문에서 구현한 분류기는 각 SNP을 평가함에 있어 기존의 방법[23, 24] 등과 같이 유전형질을 이용한 수식 계산 방법을 이용하지 않는다. 기존의 방법을 이용할 경우 각 SNP의 유전형질을 숫자로 변환함으로써 유전형질사이에 숫자의 크고 작음에 따라 영향력의 크고 작음의 부가적인 의미를 부여하게 된다. 그러나 이 논문에서는 입력 유전형질을 원 표현으로 사용함으로써 유전형질사이의 우열관계를 샘플의 분포를 기반으로 결정하게 된다. 따라서 표현방식에 따른 유전형질사이의 영향력 차이를 내포하지 않는다.

참고문헌

[1] Human Genome Project. *http://www.ornl.gov/sci/techresources/Human Genome/home.shtml*

[2] Y. S. Cho et al., "A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits," Nature Genetics, vol. 41, no. 5, pp. 527-534, May 2009.

[3] S. Ripatti et al., "A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses," Lancet, vol.376, no. 9750, pp. 1393-1400, Oct. 2010.

[4] A. C. Heath et al., "A quantitative-trait genome-wide association study of alcoholism risk in the community: Findings and implications," Biological Psychiatry, vol. 70, no. 6, pp. 513-518, Sep. 2011.

[5] H. D. Daetwyler, B. Villanueva, and J. A. Woolliams, "Accuracy of predicting the genetic risk of disease using a genome-wide approach," PLoS One, vol. 3, no. 10, p. e3395, Oct. 2008.

[6] S. Waaijenborg and A. H. Zwinderman, "Association of repeatedly measured intermediate risk factors for complex diseases with high dimensional SNP data," Algorithms for molecular biology : AMB, vol. 5, p. 17, 2010.

[7] N. P. Paynter, D. I. Chasman, J. E. Buring, D. Shiffman, N. R. Cook, and P. M. Ridker, "Cardiovascular disease risk prediction with and without knowledge of genetic variation at chromosome 9p21.3," Annals of internal medicine, vol. 150, no. 2, pp. 65-72, Jan. 2009.

[8] J. Batsis and F. Lopez-Jimenez, "Cardiovascular risk assessment - From individual risk prediction to estimation of global risk and change in risk in the population," BMC medicine, vol. 8, no. 1, p. 29, 2010.

[9] Z. Wei, K. Wang, H. Qu, H. Zhang, J. Bradfield, C. Kim, E. Frackleton, C. Hou, J. Glessner, and R. Chiavacci, "From disease association to risk assessment: an optimistic view from genome-wide association studies on Type 1 Diabetes," PLoS genetics, vol. 5, no. 10, p.e1000678, 2009.

[10] C. Kooperberg, M. LeBlanc, and V. Obenchain, "Risk prediction using genome-wide association studies," Genetic Epidemiology, vol. 34, no. 7, pp. 643-652, Sep. 2010.

[11] P. Kraft and D. J. Hunter, "Genetic risk prediction-are we there yet?" New England Journal of Medicine, vol. 360, no. 17, pp. 1701-1703, Apr. 2009.

[12] T. A. Manolio et al., "Finding the missing heritability of complex diseases," Nature, vol. 461, no. 7265, pp. 747-753, Oct. 2009.

[13] H. Siu, Y. Zhu, L. Jin, and M. Xiong, "Implication of next-generation sequencing on association studies," BMC genomics, vol. 12, no. 1, p. 322, Jun. 2011.

[14] K. Kahrizi et al., "Next generation sequencing in a family with autosomal recessive Kahrizi syndrome (OMIM 612713) reveals a homozygous frameshift mutation in SRD5A3," European journal of human genetics : EJHG, vol. 19, no. 1, pp. 115-117, 2011.

[15] S. Uhmn, D.-H. Kim, Y.-W. Ko, S. Cho, J. Cheong, and J. Kim, "A study on application of single nucleotide polymorphism and machine learning techniques to diagnosis of chronic hepatitis," Expert Systems, vol. 26, no. 1, pp. 60-69, Feb. 2009.

[16] Online Mendelian Inheritance in Man. *http://www.ncbi.nlm.nih.gov/omim*.

[17] S. J. Pocock, V. McCormack, F. Gueyffier, F. Boutitie, R. H. Fagard, and J.-P. Boissel, "A score for predicting risk of death from cardiovascular disease in adults with raised blood pressure, based on individual patient data from randomized controlled trials," BMJ, vol. 323, no. 7304, pp. 75-81, Jul. 2001.

[18] P. M. Ridker, J. E. Buring, N. Rifai, and N. R. Cook, "Development and Validation of Improved Algorithms for the Assessment of Global Cardiovascular Risk in Women The Reynolds Risk Score," JAMA, vol. 297, no. 6, pp. 611-619, Feb. 2007.

[19] A Catalog of Genome-Wide Association Studies. *http://www.genome.gov/26525384*

[20] The Jackson Laboratory, *http://www.jax.org*.

[21] C. F. Deschepper, J. L. Olson, M. Otis, and N. Gallo-Payet, "Characterization of blood pressure and morphological traits in cardiovascular-related organs in 13 different inbred mouse strains," Journal of applied physiology (Bethesda, Md. : 1985), vol. 97, no. 1, pp.

369-376, Jul. 2004.
[22] P. Pudil and J. Novovi ˇ cov´a, "Floating search methods in feature selection," Pattern recognition letters, 1994.
[23] M. H. Cho et al., "Cluster analysis in severe emphysema subjects using phenotype and genotype data: an exploratory investigation," Respiratory Research, 11:30, March 2010.
[24] Y Guan and M Stephens, "Bayesian variable selection regression for genome-wide association studies and other large-scale problems," Ann. Appl. Stat. Volume 5, Number 3, pp.1780-1815, 2011.

저 자 소 개

엄 상 용

1987 : 한림대학교 전자계산학과 이학사
1997 : 한림대학교 대학원 컴퓨터공학과 공학석사
관심분야 : 알고리즘, 병렬처리, 생물정보학
Email : suhmn@hallym.ac.kr

이 광 모

1975 : 서울대학교 공과대학 응용수학과 공학사
1984 : 서울대학교 대학원 계산통계학과 이학석사
1992 : 서울대학교 대학원 계산통계학과 이학박사
현 재 : 한림대학교 정보전자공과대학 컴퓨터공학과 교수
관심분야 : 프로그래밍언어, 병렬처리
Email : kmlee@hallym.ac.kr