SVM을 이용한 만성간염 환자 예측진단을 위한 SNP정보분석

김동회[0*], 함기백[**], 김진[*]
[*] 한림대학교 컴퓨터공학과
[**] 아주대학교 간소화기 질환센터
{kdh[0*], jinkim[*]}@hallym.ac.kr , kibaek[**]@ajou.ac.kr

Effective Analysis Of SNP Related Chronic Hepatitis Using SNP

Dong Hoi, Kim[0*], Ki Baek, Ham[**] , Jin Kim[*]
[*]Dept of Computer Engineering, Hallym University
[**]Ajou university Medical Center Genomic Research Center for Gastroenteroloy

요 약

Single Nucleotide Polymorphism(SNP)는 인간 유전자 서열의 0.1%에 해당하는 부분으로 이는 각 개인의 체질 및 각종 유전질환과 밀접한 관련이 있다고 알려져 있다. 최근 이 SNP정보의 패턴을 이용 질병의 진단 및 치료에 연관 지으려는 노력이 시도되고 있다. 그러나 아직 SNP를 이용한 효율적인 분석방법에 대한 전산학적 연구는 많지 않다. 본 논문에서는 대표적인 패턴인식기 중 하나인 Support Vector Machine(SVM)을 이용 한국인의 대표적인 유전질환으로 알려진 만성간염에 대해서 관련된 SNP에 대한 패턴 인식율 측정을 실험하였다. 실험 데이터는 간 및 소화기 질환 유전체 센터에서 얻어진 만성간염 환자와 관련 SNP정보를 사용 하였으며, 실험 결과 전체 SNP정보를 모두 가지는 환자그룹에 대한 학습인식율이 66.46%로 나타났으며, 부분그룹에서는 72.91%로 높은 인식율을 보였다. 이 결과는 SNP정보를 이용한 만성간염의 초기 진단예측에 SVM을 효율적으로 사용할 수 있음을 보인다.

1. 서 론

Single Nucleotide Polymorphism(SNP)는 인간의 유전체를 이루는 전체 30억 염기 중 0.1%에 해당하는 부분으로 집단 또는 개인 간 형질 차이를 결정하는 서열이다. 그러나 아직 SNP를 이용한 효율적인 분석방법에 대한 전산학적 연구는 많지 않다. 본 논문에서는 대표적인 패턴인식기 중 하나인 Support Vector Machine(SVM)을 이용 한국인의 대표적인 유전질환으로 알려진 만성간염에 대해서 간질환과 관련된 각 유전자에 대해 패턴 인식율 측정을 실험하였다. 실험 데이터는 간 및 소화기 질환 유전체 센터에서 얻어진 만성간염환자와 해당 환자들의 28개의 관련 유전자정보를 사용 하였으며, 실험 결과 전체 유전자를 학습요소로 사용했을 때는 66.46%로 나타났으며 각 그룹으로 나누어 실험한 결과 IL1B유전자가 72.91%로 가장 높은 인식율을 보였다. 이 결과는 간 질환 관련 유전자로 알려진 많은 유전자들 중 IL1B가 만성간염과 가장 높은 관련성을 가진다고 볼 수 있으며, SNP정보를 이용한 질환예측에 SVM을 효율적으로 사용할 수 있음을 보인다. 본 논문의 구성은 2장에서는 SNP과 SVM에 대하여 설명하고 3장에서는 실험에 사용된 데이터 및 방법에 대하여 설명한다. 4장에서는 실험에 대한 평가를 하고, 마지막으로 5장에서 결론을 맺는다.

2. 관련연구

2-1 SNP

염기서열은 DNA의 경우 {A, T, G, C}의 알파벳으로 이루어진 유한한 길이의 스트링으로 정의할 수 있다. 예를 들면 스트링 "GCCTACCGAGGC"는 DNA의 서열(sequence)이라 할 수 있다. 인간의 서열의 개수는 30억 개이며, 개개인의 서열을 비교하였을 때 99.9%가 동일하다. 하지만 인류 집단 내에서 일부분 0.1%의 차이에 의해 개인 간에 모습이나, 행동 그리고 질환감수성에 차이가 생긴다. 즉 3백만 개 정도의 염기서열 부위에서 서로 다른 염기에 의해 개인 간의 차이 또는 일정 집단이나 인종, 민족 간에 차이가 발생하게 된다. 30억 개 인간 유전체 염기서열 중에서 대략 1.0kb마다 서로 다른 염기가 올 수 있다. 이는 총 3백만 개가 되며, 이를 단일염기 다형성SNP<그림[1]>라고 부른다[2]. 즉 전체 염기서열을 분석하지 않아도 다형성을 보이는 이들 3백만 개 염기서열을 분석한다면 전체 염기서열을 분석하지 않아도, 개인 간이나 집단 간의 유전적 차이, 또는 질환군과 정상인의 차이를 알 수 있으며 따라서 질환의 조기 진단 및 개인별 맞춤의학 분야에 널리 사용될 수 있을 것이다.

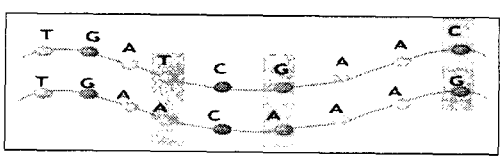

그림 1 Single Nucleotide Polymorphism

2-2 SVM

SVM은 최소의 일반화 에러로 나타나게 하는 최적의 분류 평면(Separating Hyperplane)을 결정하는 기법이라고 볼 수 있다. 일반적으로 선형적으로 분류 가능한 문제의 분류식은

본 연구는 보건복지부 보건의료기반 진흥사업(01-PJ10-PG6-01GN14-0007)의 지원에 의해 이루어진 것임

$$f_{w\cdot b} = sign(w \cdot x + b)$$

와 같이 나타낼 수 있다.

SVM에서 최적의 분류 평면은 서로 다른 클래스들을 구분하는 최대 마진(margin) 사이에 존재한다고 본다. 입력벡터 x_i에 대한 클래스 레이블(label)이 y_i라고 할 때, 최적의 분류 평면은 다음의 제약조건 최소화를 만족해야 한다.

$$Min: \frac{1}{2} w^t w \ where \ y_i(w \cdot x_i + b) \geq 1$$

선형적으로 구분이 불가능한 경우, 위의 최소화 조건은 오분류 데이터를 허용하기 위해 수정되어야 한다. 수정된 식에서 soft margin 분류기가 어느 정도의 에러를 허용하는 대신 제약조건의 위반의 측정치로 새로운 변수인 c를 포함한다. 그리고 a_i가 라그랑지(Largrangian) 계수일 때,

$$Min: L(W) = \frac{1}{2} \cdot < w.w > - \sum a_i y_i [(< w, \varphi(x_i) > + b) - 1]$$

$$0 \leq a_i \leq C$$

$$\frac{\partial L}{\partial b} = 0$$

$$\frac{\partial L}{\partial w} = 0$$

이다. 여기서, C는 ζ_i의 가중치이며, $\phi(\cdot)$는 입력 공간을 보다 고차원의 공간으로 매핑하는 비선형함수이다. 이 때, 위 식의 첫 번째 항을 최소화하는 것은 VC 차원을 최소화 하는 것과 같은 효과이다. 위 식을 풀기 위해서 라그랑지 방법을 이용하여 다음과 같이 변형한다.

$$Max: W(a) = \sum a_i - \frac{1}{2} \cdot \sum a_i a_j y_i y_j K(x_i, x_j)$$

$$0 \leq a_i \leq C$$

$$\sum a_i y_i = 0$$

이 때 $K(x_i, x_j) = <\phi(x_i), \phi(x_j)>$ 인 커널 함수이다.

두 클래스(binary class) 분류에서의 임의의 입력 벡터 X에 대한 분류 함수는 다음 식과 같다.

$$f(X) = sign(\sum_{i=1}^{l} y_i a_i (x_i \cdot X) + b)$$

본 연구에서는 SVM에서 선형 커널을 사용하여 질환예측 문제를 해결하고자 하였다.

3. 데이터 및 방법

3.1 데이터

SVM을 이용한 예측율 실험에 사용된 데이터는 아주대학교 간 소화기 질환 유전체 센터에서 얻어진 간질환 환자 중 SNP가 확보된 만성간염환자와 일반환자 데이터를 이용하였으며, SNP데이터는 간 질환과 관련성을 가진다고 알려진 28개 SNP를 사용하였다. 이들 28개 SNP를 환자그룹에 따라4그룹으로 나누어 실험하였다. 표1은 실험에 사용된 유전자의 SNP들과 해당 SNP가 가질 수 있는 염기와 환자그룹을 나타낸다.

표 1 SNPs

Group	No	SNPs	Sequence	환자수
Group1	1	CCR5(-2459)	G/A	간염:284 정상:351
	2	RANTES(-403)	G/A	
	3	MCP1(-2518)	G/A	
	4	CCRR2-V64I	G/A	
	5	CXCR1-S276T	C/G	
	6	CXCR4-I138I	G/A	
Group2	7	IL1B-31	C/T	간염:261 정상:326
	8	IL1B-511	C/T	
	9	ILIRN-S130S	C/T	
	10	IL1RN-3UTR	C/G	
	11	MBP-G54D	A/G	
Group3	12	IRF1(-410)	G/A	간염:237 정상:200
	13	IFNGR2-Q64R	G/A	
	14	IRF1(-388)	C/T	
	15	IL-10(-592)	A/C	
	16	IL-10(-1082)	G/A	
	17	IFNGR1(-56)	C/T	
	18	IFNGR1(+95)	C/T	
	19	IFNG(+874)	A/T	
	20	TNF-238	G/A	
	21	TNF-308	G/A	
Group4	22	IL18-S35S	C/A	간염:179 정상:173
	23	MMP3-E45K	G/A	
	24	MMP3-D96D	C/T	
	25	MMP3-A362A	C/T	
	26	MMP9-R279Q	G/A	
	27	MMP9-Q688R	G/A	
	28	MMP9-G607G	C/A	

SVM을 이용한 실험을 위해 SNP값을 표2와 같이 A,C,G,T를 1,2,3,4 형태의 정수로 변환하였다. 이때 환자가 위 만성간염이면 1, 정상이면 2로 결과를 변환하였다.

표 2 실험을 위한 변환된 데이터

	실제 값			변환 값			
환자레코드	SNP1	....	SNPn	SNP1	....	SNPn	간염여부
환자 1	CC	....	GG	22	....	33	2(간염)
환자 2	TT	....	CG	44	....	23	2(정상)
환자 3	CT	....	CG	24	....	23	1(간염)
.	.	....	.	.	.	.	.
.	.	....	.	.	.	.	.
환자 n	CC	....	CG	22	....	23	1(간염)

실험에 사용한 입력데이터는 표3과 같다. 표3에서의 Class는 간염의 유무를 나타내며 각 factor는 factor번호 : factor값으로 표현한다.

표 3 SVM 입력데이터 (Class 1:간염 2:정상)

Class	Factor1	Factor2	.....	FactorN
2	1:22	2:44	.....	3:33
2	1:44	2:22	.....	3:23
1	1:24	2:24	.....	3:23
1	1:44	2:22	.....	3:23
2	1:22	2:44	.....	3:33
.	.	.	.....	.
.	.	.	.....	.
1	1:22	2:44	.....	3:23

3.2 실험방법

본 논문에서는 SVM Multiclass[4]를 사용하였으며 학습데이터와 테스트데이터는 전체 환자데이터를 이용하였다. 실험은 28개 SNP정보를 공통적으로 가지는 환자그룹(간염:175명,정상:155)과 환자그룹에 따라 분류한 4개의 그룹을 각각을 실험하였다. 해당그룹의 전체 데이터 중 환자정보 가운데 첫 번째 환자정보를 간염여부를 판정하기 위한 테스트 데이터로써 사용하며, 나머지를 학습데이터로 하였다. 이러한 방식을 모든 환자정보에 라운드로빈 방식으로 적용하여, 전체 SNP에 대해서와 각 그룹에 대해 테스트 데이터를 만들어 이들을 테스트를 함으로서 어느 정도의 정확도를 가지고 질환과 정상을 구별할 수 있는가를 산출하였다. 이러한 방식을 적용한 이유는 환자의 SNP정보를 얻는 것은 매우 어렵기 때문에 기존 정보를 최대한 활용해야 하기 때문이다. 이 실험의 전체 흐름은 그림2와 같다.

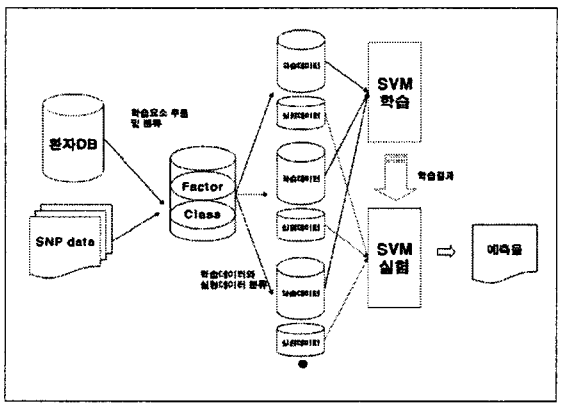

그림 2 전체 실험 구조

우선 환자데이터베이스의 검진결과와 해당 환자의 SNP정보를 이용 표3과 같이 데이터를 변환하고 변환된 데이터 가운데 테스트데이터로 사용할 1개를 제외한 나머지 데이터를 학습데이터로 하여 학습과 테스트를 하였다. 그림3은 실험결과의 일부분이다.

```
Zero/one-error on test set: 0.00% (1 correct, 0 incorrect, 1 total)
Zero/one-error on test set: 0.00% (1 correct, 0 incorrect, 1 total)
Zero/one-error on test set: 0.00% (1 correct, 0 incorrect, 1 total)
Zero/one-error on test set: 0.00% (1 correct, 0 incorrect, 1 total)
Zero/one-error on test set: 100.00% (0 correct, 1 incorrect, 1 total)
Zero/one-error on test set: 0.00% (1 correct, 0 incorrect, 1 total)
Zero/one-error on test set: 100.00% (0 correct, 1 incorrect, 1 total)
Zero/one-error on test set: 100.00% (0 correct, 1 incorrect, 1 total)
Zero/one-error on test set: 100.00% (0 correct, 1 incorrect, 1 total)
Zero/one-error on test set: 0.00% (1 correct, 0 incorrect, 1 total)
Zero/one-error on test set: 0.00% (1 correct, 0 incorrect, 1 total)
```

그림 3 실험 결과 출력

4. 평가

실험 결과 SVM을 이용 선형 커널을 사용하여 실험한 결과는 그림4와 같다

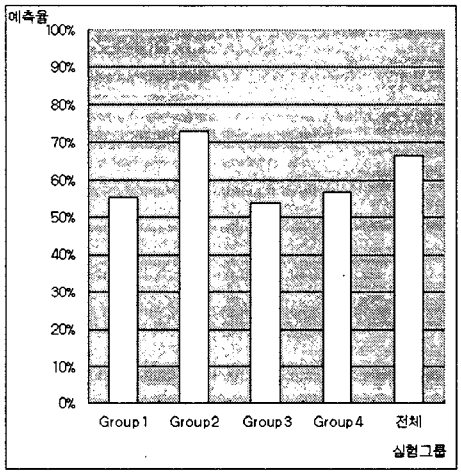

그림 4 각 그룹별 인식율

실험결과 전체 SNP를 이용한 결과 예측율이 66.46%로 나타났으며 Group2가 72.91%로 만성간염의 진단예측율이 가장 높았다.

5. 결론

본 논문에서는 한국인의 대표적인 유전질환으로 알려진 만성간염에 대한 진단예측을 위해 대표적인 패턴인식기 중 하나인 SVM을 적용하였다. SNP 정보를 이용 각종 질환의 유전적 원인규명에 대한 많은 생물학적 연구가 진행되고 있으나 아직 SNP를 이용한 효율적인 분석방법에 대한 전산학적 연구는 많지 않다. 따라서 본 연구는 다양한 유전자 분석에 적용할 수 있다. 향후연구로는 더 높은 인식율을 위하여 환자임상정보, 환자의 생활 패턴 등의 학습요소를 추가할 예정이다. 또한 다양한 질환에 대해서도 본 논문에서의 방법을 적용할 예정이다.

참고문헌

[1] Vapnik, V. N., "The Nature of Statistical Learning Theory," *Springer*, 1995.

[2] Anthony J. Brookes "The essence of SNPs" GENE 1999.

[3] Cherkassky, V., and F. Mulier, "Learning from Data - Concepts, Theory, and Methods," John Wiley & Sons, Inc., 1998.

[4] http://svmlight.joachims.org/svm_multiclass.htm